Second Annual Report

LONI Institute PKSFI Project

June 30, 2008 – June 30, 2009

Principal Investigator:
- Mark Jarrell, LSU, started as PI in May 2009
- Jarek Nabrzyski, LSU, PI from August 2008 until May 2009. Now at the University of Notre Dame

Scientific Coordinator: Bety Rodriguez-Milla, LSU

Projects Committee Lead:
- Shantenu Jha, LSU, started as the Lead on May 2009
- Daniel S. Katz, LSU, Lead until May 2009. Now at the University of Chicago

Co-Principal Investigators:

Louisiana State University	Brooks Keel, Tevfik Kosar, Steven Soper
Louisiana Tech University	Les Guice, Chokchai Leangsuksun, Bala Ramachandran, Neven Simicevic
Southern University at Baton Rouge	Michael Stubblefield, Ebrahim Khosravi, Habib Mohamadian
Tulane University	Gary McPherson, Ricardo Cortez, Lisa Fauci, Donald Gaver
University of Louisiana at Lafayette	Ramesh Kolluru, Devesh Misra, Joe Neigel
University of New Orleans	Scott Whittenburg, Vassil Roussev, Stephen Winters-Hilt

Website: http://institute.loni.org

Introduction

The LONI Institute (LI) was created to build a statewide collaborative computational science environment of faculty, staff, and students who will take advantage of the hardware investments made in LONI to advance research, education, and economic development in Louisiana.

The first year of the project was primarily aimed at recruiting faculty, staff, and students for the project. While we have not filled all the positions available within the LI, many faculty and staff have been recruited. Most of the faculty started during our second year of the project (fall 2008), and have already started addressing the LI milestones. Recruitment for the unfilled faculty and staff positions is ongoing. The student positions were successfully filled. Senior Investigators (SIs), and LI Faculty had conference calls to coordinate the search and chose one student per LI site, making a total of six students (twelve in two years). These students have used LONI facilities to advance their research.

With LI faculty and computational scientists in place, the group has started concentrating more on the research and education, and, in the future, will concentrate on economic development activities outlined in the LI strategic plan.

In terms of hires, in year 1 (Y1), we hired the first five LI faculty, who started their appointments in year 2 (Y2); while in Y2, we hired other four LI faculty members, two of them will start their appointments in Y3.

In Y2, the existing LI faculty and staff (the PIs and SIs), the staff of LONI itself, and the staff of the Center for Computation & Technology (CCT) continued to worked together on numerous projects and proposals, and have held outreach and training sessions across the state. The largest of these projects is, again, the $12M NSF-EPSCOR RII project entitled CyberTools (see http://cybertools.loni.org), which was awarded in October 2007, involves most of the LI Institutions and many of its members, and aims to develop an advanced cyberinfrastructure for statewide research. As this project keeps moving forward, and as LI personnel continue to be hired, CyberTools and the LI are becoming tightly connected, and we expect many new projects to emerge from these. One can view these two projects as layered on top of the LONI infrastructure, with CyberTools focusing on software, services, and applications, and the LI focusing on collaborations and projects.

One important proposal is led by Mark Jarrell, LI Faculty and PI. His team, consisting of a large number of researchers in the state of Louisiana, is working on the next NSF-EPSCOR RII proposal with a material science focus, and under the umbrella of the LI.

On February of this year, we kicked off what we call the LI projects. These projects are proposed by researchers of the State of Louisiana, and approved by the LI scientific committee to receive support from one of our computational scientists. In the first round, based upon technical merit and ability to further LI goals, the scientific committee decided to support 8 projects out of 23 submissions.

In summary, with the new hires in place, research in the main areas that pertain to the LI and other project activities, such as the LI projects, are ramping up, and more faculty, staff, and students are recruited.

Outline of this report

The Board of Regents has requested reports that provide information in the following sections: (1) Personnel; (2) Activities and Findings; (3) Publications and Products; (4) Contributions; and (5) Project Revision. The LI project has many milestones and deliverables that do not necessarily match such a format. Therefore, in the following sections, we discuss in detail the progress towards these specific milestones and deliverables, and we have attempted to group in roughly this order. However, we have combined sections (1) and (2), and reported on the requested activities as appropriate.

Progress on LONI Institute Specific Milestones and Deliverables

As stated in the proposal for this project, the LI has defined numerous metrics to measure project progress and success. These metrics include the hiring of faculty and researchers, creating statewide interdisciplinary research projects and obtaining federal follow-on funding for such, developing corporate partnership programs and start-up companies, developing and following interdisciplinary and multi-institutional collaborations, and creating new educational programs. The performance measures are discussed in detail below, accompanied by project milestone estimates. Deliverables for Y1 are highlighted in yellow, deliverables for Y2 are highlighted in cyan, while deliverables for "the end of year x", including Y2, are highlighted in green.

At the end of each item, we provide information collected for this second annual report. We highlight specific progress by institution where notable.

This report not only contains work done by LI faculty, staff and students, but also data from people who are associated with LI.

1. PERSONNEL, ACTIVITIES AND FINDINGS

In this section, we have combined the first two required sections of the report: 1. Personnel, and 2. Activities and Findings. We provide a description of all the personnel hired during the year, along with activities and findings, as appropriate.

1.I. Personnel Objectives, Metrics and Success Criteria

According to our strategic plan, we have the following personnel Objectives, Metrics, and Success Criteria for measuring when these objectives are achieved, as well as dates expected:

Objective	**Metric**	**Success Criteria**	**Status**
LONI Fellows	Full-time faculty hires, 2 per institution	6 by EOY2; 12 (total) by EO Y3. Nucleation of 6 new multi-institutional research groups by Y3.	9 fellows hired by the EOY2 (Sec. 1.I.a).
Development	Individual hired	1 hire, Fall Y1; new hire	Individual was

Coordinator		in 6 months if position becomes vacant	hired in Y1.
LI Graduate assistantships	Graduate students funded by Institute	6 in each 2 year period; 18 students total over life of project	12 Graduate Fellow positions funded so far (Sec. 1.I.c).
LI Computational Scientist	Individual hired	6 hired in Fall Y1; new hire in 6 months if position becomes vacant	4 CSs hired in Y1 (Sec. 1.I.d).
LI-seeded growth of LONI to national status	Receive federal funding for additional staff	12 staff funded from federal sources by EOY5	12 positions funded in Y2 (Sec. 1.I.b).

1.I.a) Full-time faculty hires

Concerning the full-time faculty hires, our success criterion is to hire of 6 faculty members by the end of Y2. During Y1, five LONI Fellows (LI Faculty) were hired, with most of them having a start date in Y2. These LI Faculty were introduced in the first annual report, and are Drs. Abdelkader Baggag, Dentcho Genov, Mark Jarrell, Damir Khismatullin, and David Mobley.

During Y2, four faculty were hired, with two having a start date in Y3. ULL and LSU continue their searches, review of applications, and interview process. Here is the list of the LONI Fellows by institution, hired in Y2:

SUBR: SU has hired two new LONI faculty members: Dr. Rachel Vincent-Finley (in the Computer Science Department starting in May 2009) and Dr. Zhenyu Ouyang (in the Mechanical Engineering Department starting in August 2009). Dr. Vincent-Finley's primary research interests include numerical analysis, particularly the interface between numerical linear algebra and numerical solution techniques, which occur in applications to biology, chemistry and biomolecular dynamics. More generally, she is interested in computational methods and high performance computing hardware and software methodology for molecular modeling including data structures and computer visualization.

Tulane: During the fall of 2008, Tulane University successfully conducted a search for the second full-time LONI Fellow. Dr. Cortez, Director of the Center for Computational Science at Tulane (CCS), acted as the liaison between departments and the CCS by participating in the search committees and meeting with the candidates. The new LONI Fellow is Dr. Caroline (Caz) Taylor, who will join the department of Ecology and Evolutionary Biology in July 2009. Her research includes computational models of bird migration that use optimization techniques to find bird stopover sites and length of stay. The approach is agent-based models with temporal and spatial features and stochastic components. Dr. Taylor will be an active participant at the CCS and is expected to initiate new computational projects, teach computation courses, and supervise multidisciplinary students.

ULL: UL Lafayette has interviewed multiple candidates for the two LONI faculty positions, one in the College of Science and the other within the College of Engineering. The finalists for both positions have been identified and job offers are being made for a starting date of Fall 2009.

UNO: Dr. Christopher Taylor started on August 2008, he joined in the Department Computer Science. His research primary research involves designing algorithms to analyze genomic data where he focuses on DNA replication, new sequencing technologies, and developmental cancer. He collaborates directly with molecular biologists to investigate biological phenomena using genomic tools such as DNA microarrays and high-throughput sequencing systems.

LA Tech: There were other hires with non-LI funding, this include, Shengnian Wang (Chemical Engineering/IfM), Eric Guilbeau (Biomedical Engineering), Jean Gourd (Computer Science), Paul Hummel (Electrical Engineering – Lecturer), Niel Crews (Mechanical Engineering/IfM), Leland Weiss (Mechanical Engineering/IfM), June Feng (Biomedical Engineering, started on January 2008).

A description of the research done by LI Faculty can be found in appendix A.

1.I.b) Federal funding for additional staff

The LI encourages submission of proposals that add federally-funded staff to LONI. We have 12 staff positions currently funded by the LONI HPCOPS project from NSF. We refer you to http://www.hpc.lsu.edu/about/staff . LA Tech, is also funding 2.5 FTE research support staff on federal grants who were already on soft-funded positions. UNO has hired numerous additional staff, such as post-docs and graduate students using federal funding from a variety of sponsors including NSF, NIH and DoD.

1.I.c) LI Graduate Fellows

The LI supports LONI Institute Graduate Student Research Assistants. Assistantships are available at all member institutions. Research can be in any area of science, engineering, social sciences, or arts and humanities, although the fellow awards are intended to support graduate students whose research projects require access to high-end computing facilities, networks, distributed data archives, and more generally cyberinfrastructure. The awards include a $20,000 stipend and tuition waiver.

The LI Graduate Student Fellows were selected on the basis of

- Excellence in research in the disciplines
- Potential to utilize and advance the infrastructure under development across LONI
- Promise for external funding in the future
- Potential to meet the metrics for success of the LI

Below, we show a list of the Y2 LI Graduate Student Fellows. We also provide a description of the research they have done in appendix D.

Graduate Fellow	Institution	Field
Christopher Clayton / Kimberlee Lyles	SUBR	Computer Science
Jeremy Dewar	Tulane	Mathematics
A. Murat Eren	UNO	Computer Science
John Jack	LA Tech	Computer Science, Institute for Micromanufacturing
Jijun Lao	LSU	Mechanical Engineering
Philip Schexnayder, Jin-Feng Chen	ULL	Physics, Computer Engineering

1.I.d) LI Computational Scientists:

A crucial component of the LI is a strong contingent of advanced staff computational scientists. The LI proposal states that "the LI will support 6 PhD level computational scientists, typically with preexisting postdoctoral experience, to help State research groups take advantage of advanced cyberinfrastructure deployed across LONI and the nation. Distributed across the 6 participating campuses, these staff will be experts in the use of LONI hardware and cyberinfrastructure, including parallel computing, networks, visualization, grids, computational mathematics, and data management. These staff will work closely together, using HD video on all campuses, and will meet biweekly at LSU (supervised by SI Katz). Each of the computational scientists will be assigned 4-5 projects, with duration of 1-2 years each, so that significant progress can be made. These projects will be based on applications from all State campuses, with the applicants being encouraged to commit some internal resources. At least 50% of the projects will be in computational biology and materials science applications; however, we expect projects from other areas of importance to the State, in disciplines ranging from astrophysics, CFD, coastal science, medicine, engineering, digital arts and humanities, and business. This is a total of 70-90 projects over 5 years. Application teams from all State campuses and all companies will be eligible to apply for *LI* partnerships to develop applications that make use of LONI hardware and the staff."

In Y1, we hired 4 computational scientists, Dr. Hideki Fujioka (Tulane), Dr. Raju Gottumukkala (ULL), Dr. Shizhong Yang (SUBR), and Dr. Zhiyu Zhao (UNO). In Y2, LA Tech has had difficulty attracting a suitable candidate for the LI Computational Scientist (non-tenured) position. An offer was made to an outstanding candidate who met all of our expectations; however, he declined. A part-time CS has been appointed (Mr. Abdul Khaliq) who is providing considerable amount of support for the modeling and simulation work in microstructures and devices, quantum chemistry, computational fluid dynamics, and electromagnetics. At LSU, Mr. Andre Merzky worked for the LI part time. LSU has been conducting interviews, and will be bringing in two candidates this summer.

A description of their research can be found in appendix B.

1.II. Research Objectives, Metrics and Success Criteria

Objective	**Metric**	**Success Criteria**	**Status**
LONI Computational Scientists	LI projects underway	12 new projects underway by EOY1; 18 new projects per year thereafter; at least 80 total; 25% projects permitted to be continued for new advances; 25% corporate	10 LI projects (Sec. 1.II.a).
State faculty, staff, and student trained and using LONI infrastructure	Number of applications for time, projects using compute, data, network, and software services	All LI projects use LONI, 12 personnel trained each year from each LI member, medical centers and community college system, 400 active LONI users from State by Y5	826 users, 168 of them have logged on in the past 6 months (Sec. 1.II.b).
National proposals	LI-funded faculty-led national funding agency proposals, submitted and funded	50% of LI projects lead to proposals to agencies outside State (e.g., NSF, DOE, NIH) or industrial funding in Y2 and subsequent years; 2 proposals submitted per year, per LI Fellow, starting in Y2, 96 total, 10 new LI Fellows projects funded total	See Sec. 1.II.c
Research computing project resources	Successful computational infrastructure/cycle applications	50% of projects lead to nationally-judged computational infrastructure awards in Y2 and subsequent years	See Sec. 1.II.c
Research publicity	Invited presentations and lectures outside LA	Each project leads to 2 presentations/lectures per year starting in Y2; 160	See Sec. 1.II.e

		total	
Scientific & Engineering Results	Peer-reviewed conference and journal publications that acknowledge LI support	3 per LONI Fellow per year; 1 per LI project per year; over 150 total	Numerous acknowledging LONI and LI (Sec. 1.II.d).
National Computing Center	LI personnel successful in obtaining federal funding	1 national federally-funded center, funded with at least $70M	None yet.
LI research impact	New non-LI-funded faculty working with LI	6 per year starting in Y2	See Sec. 1.II.f

1.II.a) LI Projects underway

Late last year, the LI put out a call for what we call LI projects http://institute.loni.org/liprojects.php), and by February, the LI scientific committee had chosen 8 out of 23 projects for the LI Computational Scientist (CS) to work on. The criteria the committee used were the relevance of the research proposed, how the project would help the LI achieve its milestones, and an interdisciplinary, inter-, and/or intra-institutional component in each project, among others. We also asked for some of these projects to be adapted so that they will have one of these components, and these turned into other accepted projects. These projects are:

1. "Infrastructure for Accurate and Efficient Binding Affinity Calculations"; PIs: David Mobley (UNO), Steve Rick (UNO), and Shantenu Jha (LSU); LI CS: Hideki Fujioka (Tulane).

2. "Spatial Modeling of the Dynamics of Invasive Nutria"; PI: Azmy Ackleh (ULL); LI CS: Raju Gottumukkala (ULL).

3. "Coupling LONI Institute Computational Scientists, CyberTools and Science Drivers at the Molecular Level"; PIs: Thomas Bishop (Tulane), Shantenu Jha (LSU), and Nayong Kim (LSU); LI CS: Andre Merzky (LSU) at the first stage of the project.

4. "Automated Data Archiving with PetaShare"; PIs: Tevfik Kosar (LSU), Gabrielle Allen (LSU), Sumeet Dua (LA Tech), Frank Löffler (LSU), and Erik Schnetter (LSU); LI CS: Hideki Fujioka (Tulane).

5. "Developing a High Performance Computational Biology and Material Science Lab at Southern University (HPC-BMSL)"; PIs: Ebrahim Khosravi (SUBR), Shuju Bai (SUBR), Rachel Vincent-Finley (SUBR), Shizhong Yang (SUBR); LI CS: Shizhong Yang (SUBR).

6. "Data Management for Disaster Management though PetaShare"; PIs: Ramesh Kolluru (ULL), Tevfik Kosar (LSU), Raju Gottumukkala (ULL), Rusti Liner (ULL); LI CS: Raju Gottumukkala (ULL).

7. "Application Profiling on LONI"; PIs: Erik Schnetter (LSU), Maciej Brodowicz (LSU), Steve Brandt (LSU), and Mayank Tyagi (LSU); LI CS: unassigned.

8. "Surface Plasmon Excitation in inhomogeneous metal-dielectric Composites"; PIs: Dentcho Genov (LA Tech), and Shizhong Yang (SUBR); LI CS: Shizhong Yang (SUBR).

9. “Refinement of Integral Membrane Protein Structure Predictions”; PIs: Christopher Summa (UNO), Steven Rick (UNO), and Zhiyu Zhao (UNO); LI CS: Zhiyu Zhao (UNO).

10. "Parallel-GIS: A High Performance Open Source Geospatial Analysis"; PIs: Ramesh Kolluru (ULL), Baker Kearfott (ULL), Raju Gottumukkala (ULL); LI CS: Raju Gottumukkala (ULL).

Details regarding these projects are provided in appendix C, with more information in appendix B, where our LI CSs explain their research.

Here we list other projects that involve LI SIs, by institution:

LA Tech:

LABRIN, the NIH/INBRE project in Louisiana, is a collaboration between several universities in Louisiana including Louisiana Tech. LA Tech is also part of the current RII project, “CyberTools.” Two PKSFI projects (in addition to the LONI Institute project) at LA Tech have the involvement of LSU and AMRI (UNO), respectively.

Sumeet Dua and Jean Gourd: "DINER: Distributed Information Discovery Laboratory," November 2008, Louisiana Board of Regents, $50,156. A BoR Enhancement grant designed to be a stepping point for students to transition from single-CPU computers to LONI.

Collin Wick: “Molecular level modeling of air/liquid and liquid/liquid interfaces.” We are carrying out investigations of the uptake of possible carcinogens onto fog droplets with molecular simulation, which is in collaboration with experimental researchers at LSU. We are calculating the conductivity in polymer electrolytes for rechargeable lithium batteries. Finally, we are developing molecular models to understand fundamental processes at the air-water interface, including the acidity of the air-water interface, how the interface influences simple reaction rates, and how different ions bind to the air-water interface.

Z. Dick Greenwood: “Monte Carlo Production and Data Analysis for the Dzero and ATLAS Experiments.” The D0 and ATLAS codes require 32-bit compatibility executables and libraries on the Intel EM64T architecture systems. The codes also require the bash/tcsh, python, and perl scripting languages, as well as standard C/C++ libraries. While the D0 binaries are not graphical applications, they do require the OpenMotif, Mesa/OpenGL, and certain X11 libraries. Also, legacy software such as an older glibc is required in some cases.

Natalia Zotov, "Data analysis for LIGO collaboration." Gravitational wave data from various simulations are analyzed to understand detector behavior and key signatures of possible cosmic events. Grid resources at Caltech, Livingston, and Hanford observatories are used in conjunction with LONI resources.

Daniela Mainardi, NSF/CAREER CTS - 0449046, "Modified-Methanol Dehydrogenase Enzymatic Catalysts For Fuel Cell Devices"– PI:, Aug 1st 2005 – Jul 31st 2010. Methanol Dehydrogenase (MDH) enzyme is investigated as a promising anode catalyst for methanol oxidation. Graduate students are studying the role of metal ions in the MDH active site, since the replacement of the original Ca2+ ion by other metal ions modify the activation energy of the enzyme upon methanol oxidation. Density Functional Theory simulations and Transition state calculations are conducted on small portions (models) of the modified-MDH active sites to provide insight on the stability and reactivity of these enzymes upon methanol oxidation.

Andrei Paun, Parallelization of the Nondeterministic Waiting Time algorithm – a biochemical simulation technique. We wish to enhance the nondeterministic component of our algorithm by dividing the biochemical system across many nodes. Dividing the system in this way will allow for a greater degree of reaction competition. Also, we need computational resources to run many biochemical simulations in tandem. Model fitting requires a large number of simulations to be generated. Running these simulations in parallel allows us to move quicker towards developing accurate and reliable models.

Neven Simicevic, Acoustic/elastic FDTD simulations for Underground Imaging Technologies, and electromagnetic simulation.

SUBR:

CS Department has 3 external and 5 internal ongoing (LONI) projects:

External:

1. Secondary structure prediction of gK and UL20 (LBRN pilot project), May 2008- April 2009;
2. NASA-EPSCoR-Dart2 Project: A Study on New Highly Reflective Thermal Barrier Coating, LSU Account # 127-85-4112 (BOR) Proposal # 33204, May 2008-Oct. 2009;
3. NASA-REA: Ab Initio and Experimental Study of A Novel Nano Ceramic Thermal Barrier Material, May 2009-June 2010.

Internal:

1. A reduced simulation method (RSM);
2. The secondary and ternary structure of protein structure prediction;
3. Thermal barrier coating simulation;
4. CFD Simulation of Nucleate Boiling Heat Transfer Enhanced by Micro-Pin Fins;
5. Minority student training

SUBR projects supported following graduate students:

Kiante Roberson: graduate student (1st semester), Computer Science Department, minority (black);
Sadque Ali Mohammed: graduate student (2nd year), Computer Science Department, International (India);
Charles Alphonce Shropshire: graduate student (1st semester), Computer Science Department, minority (black);
Kimberlee Lyles: graduate student (1st semester), Computer Science Department, minority (black);
Murali K. Ganginela: graduate student (1st year), Computer Science Department, International (India);
Frank DeTiege: graduate student (2nd year), Mechanical Engineering Department, minority (black).

Tulane:

Computational and Experimental studies of the transmission of West Nile virus began in 2008 as a collaborative project between the CCS and the department of Epidemiology at Tulane University. This seed project was started thanks to internal funding at Tulane that includes funds for a postdoctoral researcher, laboratory staff, student trainees and laboratory supplies. Dr. Ivo Foppa (Epidemiology department, School of Public Health and Tropical Medicine) and Dr. Cortez lead the project and have hired a postdoctoral researcher, Bree Cummins, who will join Tulane in the fall of 2009. In the summer of 2008, Dr. Cortez conducted a summer research program for undergraduate students, which included a project in computational epidemiology that began exploring appropriate models to use. The student group presented a poster at the SACNAS (Society for Advancement of Chicanos and Native Americans in Science) conference in October 2008 and at the Tulane Health Science Research Days in February of 2009.

CCS has led an NSF funded, focused research group for the development of analytical, computational and experimental tools to investigate the dynamics of elastic structures coupled to a complex fluid. Mucus transport by cilia in the respiratory tract, sperm penetration of the oocyte in fertilization, and peristaltic contractions of the oviduct are examples of such systems. Drs. Fauci and Cortez (Mathematics) lead this effort along with postdoctoral researcher (Dr. John Chrispell) and a graduate student (Ms. Sarah Lukens). LI faculty Fellow Damir Khismatullin has joined the group which meets weekly on this project. Additional collaborators on this project include faculty from the University of California- Los Angeles, New York University, and Washington State University.

As part of a separate EPSCoR grant that includes several of the same universities of the PKSFI project, the PKSFI funds intellectual infrastructure that is being utilized by the EPSCoR project. The latter includes the development of an antibody-based biosensor. This is being approached from theoretical, computational and experimental angles in collaboration with investigators from Tulane, University of New Orleans, Xavier University and LA Tech. The computational scientist hired at Tulane will help develop the computational tools required for the EPSCoR project. This type of synergy is quite important to our success in both projects.

Dr. Gaver continues a collaborative project with Dr. David Halpern (University of Alabama) on the computational investigation of physicochemical and fluid-structure interactions that occur during pulmonary airway reopening. This project involves the investigation of surfactant transport during the reopening of collapsed pulmonary airways, an understanding of which is critical to the development of advanced treatments of acute respiratory distress syndrome. In this project, we computationally model surfactant transport in the occlusion fluid and surfactant uptake to the air-liquid interface as a finger of air propagates through a flexible airway. This reopening process exposes the epithelial cells at the airway wall to large mechanical stresses, the magnitudes of which are predicted by computational simulation. These simulations are performed using a combined boundary element method and volume of fluid approach. The computational facilities available through the LONI are instrumental to the success of this project.

In the past year, Dr. John Perdew and his research groups have developed two new density functionals for the exchange-correlation energy of a many-electron system, which we believe may prove useful in computational materials science. Thus four graduate students and one postdoctoral fellow in our research group are learning to use the LONI and CCS computational infrastructure and the VASP and BAND codes for electronic structure calculations in solids. The first density functional, computationally efficient and accurate for "ordinary matter", is a meta-generalized gradient approximation that yields accurate lattice constants, surface energies, and atomization energies. We will test it extensively for the bulk properties of elemental solids under normal and high pressures, for the monovacancy formation energies of metals and insulators, and for the adsorption energies of molecules on transition-metal surfaces. The second functional, computationally more demanding and possibly accurate for "strongly correlated" materials, is a hyper-generalized gradient approximation, which requires further development before extensive tests can begin.

1.II.b) LONI Users:

Currently, there are 826 Louisiana users with LONI accounts, of whom 168 have logged on to a LONI system in the past 6 months. (Source: The allocations database and the user directory.)

1.II.c) External Funding

Because our LI Projects have just started early this year, we do not expect external funding for LI Projects until late Y3, however, our recently hired LI Faculty have already obtained some external funding. Here we list their grants, and the grants from other LI SIs.

From LI Faculty and LI CS, current Projects/Grants:

Dentcho Genov, DOD RFP - College of Engineering and Science, Louisiana Tech University, Title: "Surface Plasma Enhanced Solar Cell (SPESC)", Funds: $99,616, Date awarded: 06.04.2009

Mark Jarrell, Simulations of Strongly Correlated Electronic Materials, DMR-0706379, $375,000 over the three-year period 09/01/07-08/30/10 by the National Science Foundation, Materials Theory Program.

Mark Jarrell, Graduate Education in Petascale Many Body Methods for Complex Correlated Systems, OISE-0730290, $2,500,000 over the five-year period 9/1/07-8/31/12 by the National Science Foundation, Office of International Science and Engineering (OD/OISE). Investigators: Juana Moreno (PI) UND, M. Jarrell (Co-PI) and K. Tomko (Co-PI) at the Univ. of Cincinnati.

Mark Jarrell, Predictive Capability for Strongly Correlated Systems, DOE DE-FG02-04ER46129, as part of a Computational Materials Science Network, $121,200 over the three-year period 04/15/07-04/14/10 (to be approved year by year) by the Department of Energy, Basic Energy Sciences, CMSN (Warren Pickett, UC Davis, PI).

Mark Jarrell, Next Generation Multi-Scale Quantum Simulation Software for Strongly Correlated Materials DE-FC02-06ER25792 $3,000,000 over the five-year period 7/06-6/11, by the Department of Energy, SciDAC. Investigators: M. Jarrell (PI) and K. Tomko at the Univ. of Cincinnati, Th. Maier (co-PI) and E. DÁzevedo at ORNL, Z. Bai (co-PI) R.T. Scalettar and S. Savrasov at UC Davis.

Mark Jarrell, BoR, "The LONI Institute: Advancing Biology, Materials, and Computational Sciences for Research, Education, and Economic Development", Modification, Change of Principal Investigator. Date Approved: 05/19/2009.

Damir Khismatullin, Department of Defense. Title: Laser nucleation and collapse stability for advanced cavitation power technology (subcontract, completed). Role: Co-I (PI: R. Glynn Holt). Subcontract amount: $395,000 for March-December, 2008.

PI: Erik Flemington, coPI: Christopher Taylor, coPI: Dongxiao Zhu, coPI: Kun Zhang. Title: Administrative Supplements Providing Summer Research Experiences for Students and Science Educators. Source: National Institutes of Health. Supplement to Analysis of Epstein Barr virus type III latency on cellular miRNA gene expression. Amount: $216,386. Date approved: May 15, 2009. Funding Period: June 01, 2009 to August 31, 2010.

PI: Christopher M. Taylor. Title: Taylor Summer Salary Professional Service Agreement. Source: Research Institute for Children. Amount: $36,080. Date approved: March 3, 2009. Funding Period: May 17, 2009 to August 15, 2009.

Hideki Fujioka, support via NIH RO1 – HL81266, NSF EPS-0701491

Zhiyu Zhao, TeraGrid Pathways Fellowship Program, entitled "A Parallel Protein Structure Alignment Tool and a Shared Feature Database for Structures in the Protein Data Bank" was submitted in Feb 09 and approved in Mar 09. Awarded $6,075 to support a student at UNO in the fall 09 semester to develop a parallel protein structure alignment program and a protein feature database under the supervision of the PI (Dr. Zhao), and $2,000 if the PI is going to attend the TeraGrid '09 conference in June (Note: the PI will not be able to attend the conference due to her anticipated baby delivery in June).

Shizhong Yang, Secondary structure prediction of gK and UL20 (LBRN pilot project), May 2008- April 2009

Shizhong Yang, NASA-EPSCoR-Dart2 Project: A Study on New Highly Reflective Thermal Barrier Coating, LSU Account # 127-85-4112 (BOR) Proposal # 33204, May 2008-Oct. 2009.

Shizhong Yang, NASA-REA: Ab Initio and Experimental Study of A Novel Nano Ceramic Thermal Barrier Material, May 2009-June 2010.

From LI Faculty and LI CS, pending Projects/Grants:

Dentcho Genov, Louisiana Board of Regents Support Fund – RCS, Title: "Metamaterials for Applications in STEALTH Technology (MAST)", Funds requested: $ 138,652 Date Submitted: 10.21.2008

Dentcho Genov, DoD-SBIR, Title: "Tunable Electromagnetic Metamaterials Films for STEALT enhancement", Funds requested: $ 70,000 Date Submitted: 03.15.2009

Dentcho Genov, EPSCoR Research Infrastructure Improvement - Track 1 RFP, Title: "Electromagnetic Metamaterials and Active Composites (EMAC)", Funds requested: $ 1,340,444 Date Submitted: 04.20.2009

Mark Jarrell, Ohio Supercomputer Center (OSC), "Improving Developer Productivity for HPC through Cyberinfrastructure: Applications, Languages, Tools and Services", Funds requested: $294,936.00, Date Submitted: 05/26/2009.

Mark Jarrell, Board of Regents, "Louisiana Graduate Research and Education Program in Computational Materials Science", Funds requested: $20,000,000.00, Preproposal/NOI, Date Submitted: 04/29/2009.

Mark Jarrell (coPI), Randall W Hall (PI), BoR, "Planning for LONI Institute s Proposal to the 2009 Louisiana EPSCoR RII Competition", Funds requested: $10,000.00, Date Submitted: 04/20/2009.

Mark Jarrell, DOE, "Predictive Capability for Strongly Correlated Systems: Mott Transition in MnO, Multielectron Magnetic Moments, and Dynamic Effects in Correlated Materials", Funds requested: $67,657.00, Date Submitted: 04/17/2009.

Mark Jarrell, DOE, "TMS: Integrated theoretical approaches to correlated systems", Funds requested: $1,721,689.94, Date Submitted: 03/10/2009.

Mark Jarrell, DOE, "Next Generation Multi-Scale Quantum Simulation Software for Strongly Correlated Materials", Funds requested: $887,664.00, Date Submitted: 03/27/2009.

Mark Jarrell, Ohio Supercomputer Center (OSC), "An Experimental Accelerator-Based HPC System driven by High Productivity Programming Models", Funds requested: $500,000.00, Date Submitted: 11/26/2008.

Damir Khismatullin, National Institutes of Health, National Heart, Lung, and Blood Institute (NIH-NHLBI). Type: RC1 (Challenge Grant). Title: Quantitative analysis of monocyte-endothelium interactions in atherosclerosis (pending). Role: Principal Investigator. Collaborators: Klaus Ley (LIAI), Donald P. Gaver, III (Tulane U.), George A. Truskey (Duke U.).

Damir Khismatullin, National Institutes of Health, National Heart, Lung, and Blood Institute (NIH-NHLBI). Type: R21. Title: Computational studies of leukocyte dynamics using micro-PIV in collagen microchannels (in revision, to be submitted in July 2009). Role: Principal Investigator. Co-I: Sergey Shevkoplyas and Donald P. Gaver, III (Tulane U.).

Damir Khismatullin, National Science Foundation, Nano and Bio Mechanics program. Title: Thrombus rheology via noncontact measurement (in revision). Role: Principal Investigator. Co-PI: R. Glynn Holt (Boston U.).

Raju Gottumukkala, Department of Natural Resources, "Intelligent Flood Protection Monitoring, Warning and Response System", Under Review, 2,891,000 (347K subcontract as Partner Institute)

Shizhong Yang, NASA pending "Novel Nano-Structured Thermal Barrier Coatings" Co-PI;

Shizhong Yang, NSF pending "Nano Ceramic Thermal Barrier Material: Design and Fabrication" Co-PI;

Shizhong Yang, NSF pending "Predictive Quantum Computation and Design of the Catalysts for Green Energy Applications" Co-PI;

Shizhong Yang, NSF pending "Minority Serving Institutions Solar Energy Research Consortium" Co-PI;

Shizhong Yang, NSF pending "Sensor Arrays and the Interpretation of Multi-scale Data Sets" Co-PI.

Zhiyu Zhao, Innovations in Biomedical Computational Science and Technology (R01), NIH PAR-07-344, "Predicting Proteins in SCOP Classification via Alignment and Threading" was submitted in Feb 09. $431,720.00 total direct and indirect costs were proposed for the entire project period of three years ($245,720 on Dr. Fu's part and $186,000 on Dr. Zhao's part to support the proposed research and the study of three graduate students (two at UTPA and one at UNO)).

From LI SIs, Intersite Current Projects/Grants:

DOE/BOR: Ubiquitous Computing and Monitoring System (UCoMS) for Discovery and Management of Energy Resources, $3.3M (total). Allen, Co-PI at LSU. [With ULL, SUBR]

NSF OCI: The LONI Grid - Leveraging HPC Resources of the Louisiana Optical Network Initiative for Science and Engineering Research and Education, $2.2M (Total, LSU). Allen, Senior Investigator. [With LONI]

NSF OCI: Leadership-Class Scientific and Engineering Computing: Breaking Through the Limits (Blue Waters), $208M (Total), $160K (LSU). Allen, PI at LSU. [With NCSA, UIUC and RENCI]

NSF ESPCOR/BOR: Louisiana's Research Infrastructure Improvement Strategy (Includes Cyber-Tools), $12M (Total), October 2007 to September 2010, Allen, PI at LSU (subcontracts to ULL, LA TECH, Southern), Lead of CyberTools Component. [With LA TECH, ULL, SUBR]

NSF MPS: XiRel: A Next Generation Infrastructure for Numerical Relativity, $250K (Total, LSU), September 2007 to July 2010, Allen, Principal Investigator. [With RIT, Georgia Tech, AEI]

BOR PKSFI: Center of Excellence in Integrated Smart Sensor Surveillance System (CyberSpace), $3,638,000 (Total), June 207 to June 2012, Allen, co-PI. [With LA TECH]

SURA/NOAA: SURA Coastal Ocean Observing and Prediction Program, $150K (LSU), December 2006 to August 2008, Allen, PI at LSU. [With SURA, TAMU, RENCI, UAB, VIMS, GOMOOS, UNC]

NSF CNS: MRI: Development of PetaShare: A Distributed Data Archival, Analysis and Visualization System for Data Intensive Collaborative Research, $958K, August 2006 to July 2010, Kosar, PI, Allen, co-PI. [Many partners across state]

NSF LRAC Numerical Relativity and Black Hole Mergers, Computer allocation at National Centers. Over 5,000,000 CPU hours (SUs) across various NSF sites, 2008 to 2009. Allen, co-PI. [With AEI]

NSF EPSCoR/Louisiana Board of Regents, Computational Materials, 2010-2014, $20,000,000 (Perdew, one of many Louisiana Co-PI's).

Donation of current source codes for the solid-state density-functional programs VASP and BAND, from their developers in Vienna and Amsterdam, 2008, Perdew.

National Science Foundation, Density Functional Theory of Electronic Structure (DMR-0501588), June 2005 -June 2009, $372,000, Perdew.

From LI SI, Intersite Pending Proposals:

NSF EAGER, Strategies for Remote Visualization on a Dynamically Configurable Testbed, $300,000. Partners LSU, NCSA, ORNL, Internet2, LONI, August 2009 to July 2011, Allen, Principal Investigator.

NSF STCI Strategies for Remote Visualization on a Dynamically Configurable Testbed, $875,555. Partners LSU, NCSA, ORNL, Internet2, LONI. Allen, Principal Investigator.

NSF PIF Collaborative Research: Community Infrastructure for General Relativity MHD (CIGR), $600,000. Allen, Principal Investigator. [With RIT and Georgia Tech, AEI]

NSF, Center for Ubiquitous Parallel Computing Applications, $375,000, with UIUC, Allen, co-PI. [With UIUC]

LA Tech II-New: DECIDE: Decision Engine for CyberInfrastructure of Distributed agEnts, no funds associated, partnering with LA TECH. Allen, co-PI. [With LA Tech]

DHS, NIMSAT DHS Center of Excellence for Command, Control and Interoperability, $3,608,568. Allen, co-PI. [With ULL]

Recommended for funding by Condensed Matter Theory (Division of Materials Research) program officers: National Science Foundation, Density Functional Theory of Electronic Structure, June 1, 2009- May 31, 2012, $460,000, SI Perdew.

1.II.d) Publications

For publications from personnel directly supported by the LONI Institute, we refer you to Section 2, Publications and Products. The following are peer-reviewed conference or journal publications by LI SIs, which can be cross-institution, and at the interface between biology, materials, and computational science.

Jian Tao, Gabrielle Allen, Peter Diener, Frank Loeffler, Roland Haas, Ian Hinder, Erik Schnetter and Yosef Zlochower, Towards a Highly Efficient and Scalable Infrastructure for Numerical Relativity Codes, to appear, Proceedings of TeraGrid 2009.

Jason G. Fleming, Crystal W. Fulcher, Richard A. Luettich, Brett D. Estrade, Gabrielle D. Allen, and Harley S. Winer, A Real Time Storm Surge Forecasting System using ADCIRC, Estuarine and Coastal Modeling X, M. Spaulding [ed], American Society of Civil Engineers, (2008).

G. Allen, J. Nabrzyski, E. Seidel, G.D. van Albada, J.J. Dongarra and P.M.A. Sloot: in Computational Science - ICCS 2009: 9th International Conference, Baton Rouge, USA, Proceedings, Part I, in series Lecture Notes in Computer Science.

Gabrielle Allen, Philip Bogden, Gerald Creager, Chirag Dekate, Carola Jesch, Hartmut Kaiser, Jon MacLaren, Will Perrie, Gregory Stone, Xiongping Zhang, GIS and integrated coastal ocean

forecasting, Concurrency and Computation: Practice and Experience, Volume 20 Issue 14, Pages 1637 - 1651, (2008).

L.A. Constantin, J,P, Perdew, and J,M. Pitarke, Collapse of the Electron Gas to Two Dimensions in Density Functional Theory, Physical Review Letters 101, 016406 (2008); ibid. 101, 269902 (2008) (E).

E. Sagvolden, J.P. Perdew, and M. Levy, Comment on "Functional Derivative of the Universal Density Functional in Fock Space", Physical Review A 79, 026501 (2009).

Ruzsinszky, J.P. Perdew, and G.I. Csonka, Simple Charge-Transfer Model to Explain the Electrical Response of Hydrogen Chains, Physical Review A 78, 022513 (2008).

D. Lee, L.A. Constantin, J.P. Perdew, and K. Burke, Condition on the Kohn-Sham Kinetic Energy, and Modern Parametrization of the Thomas-Fermi Density, Journal of Chemical Physics 130, 034107 (2009).

J.P. Perdew, V.N. Staroverov, J. Tao, and G.E. Scuseria, Density Functional with Full Exact Exchange, Balanced Nonlocality of Correlation, and Constraint Satisfaction, Physical Review A 78, 052513 (2008).

A.V. Krukau, G.E. Scuseria, J.P. Perdew, and A. Savin, Hybrid Functionals with Local Range Separation, Journal of Chemical Physics 129, 124103 (2008).

J.P. Perdew, A. Ruzsinszky, G.I. Csonka, O.A. Vydrov, G.E. Scuseria, L.A. Constantin, X. Zhou, and K. Burke, Reply to the Comment on "Restoring the Density Gradient Expansion for Exchange in Solids and Surfaces", Physical Review Letters 101, 239702 (2008).

L.A. Constantin, J.P. Perdew, and J.M. Pitarke, Exchange-Correlation Hole of a Generalized Gradient Approximation for Solids and Surfaces, Physical Review B 79, 075126 (2009).

C.A. Jimenez-Hoyos, B.G. Janesko, G.E. Scuseria, V.N. Staroverov, and J.P. Perdew, Assessment of a Density Functional with Full Exact Exchange and Balanced Nonlocality of Correlation, Molecular Physics (special issue in honor of Fritz Schaefer) (to appear).

G.I. Csonka, J.P. Perdew, A. Ruzsinszky, P.H.T. Philipsen, S. Lebegue, J. Paier, O.A. Vydrov, and J.G.Angyan, Assessing the Performance of Recent Density Functionals for Solids, Physical Review B 79, 155107 (2009).

J.P. Perdew, A. Ruzsinszky, L.A. Constantin, J. Sun, and G.I. Csonka, Some Fundamental Issues in Ground-State Density Functional Theory: A Guide for the Perplexed, Journal of Chemical Theory and Computation 5, 902 (2009). (Invited article for the John P. Perdew special issue).

L.A. Constantin, A. Ruzsinszky, and J.P. Perdew, Exchange-Correlation Functional Based on the Airy- Gas Reference System, submitted to Physical Review B.

J.P. Perdew and E. Sagvolden, Exact Exchange-Correlation Potentials in Spin-Density Functional Theory, and their Discontinuities at Unit Electron Number, Canadian Journal of Chemistry. (Invited article for the Tom Ziegler special issue, to appear.)

J.P. Perdew, L.A. Constantin, and A. Ruzsinszky. Energy Densities of Exchange and Correlation in the Slowly-Varying Region of the Airy Gas, Progress of Theoretical Chemistry and Physics, to appear.

J.P. Perdew, A. Ruzsinszky, G.I. Csonka, L.A. Constantin, and J. Sun, Workhorse Semilocal Density Functional for Condensed Matter Physics and Quantum Chemistry, Physical Review Letters, to appear.

L. Cisneros, J. Kessler, R. Ortiz, R. Cortez and M. Bees. Unexpected bipolar flagellar Arrangements and Long-range Flows Driven by Bacteria Near Solid Boundaries. Phys. Rev. Lett. Vol. 101(16), 2008, pp. 168102.

S. Tlupova and R. Cortez. Boundary Integral Solutions of Coupled Stokes and Darcy Flows. J. Comput. Phys., Vol. 228, 2009, pp. 158-179.

1.II.e) Presentations

For presentations from personnel directly supported by the LONI Institute, we refer you to Section 2, Publications and Products.

Here we list some presentations given by LI SIs.

J. P. Perdew. Semilocal "Workhorse" Density Functional for Atoms, Molecules, and Solids, Catalysis from First Principles, Vienna, May 2009.

J. P. Perdew. Strong Correlation in Density Functional Theory, and the Hyper-GGA, Correlated Electrons in Matter , Gatlinburg, Tennessee, April 2009.

J. P. Perdew. Physics of Density Functional Theory, a two-hour tutorial for the workshop Mathematical and Algorithmic Challenges in Electronic Structure Theory (organized by Anna Krylov, John Perdew, Eric Cances, and Juan Meza), Institute for Mathematics and Its Applications, Minneapolis, September 2008.

J. P. Perdew. Promising Fifth-Rung Density Functional: Dobson's ISTLS with Tests for Uniform Gases, Planar Surfaces, and Quantum Wells, International Society for Theoretical Chemical Physics VI, Vancouver, July 2008.

J. P. Perdew. Restoring the Gradient Expansion for Exchange in a Generalized Gradient Approximation for Solids, Surfaces, and Large Organic Molecules, Quantum Systems in Chemistry and Physics XIII, East Lansing, Michigan, July 2008.

R. Cortez. Regularized Stokeslets and other elements with applications to biological flows, Joint Mathematics Meetings, Washington, D.C. (January 5, 2009).

R. Cortez. Regularization Methods for Simulations of Biological Flows, Plenary Talk, Louis Stokes Alliance for Minority Participation Fourth Transdisciplinary Research Conference, Uniersity of Puerto Rico, Mayaguez, PR, (December 5, 2008).

R. Cortez. Interaction of Rotating Helical Bacterial Flagella With Nearby Solid, APS Division of Fluid Dynamics Annual meeting, American Physical Society, San Antonio, TX, (November 25, 2008).

R. Cortez. Regularized Stokeslets and other elements with applications to biological flows, Mathematical Biology Seminar, University of Utah, Salt Lake City, (October 8, 2008).

1.II.f) LI Research Impact

There are two new non-LI faculty working with the LI:

Dr. Juana Moreno, LSU, who has a joint appointment in the Physics Department, and at the Center for Computation and Technology. In her research, she has focused on the experimentally relevant transport and magnetic properties of correlated electron systems, including diluted magnetic semiconductors, heavy fermion compounds and low-dimensional systems. These materials share in common unexpected properties which cannot been explained with conventional approaches. Computer simulations are an increasingly efficient means to study correlated systems. In her research, she has used a variety of computational tools, such as the dynamical mean-field theory and the dynamical cluster approximation in the study of diluted magnetic semiconductors and the density matrix renormalization group method in the area of low-dimensional materials. In the near future, she plans to extend these investigations to nanoscale quantum dots and heterostructures of magnetic semiconductors, where the confined geometry plays a crucial role, to incorporate orbital degrees of freedom in the modeling of heavy-fermions and cuprates, to calculate transport properties in nano systems and to simulate complex materials quantitatively using parameters extracted from first principles calculations.

Dr. Amitava Jana, SUBR, Department of Mechanical Engineering. Dr. Jana has been involved in research and educational activities in the area of Mechatronics and robotics. He has developed a Mechatronics/ robotic laboratory to offer a multi-disciplinary course for electrical and mechanical engineering students.

Xavier University is currently in the process of preparing two proposals to fund Xavier's connection to LONI. Xavier University has been working with Les Guice and Lonnie Leger on the project to lay fiber between Xavier and LSUHSC. Once funding is secured and Xavier University is being connected to LONI, they will be very interested in becoming a part of the LONI Institute team.

1.II.g) Awards

A team of 13 LSU researchers and students, led by SI Gabrielle Allen at the LSU Center for Computation & Technology (CCT), conducted a presentation and demonstration that won first prize at the SCALE 2009 challenge at CCGrid09, a premier conference for cluster and Grid computing. (http://www.hpcwire.com/offthewire/LSU-led-Black-Hole-Simulation-Wins-First-Prize-at-International-Competition-46469337.html)

1.III. Economic Development Objectives, Metrics and Success Criteria

Objective	**Metric**	**Success Criteria**	**Status**
Student internships with companies	Number of placements	2 students placed each year; 20 total (not all will be LI-funded)	From students under the supervision of faculty associated with the LI (Sec. 1.III.a).
Pilot program with Council on Competitiveness	Program established	15 students at community college trained in CS each year, 30 total placed in companies, 10 enter universities for continued study in CS	None
Industrial partnerships	Partnerships in projects with industrial partner (any company who has joint project with LONI)	25% of total projects; 20 partners in 5 years	In progress (Sec. 1.III.b).
Industry grants	Sponsored research from companies	25 by Y5 across all sites	In progress (Sec. 1.III.c).
Centers of Excellence (UIRCs)	Number formed with multi-year duration	1 by EOY3, 3 by EOY4, 5 by EOY5, all industry-funded with at least 1 industry staff member on-site (across all LI sites)	Two proposals submitted (Sec. 1.III.d).
New companies formed	Number of new companies	1 by EOY3, 3 by EOY4, 6 by EOY5	See Sec. 1.III.e.

1.III.a) Students doing Internships

Some institutions have students doing internships with companies. Even though they are/were not supported by the LI, the SIs are tightly connected with the LI.

SUBR has placed six students, Kimberlee Lyles, Tayler Washington, Marlon Gichie, Tiffany Wilkerson, Jayme Chustz, and Daniel Henderson.

UNO has numerous internship programs, which have placed students at local companies. Most internships are run through the Career Office and can be located at recruit.uno.edu. The College of Business offers a large internship program through its Link the Internship through Future Employment (LIFE) program. The Naval Architecture and Marine Engineering Department (NAME) offers a variety of internships with interns currently placed at Tidewater and Bollinger Shipyards and several placed through the Naval Research Enterprise Internship Program (NREIP). The Computer Science Department has three interns working with the local SSC-Atlantic New Orleans offices and several students from both Computer Science and Business are interns via the Student Career Experience Program (SCEP).

At LA Tech, a number of their engineering undergraduate students and some graduate students take advantage of co-op opportunities. It is hard to get a list of names but the head-count is in the range of 20-25.

In June-July 2008, Tulane's Center for Computational Science hosted 8 undergraduate students from Tulane (6), Xavier (1) and Dillard (1) Universities for a research program in computational fluid dynamics. Four of the student participants are from minority groups. The program discussed graduate school and prepared them to write technical reports and give presentations. The two minority students who graduated in May 2009 are going to graduate programs in UNC and Duke. Several of the students also attended the annual conference of SACNAS (Society for Advancement of Chicanos and Native Americans in Science). These students presented a poster with their results and participated in networking, scientific symposia and workshops. Two student participants also worked with Prof. Cortez during the 2008-09 academic year in senior projects related to computational science.

Tulane postdoctoral researcher John Chrispell and graduate student Sarah Lukens spent four weeks in May 2009 at NYU's Applied Mathematics Lab learning experimental techniques to complement their computational modeling training.

1.III.b) Partnership with Companies with a Joint Project with the LI

SUBR has a partnership with IBM.

At ULL, the partnership with Louisiana GOHSEP (Governors' Office of Homeland Security and Emergency Preparedness) would leverage LONI resources for disaster response.

LA Tech's Neven Simicevic and Erez Allouche have been working with a large company that dominates the construction equipment market (we cannot reveal the name of the company

because of a non-disclosure agreement). Neven has developed very detailed electromagnetic simulations using finite-difference-time-domain (FDTD) approach for this project. Also, one of the first TIP grants awarded last year by NIST now supports work between Louisiana Tech's Trenchless Technology Center (Allouche, Simicevic) and a company in Florida. Electromagnetic simulations using LONI hardware are a key component of this work.

Tulane has a partnership with IBM that provides 12 $4,000 fellowships to graduate students in computational science and engineering.

1.III.c) Sponsored Research with Companies

UNO has several sponsored research programs with companies. Several are through our SPAWAR contract in which UNO is the prime and we work with mostly local companies as subcontractors.

1.III.d) Formation of a Center of Excellence (UIRC)

ULL submitted a proposal to establish the NIMSAT Institute as a DHS Center of Excellence. This proposal, while ultimately not funded, was shortlisted nationally as one of the 3 finalists.

The Center for Secure Cyberspace, at LA Tech, was initiated through the BoR/PKSFI program. (http://csc.latech.edu/)

1.III.e) Progress in Forming New Companies Related to Faculty, Staff or Students

A non-LONI faculty member in the Computer Science Department at UNO is forming a new company based upon his research in bioinformatics. They have just received their DUNS number.

1.IV. Collaboration Objectives, Metrics and Success Criteria

Objective	**Metric**	**Success Criteria**	**Status**
Between computational scientists and biologists, materials	Joint papers and proposals	2 interdisciplinary papers (including preprints from a LI preprint series) per group per year; 1 at interface between bio, materials, computation per group per year; 50% of proposals have 2 of 3 disciplines	Numerous from faculty and staff associated with the LI (Sec. 1.II.c and 1.II.d).
Inter-university	Number of joint papers and proposals	2 papers, 1 proposal (including preprints from a LI preprint series) per group per year	Numerous from faculty and staff associated with the LI (Sec. 1.II.c and 1.II.d).
Inter-university	New joint projects	30 new multi-university projects proposed to SC per year	Numerous from faculty and staff associated with the LI (Sec. 1.II.a, appendices B, C)
National	Visits to national labs	3 students, 2 staff, and 6 faculty with visits to national labs per year, 2-3 each summer across all sites	From students under the supervision of faculty associated with the LI (Sec. 1.IV.a).

1.IV.a) Visits to National Labs

Dr. Ramesh Kolluru, Executive Director of the NIMSAT Institute and Mr. Dean Mallory, Assistant Director of the NIMSAT Institute, both visited the US Department of Homeland Security's National Infrastructure Simulation and Analysis Center (NISAC), located in Albuquerque, jointly hosted by Sandia National Laboratory as well as the Los Alamos National Laboratories. A Memorandum of Understanding is under development between the NIMSAT Institute and NISAC to advance joint R&D activities in critical infrastructure modeling and analysis.

Dr. Chris Taylor, UNO, will visit the DOE Joint Genome Institute in Walnut Creek, CA for a workshop from June 1-5, 2009. He is travelling with a colleague (microbiologist) who he is forming a collaboration with.

Twelve SUBR faculty members visited Oak Ridge National Lab in January 2009.

From LA Tech, there is a number of faculty and students who visited a national lab:

Faculty:

Z. Dick Greenwood to CERN, Brookhaven and LIGO
H. Lee Sawyer to FermiLab and CERN
Marcus Wobisch at FermiLab and DESY, Germany
M. Arov (postdoc) to FermiLab and CERN (extended visits – 3 to 6 months)

Students:

Ram Dhullipuddi (PhD/ENGR) to CERN (extended visit – 6 months)
Emile Frey (PhD/ENGR) at CERN (Summer)
Tracie Reed (PhD/CAM) to LIGO-Livingston (first as a trainee and then as "expert monitor")
Scott Atkins (PhD/CAM) – Summer at FermiLab
Kiran Chakravarthula (PhD/ENGR) – Summer at FermiLab
Mark Wade (BS/EE & Phys) – Summer at Fermilab
John Jack (LONI Institute Fellow 2008-09): EPA (Washington, DC) – Summer internship. He will initially try to adapt the NWT algorithm for the modeling of a virtual human liver.

In May, SI, Dr. Gabrielle Allen, LSU, visited Fermilab for an Open Science Grid meeting.

1.IV.b) Collaborations

J. P. Perdew coauthors who are Physical Chemists, not Physicists: M. Levy, A. Ruzsinszky, G.I. Csonka, K. Burke, D. Lee, V.N. Staroverov, G.E. Scuseria, A.V. Krukau, A. Savin, O.A. Vydrov, C. Jimenez-Hoyos, B.G. Janesko, S. Lebegue, J.G. Angyan

J. P. Perdew coauthors outside Louisiana: J.M. Pitarke, P.H.T. Philipsen, J. Paier, M. Levy, G.I. Csonka, K. Burke, D. Lee, V.N. Staroverov, G.E. Scuseria, A.V. Krukau, A. Savin. O.A. Vydrov, C. Jimenenez-Hoyos, B.G. Janesko, S. Lebegue, J.G. Angyan.

R. Cortez coauthors outside Mathematics: L. Cisneros, J. Kessler, M. Bees, I. Foppa, D. Gaver, D. Khismatullin.

R. Cortez coauthors outside Louisiana: L. Cisneros, J. Kessler, M. Bees, S. Tlupova, K. Leiderman, M. Shelley, J. Zhang, J. Teran, R. Dillon, D. Varela.

D. P. Gaver coauthors outside Biomedical Engineering: D. Halpern, R. Cortez.

D. P. Gaver coauthors outside Louisiana: D. Halpern

1.V. Education and Training Objectives, Metrics and Success Criteria

Objective	**Metric**	**Success Criteria**	**Status**
Statewide education	HD video courses offered	4 courses per year with students from 4 universities, and 20 total students per course receiving credit.	1 course used LONI (Sec. 1.V.a).
Statewide training	Number of training workshops, people trained	Initially 2 HPC & CSs workshops offered per year, increasing to 4 by Y5; at least 50 people trained each year, 400 total	26 tutorials (319 attendees) and 4 workshops (104+ attendees) offered (Sec. 1.V.b).
High school education	Summer camps	1 per year for LI members	2 Summer Camps (Sec. 1.V.c).
High school courses	Teachers offer LI-related material in courses	10 new teachers offer classes with LI material each, year starting in Y2	None yet.

1.V.a) HD Video Courses

In Spring 2009, LI faculty, Mark Jarrell, as part of an NSF PIRE project, taught a course entitled, "Advanced Solid State Physics". In addition to the traditional subjects, this course also covered a number of modern computational methods, such as dynamical mean field theory, quantum Monte Carlo, etc. In addition to a number of students at LSU (both registered and audits numbered around 10) the course was taught via asynchronous video to students in Germany and Switzerland (an additional 10 students).

1.V.b) Workshops and Tutorials on HPC and Computational Sciences

The LI staff computational scientists have worked with LONI staff and its member campuses to develop and hold training workshops on the use of LONI and its advanced cyber-services, as

well as annual conferences and workshops. Themes are based on overlaps between various partnerships, such as application-based workshops and tools-based workshops.

Here, we provide a list of the workshops and tutorials LONI and HPC LSU organized, as well as the number of participants. Even though we only list the events after Summer 2008, these workshops and tutorials have been offered previously as well. Many faculty, research associates, graduate and undergraduate students from across the State attended and have received training on LONI.

Semester	Training	No. Enrolled
	Tutorials	
Summer 2008		
2-July	Introduction to the HPC Environment	10
14-July	Introduction to OpenMP	12
30-July	Introduction to MPI	8
Summer 2009	Total	30
Summary	**30 people were trained in Summer 2008 in 3 tutorials**	

Semester	Training	No. Enrolled
	Tutorials	
Fall 2008		
16-Sep	Introduction to Linux and Vi	17
18-Sep	Welcome to HPC: accounts, allocations, Linux and Linux cluster environment	18
24-Sep	Introduction to MPI	10

29-Sep	MPI Part 2	5
1-Oct	Introduction to OpenMP	7
8-Oct	OpenMP Part 2	4
15-Oct	Introduction to Debugging and Profiling	15
27-Oct	Cluster Compilers and Optimization	9
27-Oct	Introduction to Debugging with Totalview	10
5-Nov	Practical MPI	5
Fall 2008	Total	100
	Workshops	
October 22 & 23	LONI HPC Workshop, at LA Tech	30
Nov. 25	LONI HPC Workshop at ULL	25
Fall 2008	Total	55
Summary	**100 people were trained in Fall 2008 in 10 tutorials and 55 people attended 2 workshops.**	

Semester	Training	No. Enrolled
	Tutorials	
Spring 2009		
28-Jan	Introduction to Linux and Vi	5

29-Jan	Welcome to HPC: accounts, allocations and the cluster environments	15
4-Feb	Introduction to MPI	5
11-Feb	Practical MPI	7
18-Feb	Introduction to OpenMP	11
26-Feb	OpenMP Part 2	5
2-Mar	Introduction to MATLAB	17
12-Mar	An introduction to the computational chemistry package, Gaussian 03	18
16-Mar	Introduction to LAPACK	22
18-Mar	Introduction to Hybrid MPI and OpenMP	7
19-Mar	Introduction to Linux and Vi	17
25-Mar	Introduction to Open Source Visualization Software	13
15-Apr	PetaShare Environment and Client Tools	47
Spring 2009	Total	189
	Workshops	
March 3 & 4	LONI HPC Workshop at SUBR	>30
April 13 & 14	LONI HPC Workshop at Tulane	19
Spring 2009	Total	>49
Summary	**189 people were trained in Spring 2008 in 13 tutorials, and more than 49 people attended 2 workshops**	

LONI High Performance Computing Workshops covered an overview of HPC environment of IBM Power5 and Linux machines at LONI, basic AIX/Linux operating system commands and editors, introduction to Parallel Computing, introduction to MPI and advanced MPI, introduction to programming with OpenMP, and the LONI Portal.

1.V.c) Summer Camps involving High School Education

The Advanced Materials Research Institute (AMRI) conducts a summer research camp each summer, which includes REU students in addition to high school students and teachers. Last summer the program had 9 undergraduate/high school students and three high school teachers. The Outreach Summer Research Program for High School Students and Teachers receives support from the Louisiana Board of Regents (LBoR), the US Army Research Office (USARO), and the National Science Foundation (NSF), through LBoR Award # LEQSF(2007-12)-ENH-PKSFI-PRS-04, USARO Award # W911NF-04-1-0226, Academy of Applied Science Subgrant 07-25 and Subgrant 07-26, and NSF Award # CHE-0611902.

LA Tech is in the middle of the second year "CyberCamp" for high school students and teachers. The workshop was attended by 46 students and 18 teachers. The students are exposed to many hands-on activities such as programming a computer, programming a robot, a treasure hunt with cyber-related clues, and they also listen to experts talking about cybersecurity, history, psychology, and politics of cyberattacks, and learn about methods used to defend against such attacks.

2. PUBLICATIONS AND PRODUCTS

Here we list the publications, presentation and other tangible products by the personnel directly funded by the PKSFI, LONI Institute grant. We also include a copy of the publications already published in appendix E. These publications can be cross-institution, and some of them are at the interface between biology, materials, and computational science.

Publications:

Supada Laosooksathit, Chokchai Leangsuksum, Abdelkader Baggag, Clayton Chandler, "Stream Experiments: Toward Latency Hiding in GPGPU", submitted to HiPC09.

D. A. Genov, S. Zhang, and X. Zhang, "Mimicking celestial mechanics in metamaterials", accepted for publication to *Nature Physics* (May 2009).

D. A. Genov, A. K. Sarychev, and V. M. Shalaev, "Adiabatic Spatially Selective Photomodification of Inhomogeneous Metal-Dielectric Composites", submitted to *Physical Review Letters* (April 2009).

N.S.Vidhyadhiraja, A.Macridin, C.Sen, M.Jarrell, Michael Ma Quantum Critical Point at Finite Doping in the 2D Hubbard Model: A Dynamical Cluster Quantum Monte Carlo Study , arXiv:0809.1477. Physical Review Letters, in press.

E. Khatami, A. Macridin, M. Jarrell The validity of the spin-susceptibility "glue" approximation for pairing in a two-dimensional Hubbard model , arXiv:0901.4802.

Karlis Mikelsons, Alexandru Macridin, Mark Jarrell The relationship between Hirsch-Fye and weak coupling diagrammatic Quantum Monte Carlo methods , arXiv:0903.0559.

C. N. Varney, C.-R. Lee, Z. J. Bai, S. Chiesa, M. Jarrell, R. T. Scalettar High Precision Quantum Monte Carlo Study of the 2D Fermion Hubbard Model , arXiv:0903.2519.

E. Khatami, C. R. Lee, Z. J. Bai, R. T. Scalettar, M. Jarrell Dynamical Mean Field Theory Cluster Solver with Linear Scaling in Inverse Temperature , arXiv:0904.1239.

D.B. Khismatullin and G.A. Truskey, "Leukocyte Rolling on P-selectin: A 3D Numerical Study of the Effects of Cell Viscosity and PSGL-1 Clustering," *Ann. Biomed. Eng. (in revision)*

D. B. Khismatullin, "The cytoskeleton and deformability of white blood cells" in Klaus Ley (Ed.), "Current Topics in Membrane. Vol. 64. Leukocyte adhesion" (Elsevier, scheduled to be published in 2009).

David L. Mobley* and Ken A. Dill, "The binding of small-molecule ligands to proteins: 'What you see' is not always 'what you get', *Structure* **17**(4), 489-498 (2009), 10 pages. * - corresponding author.

D. L. Mobley[+], C. I. Bayly, M. D. Cooper, and K. A. Dill. "Predictions of hydration free energies from all-atom molecular dynamics simulations", invited article, *Journal of Physical Chemistry B* 113: 4533-4537 (2009), special issue on "Calculation of Aqueous Solvation Energies of Drug-Like Molecules: A Blind Challenge.

D. L. Mobley[+], C. I. Bayly, M. D. Cooper, M. R. Shirts, and K. A. Dill. "Small molecule hydration free energies in explicit solvent: An extensive test of fixed-charge force fields", *J. Chem. Theory Comput.* **5:** 350-358, 2009 (DOI 10.1021/ct800409d), 9 pages. One of the top 10 most downloaded articles in JCTC between March, 2008 and March, 2009.

Encode Project Consortium. Identification and Analysis of Functional Elements in 1% of the Human Genome by the Encode Pilot Project. *Nature.* 2007 Jun 14; 447(7146): 799-816.

Neerja Karnani, Christopher Taylor, Ankit Malhotra and Anindya Dutta. Pan-S Replication Patterns and Chromosomal Domains Defined by Genome-Tiling Arrays of Encode Genomic Areas. *Genome Research.* 2007 Jun; 17(6): 865-76.

Encode Project Consortium. The Encode (ENCyclopedia Of DNA Elements) Project, *Science.* 2004 Oct 22; 306(5696): 636-40.

Anindya Dutta, Neerja Karnani, Ankit Malhotra, Gabriel Robins and Christopher M. Taylor. Extraction of Human DNA Replication Patterns from Discrete Microarray Data. *Third IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2008)*, Novotel St Kilda, Melbourne Australia, October 2008.

D. S. Katz, G. Allen, R. Cortez, C. Cruz-Neira, R. Gottumukkala, Z. D. Greenwood, L. Guice, S. Jha, R. Kolluru, T. Kosar, L. Leger, H. Liu, C. McMahon, J. Nabrzyski, B. Rodriguez-Milla, E. Seidel, G. Speyrer, M. Stubblefield, B. Voss, and S. Whittenburg, "Louisiana: A Model for Advancing Regional e-Research through Cyberinfrastructure," Philosophical Transactions of the Royal Society A, v. 367, pp. 2459-2469, 2009.

Gottumukkala, N. R., R. Nassar, C.B. Leangsuksun, M. Paun. "Reliability of a system of k nodes for high performance computing applications". To appear in the December 2009 issue of the The IEEE Transactions on Reliability.

Shizhong Yang, S.M. Guo, Guang-Lin Zhao, and Ebrahim Khosravi, " High infrared reflective nickel doped ZrO_2 from first principles simulation", ICCS May 2009. (International conference paper)

Shizhong Yang, et al. (peer-reviewed and was accepted to be published): "Doped C_{60} study from first principles simulation", New3SC- 7, (Seventh International Conference on New Theories, Discoveries and Applications of Superconductors and Related Materials), Beijing, May 2009. Also an invited talk.

Wendong Wang, Zhenjun Wang, Jinke Tang, Shizhong Yang, Hua Jin, Guang-Lin Zhao, and Qiang Li, " Seebeck coefficient and thermal conductivity in doped C_{60}", Journal of Renewable and Sustainable Energy, vol. **1**, issue 2, 23104, page1~8 (2009)

G.L. Zhao, S. Yang, D. Bagayoko, J. Tang and Z.J. Wang, **"** Electronic structure of C60 semiconductors under controlled doping with B, N, and Co atoms", Diamond & Related Materials, vol. **17,** page749~752 (2008).

Zaixin Lu, Zhiyu Zhao, Sergio Garcia, Krishnakumar Krishnaswamy, and Bin Fu, "Search Similar Protein Structures with Classification, Sequence and 3-D Alignments", to appear in the Journal of Bioinformatics and Computational Biology.

Huimin Chen and Zhiyu Zhao, "An Information Theoretic Viewpoint on Haplotype Reconstruction from SNP Fragments", to appear in the 3rd International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2009, China).

Zaixin Lu, Zhiyu Zhao, Sergio Garcia, and Bin Fu, "New algorithm and web server for finding proteins with similar 3d structures", in the Proceedings of the 2008 International Conference on Bioinformatics & Computational Biology (BIOCOMP'08, USA), pp. 674 - 680.

Eren AM, Amin I, Alba A, Morales E, Stoyanov A, and Winters-Hilt S. Pattern Recognition Informed Feedback for Nanopore Detector Cheminformatics. Submitted to BMC Biotechnology.

Eren AM & Stephen Winters-Hilt. A Visualization Tool for Nanopore Experiments. Submitted to MCBIOS Proceedings for BMC Bioinformatics.

Winters-Hilt S, Eren AM, and Armond Jr. K. Distributed SVM Learning and Support Vector Reduction. Re-submission planned to BMC Bioinformatics.

Winters-Hilt S, Eren AM, and Merat S. Unsupervised clustering using supervised support vector machines. Re-submission planned BMC Bioinformatics.

Winters-Hilt S and Eren AM. SVM-based clustering with kernel-clustering for kernel-tuning and seed cluster-region identifications.

Winters-Hilt S and Jiang Z. An Efficient Self-Tuning Explicit and Adaptive HMM with Duration Algorithm. Accepted by IEEE Transactions on Signal Processing, June 2009. (http://www.cs.uno.edu/~winters/ESTEAHMMD_preprint.pdf)

J. Jack and A. Paun, ``Discrete Modeling of Biochemical Signaling with Memory Enhancement," LNBI Transactions on Computational Systems Biology 2009, 14 pp. [accepted].

J. Jack, A. Paun, Simulation of Signaling Pathways through discrete methods, JALC, accepted 2009.

J. Jack, A. Paun, F A. Rodriguez-Paton, Discrete nondeterministic modeling of the FAS pathway, Int. Journal of Foundations of Computer Science, vol. 19 (October 2008), no. 5, pp. 1147-1162.

J. Jack, A. Paun, A. Rodriguez-Paton, Effects of HIV-1 Proteins on the Fas-Mediated Apoptotic Signaling Cascade: A Computational Study of Latent CD4+ T Cell Activation, accepted at Ninth Workshop on Molecular Computation, WMC9, Edinburgh (UK) July 28-31, 2008, 20pp.

J. Lao and D. Moldovan, “Surface stress induced structural transformations and pseudoelastic effects in palladium nanowires” Appl. Phys. Lett. 93, 093108, 2008

J. Lao and D. Moldovan, “Interfacial strain induced self-rolling of Aluminum nanotubes” In preparation, plan to submit for publication to Physical Review Letters.

Juliette W. Ioup, George E. Ioup, Lisa A. Pflag, Arslan M. Tashmukhambetov, Christopher O. Tiemann, Alan Berstein, Natalia Sidorovskaia, Philip Schnenayder et al., “Localization to verify the identification of individual sperm whales using click properties,” The Journal of the Acoustical Society of America, 125(4, pt.2 of 2), April 2009, p. 2616 (published abstract)

Natalia Sidorovskaia, Philip Schexnayder, et al., “Rhythmic analysis of sperm whale broadband acoustic signals,” The Journal of the Acoustical Society of America, 125(4, pt.2 of 2), April 2009, p. 2738 (published abstract)

S. Chu, J. Chen, Z. Wu, V. Raghavan, H. Chu. “A Treemap-based Result Interface for Search Engine Users”, 12th International Conference on Human-Computer, Interaction (HCI 2007), Volume 8, July 2007.

Presentations:

Abdelkader Baggag, LONI Institute All-Hands Meeting: “High Performance Computing and Computational Science and Engineering” with an overview of the particulate flow application.

Abdelkader Baggag, IBM Watson Research Center: “A scalable nested iterative scheme for linear systems in particulate flows”

Abdelkader Baggag, Laval University: Participation in HPC committee for NSERC evaluation of Industrial Research Chair

Dentcho Genov, “*Electromagnetic metamaterials: from imaging with super resolution to mimicking celestial phenomenon in the lab*”, The LONI Institute (LI) All-Hands Meeting, Baton Rouge, LA, October 31, 2008.

Dentcho Genov, *Electromagnetic properties of complex metamaterials: from near field imaging with super resolution to mimicking celestial phenomenon in laboratory conditions*”, Colloquium Series at the Center for Computational &Technology (CCT), Louisiana State University, Baton Rouge, LA, March 27, 2009.

Mark Jarrell, *The Phase Diagram of the Two-Dimensional Hubbard Model: A Quantum Critical Point at Finite Doping*, Invited Talk, Oct. 8, 2008, 3rd International Workshop on "Ordering Phenomena in Transition Metal Oxides" Augsburg, Germany, October 5-8, 2008

Mark Jarrell, *Bond Excitations in the Pseudogap Region of the Hubbard Model*, Invited Talk, Oct. 28, 2008, 21st International Symposium on Superconductivity, International Congress Center, Tsukuba Japan, October 27-29, 2008.

Mark Jarrell, *Massively Parallel and Multi-Scale Simulations of Strongly Correlated Electronic Systems.* , Invited Talk, March 4, 2009, Michael Dewar Memorial Symposium: Advancing Computational Chemistry Through High Performance Computing, from the Workstation to the Petascale and Beyond. March Meeting of the American Chemical Society. Salt Lake City, Utah.

Mark Jarrell, *Massively Parallel and Multi-Scale Simulations of Strongly Correlated Electronic Systems*, Keynote Lecture, The International Conference on Computational Science 2009 May 25 - 27, 2009, in Baton Rouge, Louisiana

D.B. Khismatullin, "Viscoelastic Volume-of-Fluid algorithm for multiphase flow problems", LONI HPC Workshop, Tulane University, April 13-14, 2009 — New Orleans, Louisiana.

D.B. Khismatullin, "Application of the Volume-of-Fluid algorithm to biological systems", 2009 Spring Southeastern Meeting of the American Mathematical Society, April 4-5, 2009 — Raleigh, North Carolina.

D.B. Khismatullin, "Modeling of cell adhesion using a multiphase flow approach", LONI Institute First All-Hands Meeting, Louisiana State University, October 31, 2008 — Baton Rouge, Louisiana.

D.B. Khismatullin, Tulane University School of Medicine, Department of Physiology (May 18, 2009). *Quantitative analysis of leukocyte-endothelial cell interactions in inflammation and atherosclerosis.* Host: *Dewan Majid*. Invited Talk.

D.B. Khismatullin, Tulane University, Department of Chemical and Biomolecular Engineering (April 24, 2009). *Computational modeling of receptor-mediated leukocyte adhesion to surfaces.* Host: *Noshir Pesika*. Invited Talk.

D.B. Khismatullin, Tulane University, Applied and Computational Mathematics Seminar (January 23, 2009). *Biological systems modeling using a multiphase flow approach.* Host: *Ricardo Cortez*. Invited Talk.

D.B. Khismatullin, Southern Methodist University, Department of Mathematics (October 15, 2008). *A multiphase flow approach to modeling biological systems.* Host: *Vladimir Ajaev.* Invited Talk.

David Mobley, "Lessons learned from predicting binding free energies in model binding sites" and "Quantitative predictions of protein-ligand binding affinities", American Chemical Society Meeting, Salt Lake City, UT, March 2009, contributed presentation.

David Mobley, "Predictive calculations of absolute binding free energies", American Chemical Society Meeting, August 20, 2008, Philadelphia, PA, invited presentation.

Anindya Dutta, Neerja Karnani, Ankit Malhotra, Gabriel Robins and Christopher M. Taylor. Extraction of Human DNA Replication Patterns from Discrete Microarray Data. *Third IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2008)*, Novotel St Kilda, Melbourne Austrailia, October 2008.

Christopher M. Taylor. Extraction of Human DNA Replication Timing Patterns from Discrete Microarray Data. *LONI All-Hands Meeting.* Baton Rouge, LA, October 2008. Invited Talk.

N. Raju Gottumukkala, Box Leangsuksun, Raja Nassar, Mihaela Paun, Dileep Sule, "Reliability Aware Optimal-K Node allocation of parallel applications in large scale HPC systems", High Availability and Performance Computing Workshop (HAPCW 2008), Denver, Colorado.

Raju Gottumukkala, Ramesh Kolluru, "Improving Disaster Response: NIMSAT", The 2009 gulf Coast Marine Conference.

Raju Gottumukkala, Rusti Liner, "GIS Projects at NIMSAT Institute" The 25th Annual Remote Sensing and GIS Workshop, April 14-16 2009, Louisiana.

Presentation at 2009 LAS 83rd annual conference: "First principles molecular dynamics simulation of nano gold adsorption on (0001) surface of Ruthenium", Shizhong Yang, Shuju Bai, Ebrahim Khosravi, and Guang-Lin Zhao.

Invited talk: "Doped C60 study from first principles simulation", New3SC- 7, (Seventh International Conference on New Theories, Discoveries and Applications of Superconductors and Related Materials), Beijing, 2009.

Zhiyu Zhao (author & presenter), "Intermediate MATLABhttp://www.hpc.lsu.edu/training/tutorials/presentations/Intro-MATLAB-0309.pdf", Stanley Thomas Hall, Tulane University; An invited tutorial session of the LONI HPC Workshop, Spring 09, hosted by the Tulane University and open to all the LI research community, see http://www.hpc.lsu.edu//training/20090413/index.php. (04/14/09)

Zhiyu Zhao (author & presenter), "Introduction to LAPACK", Liberal Arts Building, UNO; A tutorial session of the LONI HPC Training, Spring 09, open to all the LI research community via UNO's Access Grid facilities, see http://www.hpc.lsu.edu/training/tutorials/index.php#spring09lapack. (03/16/09)

Zhiyu Zhao (author & presenter), "Introduction to MATLAB", Liberal Arts Building, UNO; A tutorial session of the LONI HPC Training, Spring 09, open to all the LI research community via UNO's Access Grid facilities, see http://www.hpc.lsu.edu/training/tutorials/index.php#spring09matlab. (03/02/09)

Zhiyu Zhao (author & presenter), "Introduction to the Supercomputing Resources at LONI & TeraGrid", Math Building, UNO; A presentation open to all the UNO research community as required by the chair of the Department of Computer Science, see http://www.cs.uno.edu/special/seminars.xml#Introduction%20to%20Supercomputing%20Resources%20at%20LONI%20and%20TeraGrid and http://www.cs.uno.edu/~sylvia/LONI&TeraGrid.pdf. (02/06/09)

Zhiyu Zhao (author & presenter), "Protein 3D Structure Alignment and Searching for Similar Structures in the Protein Data Bank", Engineering Building, UNO; A seminar talk invited by the Department of Electrical Engineering, UNO and open to all the EE faculty/staff and students, see http://www.cs.uno.edu/~sylvia/ProteinStructure.pdf. (01/29/09)

Zhiyu Zhao (author & presenter), "Feedback Algorithm and Web-Server for Protein Structure Alignment", CERM Building, UNO; A seminar talk invited by the Department of Computer Science and open to all the CS faculty/staff and students, see http://www.cs.uno.edu/special/seminars.xml#Feedback%20Algorithm%20and%20Web-Server%20for%20Protein%20Structure%20Alignment and http://www.cs.uno.edu/~sylvia/SLIPSA.pdf. (11/21/08)

Scott Whittenburg (author) and Zhiyu Zhao (author & presenter), "Computational Research at UNO", LSU Union, LSU; A presentation required by Dr. Whittenburg (vice chancellor of research at UNO) and open to all attendees of the LI All Hands Meeting '08, see http://institute.loni.org/FirstAllHandsMeeting.php. (10/31/2008)

Zhiyu Zhao (author & presenter), "Research on Protein 3-D Structure and Genome Sequence Related Problems", Liberal Arts Building, UNO; A talk invited by the Director of the University Honors Program and open to all the UNO honors students of fall 08, see http://www.cs.uno.edu/~sylvia/Protein&Genome.pdf. (10/28/08)

Zhiyu Zhao (author & presenter), "Linear Time Probabilistic Algorithms for the Singular Haplotype Reconstruction Problem from SNP Fragments", Engineering Building, UNO; A seminar talk invited by the Department of Electrical Engineering, UNO and open to all the EE faculty/staff and students, see http://www.cs.uno.edu/~sylvia/HapRec.pdf. (10/02/08)

A. Murat Eren & Stephen Winters-Hilt. Pattern recognition-informed sampling for nanopore biosensing. MidSouth Computational Biology and Bioinformatics Society (MCBIOS), Starkville, MS, Feb. 20-21, 2009.

A.Murat Eren & Stephen Winters-Hilt. A Visualization Tool for Nanopore Experiments. MidSouth Computational Biology and Bioinformatics Society (MCBIOS), Starkville, MS, Feb. 20-21, 2009.

Joshua Morrison, A. Murat Eren, and Stephen Winters-Hilt. Machine Learning Web Interfaces for Bioinformatics & Cheminformatics. MidSouth Computational Biology and Bioinformatics Society (MCBIOS), Starkville, MS, Feb. 20-21, 2009.

Amanda Alba, Eric Morales, A. Murat Eren, Joshua Morrison and Stephen Winters-Hilt. Nanopore-transduction based study of individual molecular binding events. MidSouth Computational Biology and Bioinformatics Society (MCBIOS), Starkville, MS, Feb. 20-21, 2009

John Jack gave a presentation at the EPA's National Center for Computational Toxicology (NCCT) on the research involving the Nondeterministic Waiting Time (NWT) algorithm. The talk was given in May 2009 and was one hour in length.

Dr. Andrei Paun (John Jack's dissertation advisor) is presenting various aspects of their research at "Descriptional Complexity of Formal Systems" in Magdeburg, Germany. He is an invited speaker at the conference which takes place July 6th - 9th.

J. Lao and D. Moldovan, "Molecular dynamics simulation study of pseudoelastic effects in palladium nanowires" The Fourth International Conference on Multiscale Materials Modeling, Tallahassee, Florida, October 27-31, 2008.

Philip Schnenayder, Physics Department seminar, April 2009.

Philip Schnenayder, Oral presentation at the 157th meeting of the Acoustical Society of America, Portland, Oregon, May 22 2009: "Rhythmic analysis of sperm whale broadband acoustic signals".

Invited Book Chapters:

Neerja Karnani, Christopher M. Taylor and Anindya Dutta. Microarray Analysis of DNA Replication Timing. Microarray Analysis of the Physical Genome. *Methods in Molecular Biology.* Vol 556, ISBN: 978-1-60327-191-2, Humana Press. June 16, 2009.

Book:

Abdelkader Baggag, "Parallel Numerical Algorithms and Applications", Book in progress.

Posters:

Christopher Taylor, Neerja Karnani, and Anindya Dutta. Analyzing DNA Replication Timing in the Human Genome. *Experimental Biology.* Ernest N. Morial Convention Center, New Orleans, LA, April 2009.

Patent or licensing agreements:

Shixian Chu, Jinfeng Chen, Zonghuan Wu, Chee-Hung Henry Chu, Vijay Raghavan, "Method and Apparatus for Information Visualized Expression and Visualized Human Computer Interactive Expression Thereof", PCT/CN2008/000168

3. CONTRIBUTIONS

The LI is still in the formative stages as we hire personnel. We do not expect its faculty to be fully in place (and therefore its contributions to the state research and education capacity will not be fully felt) until Y3 and beyond. However, at this stage the LI already has made important contributions:

3.I. Contributions to the state research and education capacity.

Through a collaborative search procedure we have hired nine faculty, seven of which started in Y2, and two which will start on Y3 (with 3 more to be hired in Y3). These faculty are outstanding, and have been attracted to the state through activities on LONI and through LI member cooperation. All are highly computationally oriented, and some are international leaders in their fields. During the recruitment phase, members of the LI sites discussed candidates, coordinated recruitment across the universities, and even participated in interviews. The result is that these new faculty are already in discussions with various LI members about collaborative, multi-institutional research projects that utilize LONI, setting the stage for larger collaborative funding projects in the future. These faculty will now be effective in recruiting the second round of faculty in Y3.

Through recruitment of the four computational scientists (with two more still to be recruited), the existing faculty have come to know each other and the interests of each site. These scientists are already involved in the advanced projects across the state (LI Projects). This contributes directly to the research and educational capacity of the state, by making the projects our faculty carry out more advanced, and grant proposals we write more competitive. These staff have also helped train the university base on cyberinfrastructure (CI) we develop, making them more competitive as well. Finally, these staff will be critical to advances in corporate partnerships as they are developed, as they will be able to support research projects by companies that are carried out, using LONI and other CI we develop and deploy, in collaboration with our LI university partners.

The LI faculty have initiated and led a new EPSCOR grant proposal on computational material science. This proposal, led by Mark Jarrell, involves all the LI sites, and Xavier University. The proposal was recently submitted to the LA Board of Regents, and is expected to also build upon the current NSF EPSCOR "CyberTools" project.

Finally, in supporting our six graduate students each year, for work on LONI, we are clearly supporting the state research and education capacity. Starting Y3, there will be 12 graduate students funded, that is, 2 students per LI site.

3.II. Securing external federal and private sector funding.

As reported above, the pre-existing (PIs and SIs) and newly hired LI faculty and staff have received external funding for numerous projects during the second year. Collaborations

developed will be strengthen and enhanced in the future years of the LI gains. The $12M CyberTools project is expected to provide advanced CI for the entire state, and the $2.2M NSF HPCOPS project brings the LONI infrastructure to the TeraGrid. The latter award funds 12 staff positions to better develop and support the LONI environment, while integrating this national environment with the state.

3.III. Infrastructure.

Two projects are providing advanced Cyber-Infrastructure to the state upon which other projects can build. The most important of these projects from the point of view of infrastructure for the state is the CyberTools project, which we expect will provide advanced CI for the entire state, not only the LI members. The NSF HPCOPS project connects the LONI infrastructure to the TeraGrid, which is the NSF's national backbone for advanced CI. This award also funds staff positions which support the LONI environment as it integrates this national environment with the state. This makes it much easier for all state researchers to take advantage of the national CI. Both HPCOPS and CyberTools awards are providing advanced CI for the state to build on.

3.IV. Economic Development.

As described above, corporate partnerships are being explored. For example, deep discussions have been held with Schlumberger for a series of pilot projects with staff and faculty at LSU, which if funded would make use of the projects funded by CyberTools and CCT. Tulane and SUBR have partnerships with IBM, and ULL has a partnership with Louisiana GOHSEP. These are examples of the kinds of economic development partnerships that we expect to develop in the future once the LI is fully developed.

4. PROJECT REVISIONS.

We do not at this time expect any revisions to the project deliverables. As additional faculty and staff are recruited, we expect to continue on track to deliver what was originally promised. However, the Project has undergone important changes in leadership. The original PI (Seidel) accepted the position of Director of the Office of Cyberinfrastructure at NSF; Seidel was succeeded by Jarek Nabrzyski, who in turn has moved to Notre Dame as the Director of the Center for Research Computing. Prof. Mark Jarrell has now taken on PI's responsibility.

Appendix A

LI Faculty White Papers

White Paper for Abdelkader Baggag, Associate Professor of Computer Science, LA Tech University

LI FACULTY (September 2008 – present)

Abdelkader Baggag, PhD

Associate Professor of Computer Science
Louisiana Tech University, Department of Computer Science
600 West Arizona, Nethken 149, Ruston LA 71272
Phone: (318) 257-2147, Fax: (318) 257-4922
Email: abaggag@latech.edu

EXPERTISE

Parallel numerical algorithms for large scale engineering applications and their efficient implementation on massively parallel computers.

RESEARCH INTERESTS

With academic background in computer science, applied mathematics, and engineering, and exposure to industrial and US national laboratory applications, such as NASA Langley Research Center (LaRC), I work at the interface between the numerical solution of partial differential equations, science and engineering applications, and parallel computing. I have experience, gained in particular by interacting with NASA scientists, in the numerical simulation of many disciplines such as (parallel) computational aero-acoustics (CAA), and (parallel) computational fluid dynamics (CFD): modeling the physical processes, designing and analyzing iterative algorithms, and implementing them on advanced massively parallel computers, using the message passing paradigm.

I have developed a comprehensive curriculum in HPC, and my goal is to use my prior expertise in curriculum development to set up a similar set of HPC courses within the LONI Institute.

RESEARCH GROUP MEMBERS

1. Chokchai (Box) Leangsuksun, Associate Professor, Computer Science
2. James Elliott, PhD Student, Computer Science
3. Supada Laosooksathit, PhD Student, Computer Science
4. Timothy Lindsay, Master's Student, Computer Science
5. Moayad Almohaishi, Master's Student, Computer Science

ASSOCIATE GROUP MEMBER

1. Dick Greenwood, Associate Professor, Physics

RESEARCH PROJECTS

1. Parallel Algebraic Multigrid Methods on Distributed Memory Computers
2. GPGPUs: Toward Latency Hiding
3. Porting the particulate flows code, and changing it to handle new physics, namely the simulation of the blood
4. General Purpose Graphics Processor Units and High Energy Physics (with Greenwood and Leangsuksun)

RESEARCH PROJECT 1:

PARALLEL ALGEBRAIC MULTIGRID METHODS ON DISTRIBUTED MEMORY COMPUTERS FOR LARGE-SCALE INDUSTRIAL APPLICATIONS

Massively parallel computer systems, incorporating thousands of processors, are becoming a reality in Louisiana, through the LONI institute. Thus, one can hope that we are now able to test more detailed mathematical models with millions of degrees of freedom. Massively parallel computing is a major way to analyze many challenging problems in engineering, and offers a more flexible and scalable approach than experimentation. This, however, places a heavy burden on the design of suitable algorithms, their efficient implementation on parallel architectures and the development and maintenance of very large software.

The fast and efficient solution of linear systems of equations is a key task in the process of many industrial problems, as it represents more than 90% of the computational time. It is therefore necessary to develop algorithms of optimal complexity where both memory and compute time should depend only linearly on the problem size. Moreover, the increasing demands of computationally challenging applications on complex geometries, which run on high-performance computers with tens or hundreds of thousands of processors, have necessitated the development of scalable solvers and preconditioners.

One of the most effective ways to achieve scalability is the use of multilevel techniques. Algebraic multigrid (AMG) methods have proved to be efficient algorithms for solving large linear systems on unstructured grids. By the term “algebraic multigrid,” we mean the class of solvers based on multigrid principles, but which do not depend on the availability of the underlying mesh about the problem. AMG methods use only the information available in the linear system of equations, and therefore are of special interest especially for unstructured grids where no hierarchy of meshes is provided, e.g. in finite element discretizations.

The first aim of the proposed project is to efficiently implement the (sequential) Ruge-Stüben classical AMG algorithm, and extensively test it on real life engineering applications. The second aim is to investigate new parallel approaches suitable for distributed memory computers, and implement a general parallel algebraic multigrid algorithm for finite element discretizations based on domain decomposition ideas. In particular, different parallel strategies for the coarsening phase and the smoothing operator will be examined. As for the robustness of the algorithm, an adaptive process will be utilized.

While the AMG algorithm works well for a wide range of problems, there are situations in which it requires special care, e.g., on stretched grids where obtaining a robust and efficient solution still remains problematic. One reason for this is that, sometimes, the interpolation

operator does not interpolate adequately the smooth modes of the error. Hence, a characterization of the smooth error will be provided, analyzed and implemented along the lines of element-based algebraic multigrids (AMGe), and element-free AMGe. Lastly, a careful analysis and (implementation) of the compatible relaxation algorithm, will be undertaken, as it holds much promise for parallel computation.

The project will provide significant training of highly qualified personnel at the Masters and PhD levels. The outcome of the current project will be the development of efficient and robust parallel algebraic multigrid algorithms which scale both with machine and problem size, and an object-oriented software that is portable, modular and MPI-based.

RESEARCH PROJECT 2:

GPGPUS: TOWARD LATENCY HIDING

Abstract: In multithreaded programming on GPUs, data transfer between CPU and GPUs is a major impendence that prevents GPU to achieve its potential. Hence, stream management framework – a latency hiding strategy introduced by CUDA, becomes our attention. Streaming allows overlapping between kernel execution time and transfer time of independent data between CPU and GPUs. For this reason, the total execution time can potentially be reduced. In this project, we introduce performance models in order to study the utilization of streams. Moreover, we study two methods that are used for timing operations in CUDA, namely CUDA calls and CUDA events. CUDA call functions are functions implemented in C++, while CUDA events method is an API. Our finding shows that CUDA events method is more accurate for timing operations than CUDA call functions

Description: Due to the rapid increase in computational performance required to handle massive data sets, Graphic Processing Units (GPUs), which were first designed for specific graphics computation, have been used for non-graphics applications. This process is known as GPGPU, or General Purpose Computation using GPUs. GPUs are specialized for compute intensive, highly parallel computation via several multithreaded processors. As such, GPUs are commonly classified as a collection of many-core Central Processing Units (CPUs). Today, GPUs have been exploited in many application areas including oil exploration, scientific data processing, and stock options pricing determination. The operations driving these applications require efficient data management and the processing of massive data sets, such that these operations can be simultaneously performed in an extremely fast fashion. However, data transfer from host to GPU devices and back slows down the application executions. To fulfill these requirements, a latency hiding strategy is needed. Due to its coprocessor status, the kernel grid aboard the GPU is invoked by a program running on CPU, which in turn distributes thread blocks of the grid to the multiprocessors. Each thread of a thread block concurrently executes its instructions on one multiprocessor. New blocks are launched on the multiprocessors again as previous thread blocks terminate. Even though the cost of each memory copy between the CPU (host) and GPU (device) is relatively very low, the millions of bytes of data and thousands of transfers of each chunk of this data required by applications can multiply those costs and lead to exceptionally high values. To reduce these costs, the GPU allows for the applications to be managed concurrency through streams. Unlike stream processing – a technique to operate on streams or sequences of data, streaming is a latency hiding strategy which allows for sequences of operations to execute

successively. Different streams execute particular operations in parallel. With streams, the kernel launches the current operation and copies the memory chunk of the next operation asynchronously. In this paper we focus on memory transfer between main memory which works with the CPU and the global memory of the GPU. The internal memory transfer of the GPU (between global memory and shared memory which works directly with threads) will be the scope of another project. Besides the streaming technique, we also study two methods for timing operations in CUDA, namely CUDA call functions and CUDA events. CUDA call functions are implemented in C++ and provided in CUDA SDK. CUDA events method is an API and has to be implemented with stream.

BOOK IN PROGRESS: "Parallel Numerical Algorithms and Applications"

COURSE DEVELOPMENT:

1. **"High Performance Computing and large Scale Numerical Modeling"**

In this course, I will be teaching the essence of what is needed for researchers to take advantage of the machines' power, and I will spend an extra effort to hold the researchers' hand in parallelizing, debugging and optimizing some of their respective codes. The scaling of parallel algorithms has not yet matched peak speed, and the programming burden for parallel machines remains heavy.

Hence the applications must be programmed to exploit parallelism in the most efficient way possible. Today, the responsibility for achieving the vision of scalable parallelism remains in the hands of the application developers. This course illustrates the state-of-the-art of parallel computing, and links theory to applications, through demonstrations and training.

This course should be of interest to engineers, programmers and code developers. As a first step, I will include it as part of the CAM program, and as a follow up step, I will deliver it to LONI scientists.

2. **"Multi-institution, interdisciplinary courses for Computational Material Science"**

By Juana Moreno, Karen Tomko, Mark Jarrell, and Abdelkader Baggag

This effort has not been funded and should be re-submitted.

PRESENTATIONS

1. LONI Institute All-Hands Meeting: "High Performance Computing and Computational Science and Engineering" with an overview of the particulate flow application
2. IBM Watson Research Center: "A scalable nested iterative scheme for linear systems in particulate flows"
3. Laval University: Participation in HPC committee for NSERC evaluation of Industrial Research Chair

PEER REVIEWED PUBLICATION

“Stream Experiments: Toward Latency Hiding in GPGPU” by Supada Laosooksathit, Chokchai (Box) Leangsuksun, Abdelkader Baggag, Clayton Chandler

DIRECTED STUDENT LEARNING

1. Master's Thesis Committee Member, "Wireless Sensor Network Protocols," CSC. (February 26, 2009). Advised: Mohamed Faisal Baig
2. PhD Dissertation External Committee Member, "Modélisation du changement de phase de la cryolite dans une cuve Hall- Héroult," CSC. (October 10, 2008). Advised: Edith Laliberte

TEACHING AT LOUISIANA TECH UNIVERSITY (LECTURER AND INSTRUCTOR)

ENGR 501 “Research Methods
CSC 581 “Parallel Algorithms”

Electromagnetic Metamaterials and Nanophotonics

LI FACULTY (09.2008 - present)

Dentcho A. Genov, PhD

Assistant Professor of Physics & Electrical Engineering
Louisiana Tech University, Engineering Annex, Room 220
599 W Arizona Ave, Ruston LA 71272
Phone: (318) 257-4190, Fax: (318) 257-2777
Webpage: http://www.phys.latech.edu/~dgenov/

RESERCH INTERESTS

- Electromagnetic properties of nano-structured complex media including: metal composite materials, rough surfaces, fractal aggregates, and ordered media
- Solid state and condensed matter physics: geometrical phase transitions, scaling theory, classical and quantum wave localization phenomena
- Nanophotonics and quantum optics, nonlinear optics and spectroscopy, quantum dots, nanoscopic lasers and optical elements, light scattering from metal particles
- Artificial materials: metamaterials and negative index media, electric and magnetic plasmon waveguides, plasmonic and ordinary band gap materials
- Numerical code development and algorithm optimization, large-scale computer simulations in electrodynamics, plasma physics, and material science

GROUP MEMBERS

Venkatesh Kumaran, PhD student (theory/computation)
Pattabhiraju Mundru, PhD student (theory/computation)
Shravan Rakesh, PhD student (theory/computation)
Rajivalochan Subramanian, PhD student (theory and experiment)

Projects Description: The main focus of our research is a rapidly developing field of artificial optical materials, referred to as electromagnetic metamaterials (EMMs). The phenomenal progress in nanofabrication now provides the enabling technology to develop EMMs with unlimited range of optical properties opening the possibility to manipulate light at will. This is accomplished by precise engineering of the microscopic magnetic and electric response of the media and is equivalent to virtually creating *new* types of quasi-atoms and quasi-molecules. As a result EMMs have been proposed to create a negative refraction index media, invisibility devices and lenses with super resolution. The LI faculty (Genov) has substantially contributed to this field publishing more than 30 papers in top peer reviewed scientific journals, including; *Nature*, *Nature Photonics*, *Nature Physics*, *Physical Review Letters*, *IEEE*, *Nano Letters*, est.. Currently our group is pursuing the following five related research projects:

a. Computational engineering of EMMs for optical invisibility (a PhD student involved)
b. Surface Plasma Enhanced Solar Cells (SPESC) (a PhD students involved).
c. Surface Plasmon based 100THz transistor (a PhD students involved).
d. Reversal of Casimir force in Metamaterials: (a PhD student involved).
e. The optical-mechanical analogy and it broad impact.

a. Computational engineering of EMMs for optical invisibility: Cloaking is an advanced stealth technology that utilizes EMMs to render objects invisible from arbitrary electromagnetic fields. The most popular methods for achieving invisibility or cloaking are based on encapsulating the object in EMM cloaking shells which guides the impinging waves away from the object rendering it invisible.[1,2] Although the proposed methods promise to provide substantial invisibility, they all suffer from substantial energy dissipation which makes true invisibility virtually impossible. The goal of our current efforts is to study prospective designs for low loss metamaterials to achieve high levels of electromagnetic invisibility both at the macroscopic and microscopic scales. To achieve this we rely on a mathematical technique called transformational optics (see Fig. 1) that allows the determination of the EMMs that provide a set of desired light paths. Specifically, we study a class of conformal maps that lead to new EMMs that may realize clocking of an object without involvement of magnetism, and concurrently under lower dissipation. These studies will also aid in developing new mathematical and numerical tools for treating electromagnetic interaction with metamaterials both in isotropic and anisotropic regime. The investigation of the local material response of strongly interactive optical elements requires utilization of parallel computational techniques such as finite difference frequency domain (FDFD) and consequently a high performance computing (HPC), which is provided by LONI. The developed numerical tools are also expected to contribute to ongoing projects such as Cyber Tools.

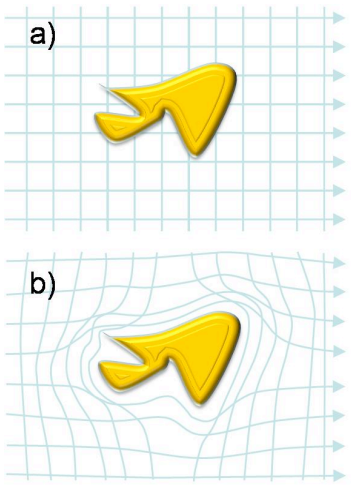


Figure 0 In contrast to normal free space (**a**), in EMMs the light rays curve around the object making (**b**), it invisible.

b. Surface Plasma Enhanced Solar Cells (SPESC): Solar radiation provides a source of free energy, which if efficiently taped could solve the most urgent problem facing the industrialized world, namely its reliance on fossil fuels for generation of electricity. The principal objective of this project is to develop a new approach toward inexpensive and highly efficient solar cells based on nano-engineered media. Specifically, a new photo photovoltaic cell is proposed that merges current technology with an Active Plasmonics Composite (APC) (Fig. 2a) to achieve enhanced performance. In recent works, we showed that in the optical and near-infrared frequency ranges the radiation reservoir associated with

the APC presents drastic departures from any conventional media, resulting in new phenomenon such as enhancement of the spontaneous emission, strong localization of light, and dramatic enhancement of nonlinear optical processes (Fig. 2b). In this project the enhancement of the SP density of states are utilized to engineer the photovoltaic properties of the SPESC. Substantial improvement of the current yield and conversion efficiencies (by a factor of two) are expected with the enhanced performance sustained over a broad frequency range. The SPESC design and optimization involves calculation of random systems of up to 1 million strongly interacting particles and required 1-10Tbits of shared memory. Thus the utilization of the HPC provided by LONI is crucial. Furthermore, to experimentally test the theoretical predictions collaboration has been initiated with Dr. S. Selmic (LA Tech) that will result in the creation of a SPESC prototype.

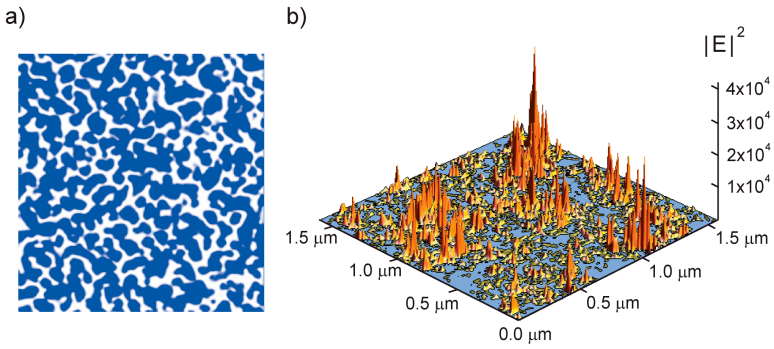


Figure 2. Metal-dielectric composite thin films (**a**) support morphology dependent surface plasmon (SP) resonances characterized by (**b**) highly enhanced local field densities over wide [3]

c. Surface Plasmon based 100THz transistor: In this project we use surface electromagnetic modes propagating between the interfaces of metal/semiconductors and air to develop an all-optical transistor. Specifically, we rely on a highly doped hetero-junction in connection with optical waveguides to excite and propagate confined SP modes. Two distinct mechanisms, one based on temperature switching and another on charge depletion, will be tested to simulate an effective transistor "I-V" response. Prospective designs will be identified and a prototype will be tested in the Institute for Micromanufacturing (LA Tech).

d. Reversal of Casimir force in Metamaterials: The Casimir force is a manifestation of a unique phenomenon due to existence of an "infinite ocean" of quantum electromagnetic vacuum fluctuations. For ordinary materials this force is always positive. However, with the invention of the EMMs it may be possible to reverse it sign from attraction to repulsion. In this project we will study numerically and analytically the conditions on the EMMs for such reversal to take place. Apart from the fundamental ramification of this project possible applications are envisioned for development of new thin film coatings to address contamination issues in clean rooms, thus lowering the coast of operations and increasing microprocessor chip production efficiency.

e. The optical-mechanical analogy and it broad impact: The optical-mechanical analogy recently demonstrated by Genov *et al.* provides a useful link between the study of light propagation in inhomogeneous media and the motion of massive bodies or light in gravitational potentials[4]. Specifically, we have shown that it possible to directly map in metamaterials the light interaction around gravitational black hole (Fig. 3a) or development novel Photon Traps (CIPTs) (Fig. 3b) as a direct manifestation of a planetary motion, but for light. Our immediate research efforts are focused at improving the existing EMMs designs

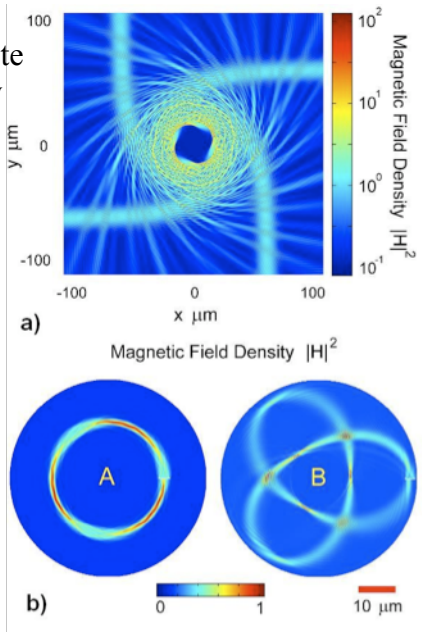


Figure 3 a) Light scattering in curved space-time (around a black hole) and b) closed photon trajectories in CIPT metamaterials[4]

and establishing collaborations that will allow the experimental validation of the discovered phenomenon.

References:

1. J. B. Pendry, D. Schurig and D. R. Smith, "Controlling Electromagnetic Fields", *Science* **312**, 1780 (2006).
2. J. Valentine, S. Zhang, T. Zentgraf, E. Ulin-Avila, D. A Genov, G. Bartal and X. Zhang, "Three Dimensional Optical Metamaterials Exhibiting Negative Refractive Index", *Nature* **455**, 376 (2008).
3. D. A. Genov, K. Seal, X. Zhang, V. M. Shalaev, A. K. Sarychev, Z. C. Ying, H. Cao, "Collective Electronic States in Inhomogeneous Media at Critical and Subcritical Metal Concentrations", *Phys. Rev. B* **75**, 201403 (2007).
4. D. A. Genov, S. Zhang, X. Zhang, "Mimicking celestial mechanics in metamaterials", in press *Nature Physics* (2008).

Grant Proposals Pending:

1. Louisiana Board of Regents Support Fund – RCS
Title: "Metamaterials for Applications in STEALTH Technology (MAST)"
Funds requested: $ 138,652 Date Submitted: 10.21.2008

2. DoD-SBIR
Title: "Tunable Electromagnetic Metamaterials Films for STEALT enhancement "
Funds requested: $ 70,000 Date Submitted: 03.15.2009

3. EPSCoR Research Infrastructure Improvement - Track 1 RFP
Title: "Electromagnetic Metamaterials and Active Composites (EMAC) "
Funds requested: $ 1,340,444 Date Submitted: 04.20.2009

Grant Proposals Approved:

1. DOD RFP - College of Engineering and Science, Louisiana Tech University
Title: "Surface Plasma Enhanced Solar Cell (SPESC) ", Funds: $99,616
Date awarded: 06.04.2009

Peer previewed publications acknowledging the LONI Institute:

1. D. A. Genov, S. Zhang, and X. Zhang, "Mimicking celestial mechanics in metamaterials", accepted for publication to *Nature Physics* (May 2009).
2. D. A. Genov, A. K. Sarychev, and V. M. Shalaev, "Adiabatic Spatially Selective Photomodification of Inhomogeneous Metal-Dielectric Composites", submitted to *Physical Review Letters* (April 2009).

Presentations:

1. "*Electromagnetic metamaterials: from imaging with super resolution to mimicking celestial phenomenon in the lab*", The LONI Institute (LI) All-Hands Meeting, Baton Rouge, LA, October 31, 2008.

2. *Electromagnetic properties of complex metamaterials: from near field imaging with super resolution to mimicking celestial phenomenon in laboratory conditions*", Colloquium Series at the Center for Computational &Technology (CCT), Louisiana State University, Baton Rouge, LA, March 27, 2009.

Teaching at Louisiana Tech University:

Instructor:

ENGR 501/MSNT 502 Research Methods

Lecturer:

ENGR592/PHYS540/PHYS470s (Computational Methods)
ELEN411 (Electricity and Magnetism)

White Paper for Mark Jarrell, Professor of Physics and Astronomy, LSU

Group Members (Computation and Theory of Strongly Correlated Materials)

M Jarrell, Professor

J. Moreno, Assistant Professor

C. Slezak, Visiting Professor

C. Pao, Visiting Professor

C. Sen, Postdoc

D. Galanakas, Postdoc

S. Su

U. Yu, Postdoc

Z. Xu, Postdoc

J. Lu, Postdoc

V. Rousseau, Postdoc

E. Khatami, Student

K. Mikelsons, Student

M. Nili, Student

H. Fotso, Student,

P. Reis, Student

S. Yang, Student

K. Chen, Student

D. Norman, Administrative Assistant

Current Research: Strongly correlated materials display complex emergent phenomena, or behavior that emerges when many units are assembled that would not be predicted from a complete understanding of the units. This includes transition metal oxides, heavy fermion materials, organic magnets, and spintronic materials. The study of these systems is complicated by the competition of different ground states, including spin, charge and orbital ordering and by the lack of a small parameter. As a result, little progress has been made with conventional theory, and large-scale simulations are needed to form a more complete understanding of models of these systems.

We employ a variety of computational methods to study these systems. The Dynamical Mean Field Approximation (DMFA) [16,17] and its cluster extensions, including the Dynamical Cluster Approximation (DCA)[18-21] are at the heart of this approach. These approaches map the lattice onto a cluster embedded in a self-consistently calculated effective medium. Correlations within the cluster are treated explicitly while those at longer length scales are treated in a mean field approximation. The embedded cluster problem is solved using a quantum Monte Carlo (QMC)[22-24] simulation while disorder can be included by averaging over configurations. [18-20] Either a perfectly parallel (MPI) or a hybrid parallel (MPI+OpenMP) calculation is used. Nevertheless, the calculation is limited by the amount of memory available on each shared-memory node. A far more significant limitation of this technique is the minus sign problem, which is non-polynomial hard.[25] This means that all simulations of correlated electrons will grow exponentially with the inverse temperature and cluster size making very difficult to treat correlations on the important length scales.

To treat more complex problems a third length scale must be introduced as in the multi-scale many body (MSMB) approach.[26,27] This is accomplished by a multiple embedding scheme in which correlations over each length scale are treated with an appropriate approximation. Strong correlations at short length scales are treated with an explicit (numerically exact) QMC simulation on a small cluster. This cluster is embedded in the larger cluster where the weaker correlations at intermediate length scales are treated using the parquet approximation[28-32] which requires both contractions and rotations of rank three tensors (vertices) and a massively parallel computer with at least tens of thousands of processors. This cluster is embedded in an effective medium, which is used to treat correlations on the longest length scales.

Density functional theory is also an essential component of this project. Both the DMFA/DCA and the MSMB approach are parameterized by down folding LDA calculations.[27,33]

These materials and systems are of great technological importance. Correlated electron materials and especially transition-metal oxides show great promise for novel applications in the semiconductor industry to go beyond CMOS devices for future information processing technologies, which could be based on “state variables” such as spin. The chapters “Emerging Research Devices” and “Emerging Research Materials” in the 2007 International Technology Roadmap for Semiconductors (ITRS 2007)[11] stress that highly correlated electron systems exhibit coupling between orbital, charge, and spin ordering may enable new devices by greatly enhancing their sensitivity to different applied fields.

Our work will lead to a better understanding of these materials, which may lead to

better devices based upon them. We also develop and distribute a number of codes, which employ architectures at the forefront of computer science, including hyperparallel and multicore machines.

Our work relies upon the large-scale supercomputers available through LONI, the NSF Teragrid, and the DOE NLCS facilities at ORNL.

We enhance the impact of our work by distributing codes, and related courseware. Two complete courses, Solid State Physics and Classical Electrodynamics are distributed on the group web page, http://www.phys.lsu.edu/~jarrell . In collaboration with my student Cyrill Slezak, who visits LSU each summer for an extended period, we also participate in Inquiry based RET programs, and we are working with our colleagues to incorporate Inquiry and Active Classroom Teaching techniques into elementary courses in Physics and Astronomy.

As part of our NSF PIRE program, we teach a complete set of courses in Computational materials Science which are broadcast via interactive synchronous video to a number of schools in Germany and Switzerland. Our group is also the lead of a DOE SciDAC project involving researchers at LSU, OSC, UC Davis, and ORNL. The goal of this SciDAC is to develop the MSMB formalism mentioned above.

Recent Publications:

- N.S.Vidhyadhiraja, A.Macridin, C.Sen, M.Jarrell, Michael Ma Quantum Critical Point at Finite Doping in the 2D Hubbard Model: A Dynamical Cluster Quantum Monte Carlo Study , arXiv:0809.1477. Physical Review Letters, in press.
- E. Khatami, A. Macridin, M. Jarrell The validity of the spin-susceptibility "glue" approximation for pairing in a two-dimensional Hubbard model , arXiv:0901.4802.
- Karlis Mikelsons, Alexandru Macridin, Mark Jarrell The relationship between Hirsch-Fye and weak coupling diagrammatic Quantum Monte Carlo methods , arXiv:0903.0559.
- C. N. Varney, C.-R. Lee, Z. J. Bai, S. Chiesa, M. Jarrell, R. T. Scalettar High Precision Quantum Monte Carlo Study of the 2D Fermion Hubbard Model , arXiv:0903.2519.
- E. Khatami, C. R. Lee, Z. J. Bai, R. T. Scalettar, M. Jarrell Dynamical Mean Field Theory Cluster Solver with Linear Scaling in Inverse Temperature , arXiv:0904.1239.

Presentations and Talks:

- *The Phase Diagram of the Two-Dimensional Hubbard Model: A Quantum Critical Point at Finite Doping*, Invited Talk, Oct. 8, 2008, 3rd International Workshop on "Ordering Phenomena in Transition Metal Oxides" Augsburg, Germany, October 5-8, 2008
- *Bond Excitations in the Pseudogap Region of the Hubbard Model*, Invited Talk, Oct.

28, 2008, 21st International Symposium on Superconductivity, International Congress Center, Tsukuba Japan, October 27-29, 2008.

- *Massively Parallel and Multi-Scale Simulations of Strongly Correlated Electronic Systems.* , Invited Talk, March 4, 2009, Michael Dewar Memorial Symposium: Advancing Computational Chemistry Through High Performance Computing, from the Workstation to the Petascale and Beyond. March Meeting of the American Chemical Society. Salt Lake City, Utah.
- *Massively Parallel and Multi-Scale Simulations of Strongly Correlated Electronic Systems.* , Keynote Lecture, The International Conference on Computational Science 2009 May 25 - 27, 2009, in Baton Rouge, Louisiana

External Funding:

- *Simulations of Strongly Correlated Electronic Materials*, DMR-0706379, $375,000 over the three-year period 09/01/07-08/30/10 by the **National Science Foundation, Materials Theory Program**.
- *Graduate Education in Petascale Many Body Methods for Complex Correlated Systems*, OISE-0730290, $2,500,000 over the five-year period 9/1/07-8/31/12 by the **National Science Foundation, Office of International Science and Engineering (OD/OISE)**. Investigators: Juana Moreno (PI) UND, M. Jarrell (Co-PI) and K. Tomko (Co-PI) at the Univ. of Cincinnati.
- *Predictive Capability for Strongly Correlated Systems*, DOE DE-FG02-04ER46129, as part of a Computational Materials Science Network, $121,200 over the three-year period 04/15/07-04/14/10 (to be approved year by year) by the **Department of Energy, Basic Energy Sciences, CMSN** (Warren Pickett, UC Davis, PI).

Next Generation Multi-Scale Quantum Simulation Software for Strongly Correlated Materials DE-FC02-06ER25792 $3,000,000 over the five-year period 7/06-6/11, by the **Department of Energy, SciDAC**. Investigators: M. Jarrell (PI) and K. Tomko at the Univ. of Cincinnati, Th. Maier (co-PI) and E. DÁzevedo at ORNL, Z. Bai (co-PI) R.T. Scalettar and S. Savrasov at UC Davis.

Recent Applications for External Funding:

Proposal	PI Name	Deadline	Amount Requested	Status
34971 - 1	Jarrell, Mark	05/28/2009	$294,936.00	Submitted

Sponsor: Ohio Supercomputer Center (OSC) *Type:* New

Date Approved: 05/26/2009

Project Title: Improving Developer Productivity for HPC through Cyberinfrastructure: Applications, Languages, Tools and Services

Proposal Specialist: Young, Shirley Langford

Award Specialist:

Jarrell, Mark 04/30/2009 $20,000,000.00 Submitted

Sponsor: Board of Regents - BOR *Type:* Preproposal/NOI

Date Approved: 04/29/2009

34891 - 1 *Project Title:* Louisiana Graduate Research and Education Program in Computational Materials Science

Proposal Specialist: Williams, Rhonda Meyers

Award Specialist:

Hall, Randall W 04/20/2009 $10,000.00 Submitted

Sponsor: Board of Regents - BOR *Type:* New

Date Approved: 04/20/2009

34809 - 1 *Project Title:* Planning for LONI Institute s Proposal to the 2009 Louisiana EPSCoR RII Competition

Proposal Specialist: Hakes, Jonathan

Award Specialist:

Jarrell, Mark 04/17/2009 $67,657.00 Submitted

Sponsor: Dept of Energy - DOE *Type:* New

Date Approved: 04/17/2009

34778 - 1 *Project Title:* Predictive Capability for Strongly Correlated Systems: Mott Transition in MnO, Multielectron Magnetic Moments, and Dynamic Effects in Correlated Materials

Proposal Specialist: Impson, Dana Tuminello

Award Specialist:

Sheehy, Daniel E 02/27/2009 $1,721,689.94 Submitted

Sponsor: Dept of Energy - DOE *Type:* New

Date Approved: 03/10/2009

34627 - 1 *Project Title:* TMS: Integrated theoretical approaches to correlated systems

Proposal Specialist: Williams, Rhonda Meyers

Award Specialist:

Jarrell, Mark 03/26/2009 $887,664.00 Submitted

Sponsor: Dept of Energy - DOE *Type:* New

Date Approved: 03/27/2009

34579 - 1 *Project Title:* Next Generation Multi-Scale Quantum Simulation Software for Strongly Correlated Materials

Proposal Specialist: Impson, Dana Tuminello *Award Specialist:*

Jarrell, Mark 11/26/2008 $500,000.00 Submitted

Sponsor: Ohio Supercomputer Center (OSC) *Type:* New

Date Approved: 11/26/2008

34304 - 1 *Project Title:* An Experimental Accelerator-Based HPC System driven by High Productivity Programming Models

Proposal Specialist: Li, Ping *Award Specialist:*

Jarrell, Mark $0.00 Submitted

Sponsor: Board of Regents - BOR *Type:* Modification

Mod Type: Change of Principal/Co-Investigator, *Date Approved:* 05/19/2009

31914 - 6 *Project Title:* The LONI Institute: Advancing Biology, Materials, and Computational Sciences for Research, Education, and Economic Development

Proposal Specialist: Courville, Darya Delaune *Award Specialist:*

2009 ANNUAL REPORT FOR THE LONI INSTITUTE GRANT

Damir B. Khismatullin

Department of Biomedical Engineering
Tulane University, New Orleans, LA 70118

Graduate students (2009-current):
Hongzhi Lan, Chong Chen

Undergraduate students (2008-2009):
Daniel Haber, Joseph Berenblit

Specialization:

Computational modeling and experiments

Research fields:

1) Bio-transport and cellular biomechanics;
2) Cell-cell interactions in inflammation, atherosclerosis, and thrombosis;
3) Medical ultrasound and biomedical applications of gas microbubbles;
4) Multiphase flows and non-Newtonian fluid mechanics

I started my appointment as Associate Professor of Biomedical Engineering at Tulane University, with 50% support from the LONI Institute Grant, in August 2008. My research interests focus upon 1) modeling the mechanical behavior of biological systems at cellular and tissue levels and 2) the numerical solution of the multiphase fluid flow problems where the deviation from the Newtonian law is significant. Specific points of current interest include the biomechanics of leukocytes, platelets and endothelial cells; leukocyte-endothelial cell interactions in inflammation and atherosclerosis; thrombus formation and rheology; microvascular and arterial blood flow; contrast-enhanced ultrasound imaging, and shock-wave lithotripsy. In my research laboratory, we also conduct in vitro experiments on adhesive interactions of living cells. The main objective of my research is to integrate computational modeling, in vitro and in vivo experiments to improve understanding of the behavior of biological systems under both physiological and pathophysiological conditions. **Achieving this objective will have a major impact in development of therapy against inflammation, atherosclerosis, and thrombosis (pathologies responsible for the majority of death and hospitalization in Louisiana and other states) and thus will be of benefit to the majority of Louisiana population. It also helps to establish the strength of the LONI, and Louisiana as a whole, in biomedical computational science.**

During the second year of the grant period (August 2008 - June 2009), my activities included: 1) development of a research laboratory equipped with both experimental and computational facilities; 2) writing grant proposals for NIH, NSF, and BoR; 3) running numerical simulations and conducting experimental studies for the projects listed below; 4) establishing new external and internal collaborations; 5) teaching the cell mechanics course in Fall 2008 and the bio-transport course in Spring 2009; 6) giving invited presentations and conference talks; and 7) serving as a reviewer for NIH, California HIV/AIDS Research Program, and various scientific journals. My current research projects are

Project 1: Quantitative analysis of monocyte-endothelium interactions in atherosclerosis. NIH Challenge Grant Application (PI: Khismatullin; pending: submitted in April 2009)

External Collaborators: Klaus Ley (La Jolla Institute for Allergy & Immunology, CA)
George A. Truskey (Duke University, Durham, NC)
Internal Collaborator: Donald P. Gaver, III

The goal of this proposal is a) to examine the combined effect of hypercholestermia (oxLDL) and disturbed flow on monocyte adhesion to endothelium at atherosclerosis-prone sites and b) to develop robust quantitative models of monocyte-endothelium interactions that can be used as a tool to explore therapies for atherosclerosis.

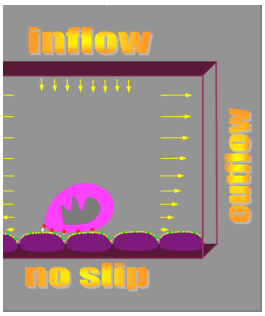


Figure 1. Schematic drawing of the flow domain used to study numerically the effect of recirclating flow on monocyte-endothelial cell adhesion.

Preliminary results: We have developed the algorithm for a fully three-dimensional, direct numerical simulation of receptor-mediated adhesion of human monocyte in a rectangular micro-channel. In addition, we implemented stagnation-point flow conditions, schematically shown in Figure 1, in the algorithm to study the effect of recirculating flow on monocyte rolling and adhesion. Figure 2 shows the disturbed velocity field and monocyte shape in the midsagittal plane, according to our simulations.

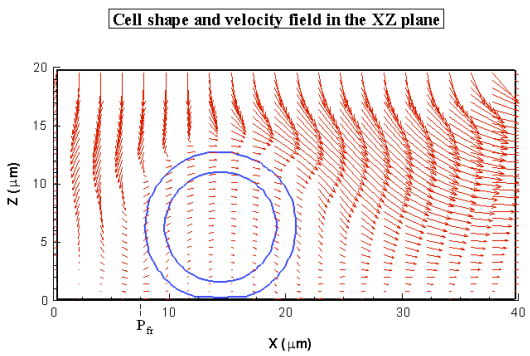


Figure 2. Cell shape and velocity field at the mid-plane of the cell when the reattachment point P_{fr} is at 7.5 μm from the left boundary of the flow domain and t = 0.036 s. The cytoplasmic viscosity is 200 poise. The nucleus-to-cytoplasm viscosity ratio is 2.5. The microvillus length and radius are 0.3 μm and 0.1 μm. The number of microvilli is 1016. The on and off rates are 1.0 $\mu m^2/s$ and 0.5 s^{-1}. The surface densities of adhesion molecules are 1000 μm^{-2}.

Our computational model is a custom incompressible computational fluid dynamics (CFD) code, written in Fortran with OpenMP directives, in which the Volume-of-Fluid (VOF) method is used on a Marker-and-Cell (MAC) grid for tracking the position and shape changes of the cell over time. The code finds the solution of the continuity and Navier-Stokes equations in three dimensions and takes into account viscoelasticity of the cell through the Giesekus model. The kinetic model of receptor-ligand binding is used to calculate the adhesive force on the cell surface. Adhesive interactions are initiated if the separation distance between the substrate and the free end of a ligand molecule is less than or equal to the unstressed length of receptor.

Currently, we run the code on the eight-processor node I purchased for the Tulane CCS. One of my students (Hongzhi Lan) works on the implementation of MPI into the algorithm to run it on a LONI cluster (his research is supported by 2009-2010 LONI Institute Graduate Fellowship).

As about the experimental part of the project, my research team (graduate student Chong Chen and undergraduate students Daniel Haber and Joseph Berenblit) has developed a parallel-plate flow chamber assay for analysis of monocyte-endothelial cell interactions under flow (Figure 3). The cell systems we currently study are THP-1 (monocytic cell line) and HUVEC (human umbilical vein endothelial cell).

My research laboratory is now fully equipped to conduct cell-cell interaction studies in vitro.

Figure 3. Complete dual pump flow chamber system installed in the laboratory.

We are also in process of purchasing the Bioflux 200 flow system (Fluxion Biosciences, San Francisco, CA) which utilizes Fluxion's Well Plate Microfluidics™ (straight micro-channels in 24- or 48-well plates). This system will be used in the project to study monocyte-substrate and monocyte-endothelium interactions in straight channels under steady flow conditions.

Project 2: Computational studies of leukocyte dynamics using microPIV in collagen microchannels. NIH R21 Grant Application (PI: Khismatullin; in revision: to be submitted in July 2009)

Internal Collaborators: Sergey Shevkoplyas, Donald P. Gaver, III

The goal of this proposal is to develop a) a well-controlled in vitro system that could mimic leukocyte behavior in vivo and b) a computational algorithm that realistically describes leukocyte rolling in this system. The studies proposed in this R21 application form the basis for an independent R01 proposal to investigate leukocyte-endothelium interaction properties in ex vivo collagen-tube assay and in vivo.

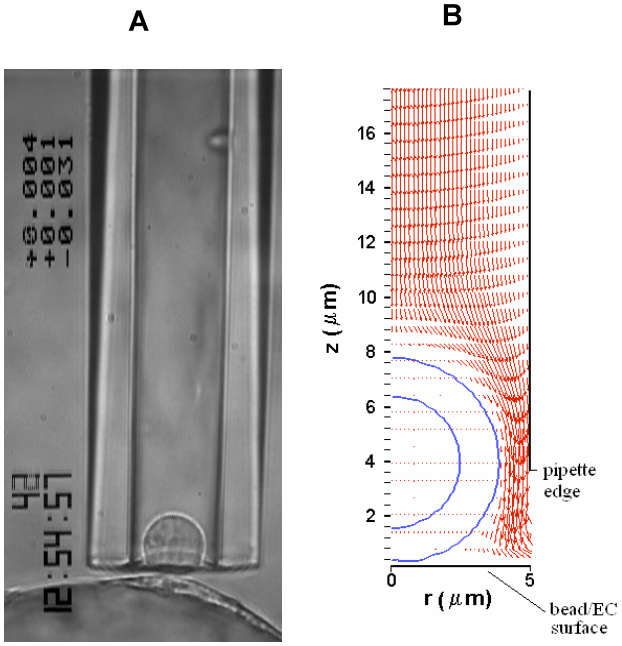


Figure 4. (A) A videomicrograph frame showing adhesion of THP-1 cell to an endothelial cell in the micropipette system. The leukocyte motion toward the endothelial cell and subsequent leukocyte-EC interaction were initiated by negative suction (J. P. Irick and G. A. Truskey, unpublished). (B) Numerical image showing the velocity field during negative suction and the leukocyte interacting with the substrate (bead/EC surface) in the micropipette system.

Preliminary results: As discussed in Project 1, we have developed a 3-D computational model of receptor-mediated leukocyte adhesion to the endothelium or a ligand-coated surface in a parallel plate flow chamber. We are currently working on extension of this model to the case of cylindrical geometry. Two codes are now in the testing stage: the computational model for leukocyte dynamics in a micropipette system (axisymmetric flow problem), which is illustrated in Figure 4B and the 2-D tissue-blood model in which the adhesive surface is the interface between the tissue and blood phases. These two algorithms will then be combined to develop the computational model for leukocyte adhesion to the surface of a collagen microchannel with a circular cross section.

Our experimental efforts in this proposal will be lining the microchannels with endothelial cells and conducting flow assays. The microchannels will be fabricated by Dr. Shebkoplyas.

Other projects that are initiated or submitted in 2008-2009 are "**Thrombus rheology via noncontact measurement**" (external collaborator is Dr. R. Glynn Holt from Boston University; the first submission was highly recommended; the second submission was sent to NSF in October 2008), where my role is to perform computational study of blood clot shape oscillation in an acoustic levitator to determine the blood clot material constants; "**Computational and experimental studies of thrombus growth and deformation**" (external collaborator is Dr. Steven Jones from Louisiana Tech; the proposal was submitted for consideration in the LONI proposal on Computational Materials), where my role is to develop the computational model for thrombus growth and deformation; "**Fluid mechanics of capillary sprouting**" (internal collaborator: Dr. W. Lee Murfee), where my role is to develop the computational model of blood flow in capillaries which takes into account vessel sprouting (I have already developed a 2-D model for this project as mentioned in Project 2). My research in 2008 was funded by the Department of Defense (the project on "**Laser nucleation and collapse stability for advanced cavitation power technology**") where I performed mathematical modeling of bubble cluster dynamics in a spherical flask.

During the second year of the grant period, I have also established internal collaboration with Drs. Ricardo Cortez and Lisa Fauci (our interests are to combine the immersed boundary and viscoelastic VOF algorithms to study numerically the effect of viscoelasticity on the motion of microorganisms) and Dr. Noshir Pesika (in vitro experiments and computational analysis of the effect of channel height on cell adhesion to surfaces using the micro-channels fabricated by Dr. Pesika) as well as external collaboration with Dr. Wendy Thomas (University of Washington) to study bacterial adhesion to surfaces.

Publications:

D.B. Khismatullin and G.A. Truskey, "Leukocyte Rolling on P-selectin: A 3D Numerical Study of the Effects of Cell Viscosity and PSGL-1 Clustering," *Ann. Biomed. Eng. (in revision)*

D. B. Khismatullin, "The cytoskeleton and deformability of white blood cells" in Klaus Ley (Ed.), "Current Topics in Membrane. Vol. 64. Leukocyte adhesion" (Elsevier, scheduled to be published in 2009).

Presentations:

D.B. Khismatullin, "Viscoelastic Volume-of-Fluid algorithm for multiphase flow problems", LONI HPC Workshop, Tulane University, April 13-14, 2009 — New Orleans, Louisiana.

D.B. Khismatullin, "Application of the Volume-of-Fluid algorithm to biological systems", 2009 Spring Southeastern Meeting of the American Mathematical Society, April 4-5, 2009 — Raleigh, North Carolina.

D.B. Khismatullin, "Modeling of cell adhesion using a multiphase flow approach", LONI Institute First All-Hands Meeting, Louisiana State University, October 31, 2008 — Baton Rouge, Louisiana.

Invited Talks:

1. Tulane University School of Medicine, Department of Physiology (May 18, 2009). *Quantitative analysis of leukocyte-endothelial cell interactions in inflammation and atherosclerosis.* Host: *Dewan Majid*.
2. Tulane University, Department of Chemical and Biomolecular Engineering (April 24, 2009). *Computational modeling of receptor-mediated leukocyte adhesion to surfaces.* Host: *Noshir Pesika*.
3. Tulane University, Applied and Computational Mathematics Seminar (January 23, 2009). *Biological systems modeling using a multiphase flow approach.* Host: *Ricardo Cortez*.
4. Southern Methodist University, Department of Mathematics (October 15, 2008). *A multiphase flow approach to modeling biological systems.* Host: *Vladimir Ajaev.*

External Funding:

1. National Institutes of Health, National Heart, Lung, and Blood Institute (NIH-NHLBI). Type: RC1 (Challenge Grant). Title: *Quantitative analysis of monocyte-endothelium interactions in atherosclerosis* (pending). Role: Principal Investigator. Collaborators: Klaus Ley (LIAI), Donald P. Gaver, III (Tulane U.), George A. Truskey (Duke U.).
2. National Institutes of Health, National Heart, Lung, and Blood Institute (NIH-NHLBI). Type: R21. Title: *Computational studies of leukocyte dynamics using micro-PIV in collagen microchannels* (in revision, to be submitted in July 2009). Role: Principal Investigator. Co-I: Sergey Shevkoplyas and Donald P. Gaver, III (Tulane U.).
3. National Science Foundation, Nano and Bio Mechanics program. Title: *Thrombus rheology via noncontact measurement* (in revision). Role: Principal Investigator. Co-PI: R. Glynn Holt (Boston U.).
4. Department of Defense. Title: Laser nucleation and collapse stability for advanced cavitation power technology (subcontract, completed). Role: Co-I (PI: R. Glynn Holt). Subcontract amount: $395,000 for March-December, 2008.

Computational free energy studies from molecular simulations

PI: David L. Mobley

Research group: Pavel Klimovich (graduate student), Matthew Hellmers (undergraduate student), Christopher Savoie (undergraduate student)

Affiliations: Department of Chemistry, University of New Orleans; LONI

Free energies govern a huge number of interesting chemical and biological processes. Free energies drive biomolecular association and dissociation, solubility, permeation, and transfer between different environments. My research group is particularly interested in understanding, predicting, and manipulating these free energies. We develop and apply computational methods to predict binding free energies, transfer free energies, and solubilities based on computer simulations of the molecules involved.

Our work on binding free energies between proteins and small-molecule ligands was essentially the first work to compute rigorous binding free energies between proteins and small molecule ligands, without requiring the bound structure of the protein and ligand as input. Using molecular dynamics simulations, we computed binding free energies beginning from the unbound protein structure, and predicted ligand binding modes. We successfully tested the approach we developed for making blind predictions in a model binding site, and are currently applying the approach to other protein binding sites. Ultimately, the methods we are developing for studying protein-ligand binding free energies will have application to computational drug discovery, biomolecular association generally, and to guide design of new enzymes.

We also have invested significant effort in predicting small molecule solvation (gas-to-water transfer) free energies. These are now straightforward to calculate; we recently computed them for a set of 504 different small molecules. We also predicted these in several blind tests. These solvation free energies are seen as a proxy for the ability of current molecular dynamics force fields to accurately describe binding interactions, as removal of a small molecule from solvent is an important part of the binding process. We have largely been fairly successful in predicting solvation free energies, though failures have also guided us to deficiencies in the force field and point the way towards further force field developments. Consequently, this work has the potential for large payoffs in diverse areas, from protein-ligand binding, to protein folding and protein structure prediction, to understanding surface interactions and properties of materials.

A new focus in the group is predicting solubilities of small molecules – the concentration above which a molecule will fail to dissolve in water. This is important in a huge number of contexts, from oil extraction (where pipelines can be blocked by solid deposits) to drug discovery (where most drugs must dissolve after being taken in pill form). Solubility is simple in theory – it is determined by the balance of favorable interactions within a solid form, with favorable interactions between the molecule and water. But this is easier said than predicted, and computational methods are only just reaching the point where this is becoming a tractable problem. Given our expertise in solvation, we have a handle on half of the solubility problem, and are now beginning work to handle the solid state. Improved methods for predicting solubilities will help guide efforts to control solubility, for example in a drug discovery or chemical reaction context.

Overall, the research has the potential to transform a variety of fields that are currently governed by experimental trial and error. Computational methods have so far been unreliable enough for these problems that it is typically preferred to simply do the experiment. This research will help bring computation to the point where computational results can reliably predict experiments, paving the way for computers to guide scientific discovery rather than trial and error. Experiment could be used to confirm computational predictions, rather than the current approach of merely using computation to help rationalize experimental results.

This work is heavily dependent on existing Louisiana cyberinfrastructure, in particular the Louisiana Optical Network Initiative (LONI). It is tremendously demanding computationally, and so high performance computing is key to pushing these models forwards.

Much of my work is collaborative. I am beginning to develop collaborations with experimental groups at Louisiana State University and pharmaceutical companies. I already have existing collaborations with an experimental group at the University of California, San Francisco (UCSF), and other computational groups at UCSF, Merck, the University of Notre Dame, and others.

I currently supervise a graduate student and two undergraduate students, and will be developing and teaching a graduate level course in computational chemistry/molecular modeling in Fall 2009.

Publications since joining the University of New Orleans:

David L. Mobley* and Ken A. Dill, "The binding of small-molecule ligands to proteins: 'What you see' is not always 'what you get', *Structure* **17**(4), 489-498 (2009), 10 pages. * - corresponding author.

D. L. Mobley[+], C. I. Bayly, M. D. Cooper, and K. A. Dill. "Predictions of hydration free energies from all-atom molecular dynamics simulations", invited article, *Journal of Physical Chemistry B* 113: 4533-4537 (2009), special issue on "Calculation of Aqueous Solvation Energies of Drug-Like Molecules: A Blind Challenge.

D. L. Mobley[+], C. I. Bayly, M. D. Cooper, M. R. Shirts, and K. A. Dill. "Small molecule hydration free energies in explicit solvent: An extensive test of fixed-charge force fields", *J. Chem. Theory Comput.* **5:** 350-358, 2009 (DOI 10.1021/ct800409d), 9 pages. One of the top 10 most downloaded articles in JCTC between March, 2008 and March, 2009.

Presentations/talks:

"Lessons learned from predicting binding free energies in model binding sites" and "Quantitative predictions of protein-ligand binding affinities", American Chemical Society Meeting, Salt Lake City, UT, March 2009, contributed presentation.

"Predictive calculations of absolute binding free energies", American Chemical Society Meeting, August 20, 2008, Philadelphia, PA, invited presentation.

External funding:

None received yet.

LONI Institute Whitepaper

for the LONI Institute Second Annual Report

Principal Investigator:
Christopher M. Taylor
Assistant Professor
Department of Computer Science
University of New Orleans

Group Members:
Jack Torres
Tulane University Undergraduate
Mathematics Major, Rising Senior
Specialization: Analysis of Microbial Diversity

Nathan Simpson
University of New Orleans Undergraduate
Computer Science Major, Rising Senior
Specialization: Analysis of MicroRNA Sequencing Data

Project Description

Dr. Christopher Taylor is a new assistant professor in the Department of Computer Science at the University of New Orleans. He is a member of the bioinformatics group and has a joint appointment at the Research Institute for Children (RIC), part of the Children's Hospital of New Orleans. RIC provides Dr. Taylor with research support and collaborative opportunities with a staff of LSU-HSC faculty performing research at the institute. Dr. Taylor's research is highly collaborative in nature and he works with a variety of biologist to design algorithms for analysis of experimental data. In particular his group focuses on:

- Developing Algorithms for Analysis of Genomic Data
    - High-throughput Sequencing Data
    - Genome Tiling Microarray Data
- Analysis of Human Microbiome Data
    - Classify Diverse Constituents Present in Samples
    - Phylogentic Analysis of Metagenomes
- Application of GPU Hardware to Sequence Mapping
    - Acceleration of Existing Mapping Algorithms
    - Technology and Application Specific Mapping Algorithms

High-throughput sequencing and DNA Microarray technologies have transformed the landscape of research in biology from the single experiment-single result model to an interrogation of the entire genome from a single experiment. These technologies provide researchers with an abundance of data, but also necessitate the development of specialized analysis algorithms to process the results. Many of the emerging technologies are rapidly changing and present new computational challenges with each new generation of equipment. Our research group collaborates closely with the biologists who perform these experiments to develop new algorithms and analysis techniques to tackle these constantly evolving problems.

One of our ongoing collaborations involves Human DNA replication. Genome tiling microarrays were used to assay DNA replication timing. The data from these experiments provides a very noisy and discrete view of the replication timing. We developed algorithms to reconstruct a continuous profile of the DNA replication timing. We used this profile to identify possible sites of replication origin, investigate correlation with other genomic markers, and to design additional biological experiments. We are currently designing algorithms to analyze other aspects of replication activity such as nascent strand DNA, ORC binding sites, and nuclear matrix attachment regions. This collaboration involves researches at the University of Virginia and is supported through funding from NIH.

Another collaboration that is beginning this Summer involves researchers at Tulane University and Xavier University. Supplemental funding from NIH is in place to support the work for this collaboration to study the effects of a pair of MicroRNAs transfected into cells. High-throughput sequencing technology is being used to interrogate on a genomic scale and we

are helping to design algorithms to analyze this data. We wish to investigate the synergy effect of MicroRNA targets, and the effect on biological pathways in the cell.

Our research group is also beginning collaborative work with a microbiologist at RIC who is affiliated with LSU-HSC. We are studying the Human microbiome to assess the affects of the diverse colonization of bacteria that lives on and inside of the human body. This work requires development of algorithms to interpret sequencing data and downstream analysis of the results. We are currently in the process of applying for a grant to fund this research.

Finally, our group is interested in exploring the application of GPU hardware to improve the accuracy and speed of mapping new high-throughput sequencing reads to a reference genome. We are in the process of procuring a laptop with GPU hardware to perform initial experiments and collect preliminary results for future funding opportunities.

Publications:

Invited Book Chapters:

- Neerja Karnani, Christopher M. Taylor and Anindya Dutta. Microarray Analysis of DNA Replication Timing. Microarray Analysis of the Physical Genome. *Methods in Molecular Biology.* Vol 556, ISBN: 978-1-60327-191-2, Humana Press. June 16, 2009.

Refereed Journal Articles:

- Encode Project Consortium. Identification and Analysis of Functional Elements in 1% of the Human Genome by the Encode Pilot Project. *Nature.* 2007 Jun 14;447(7146):799-816.

- Neerja Karnani, Christopher Taylor, Ankit Malhotra and Anindya Dutta. Pan-S Replication Patterns and Chromosomal Domains Defined by Genome-Tiling Arrays of Encode Genomic Areas. *Genome Research.* 2007 Jun;17(6):865-76.
- Encode Project Consortium. The Encode (ENCyclopedia Of DNA Elements) Project. *Science.* 2004 Oct 22;306(5696):636-40.

Refereed Conference Papers:

- Anindya Dutta, Neerja Karnani, Ankit Malhotra, Gabriel Robins and Christopher M. Taylor. Extraction of Human DNA Replication Patterns from Discrete Microarray Data. *Third IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2008)*, Novotel St Kilda, Melbourne Austrailia, October 2008.

Presentations:

Conference Talks:

- Anindya Dutta, Neerja Karnani, Ankit Malhotra, Gabriel Robins and Christopher M. Taylor. Extraction of Human DNA Replication Patterns from Discrete Microarray Data.

Third IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2008), Novotel St Kilda, Melbourne Austrailia, October 2008.

Posters:

- Christopher Taylor, Neerja Karnani, and Anindya Dutta. Analyzing DNA Replication Timing in the Human Genome. *Experimental Biology.* Ernest N. Morial Convention Center, New Orleans, LA, April 2009.

Invited Talks:

- Christopher M. Taylor. Extraction of Human DNA Replication Timing Patterns from Discrete Microarray Data. *LONI All-Hands Meeting.* Baton Rouge, LA, October 2008.

Funding:

- PI: Erik Flemington, coPI: Christopher Taylor, coPI: Dongxiao Zhu, coPI: Kun Zhang. Title: Administrative Supplements Providing Summer Research Experiences for Students and Science Educators. Source: National Institutes of Health. Supplement to Analysis of Epstein Barr virus type III latency on cellular miRNA gene expression. Amount: $216,386. Date approved: May 15, 2009. Funding Period: June 01, 2009 to August 31, 2010.
- PI: Christopher M. Taylor. Title: Taylor Summer Salary Professional Service Agreement. Source: Research Institute for Children. Amount: $36,080. Date approved: March 3, 2009. Funding Period: May 17, 2009 to August 15, 2009.

Developing a High Performance Computational Biology and Material Science Lab at Southern University (HPC-BMSL)

Ebrahim Khosravi, Ph.D.	Chairman and Professor of Computer Science
Shuju Bai, Ph.D.	Associate Professor of Computer Science
Rachel Vincent-Finley, Ph.D.	LONI Faculty, Assistant Professor of Computer Science
Shizhong Yang, Ph.D.	LONI Institute Computational Scientist
Kimberlee Lyles	LONI Institute Graduate Fellow

Research Subproject:

A novel reduced coordinate space method for molecular dynamic simulations

Investigator:

Rachel E. Vincent-Finley, Ph.D., LONI Institute Faculty

Assistant Professor of Computer Science

Southern University and A & M College

Baton Rouge Louisiana

Today's challenges in molecular dynamics (MD) include the simulation of protein molecules consisting of tens of thousands to millions of atoms over microsecond time intervals. Achieving this goal will require state-of-the-art innovative technologies in protein modeling, computational methods, high performance computing hardware and software methodology, data structures, and computer visualization.

My research involves the development of methods for performing molecular simulations with respect to a reduced coordinate space. Consider a molecule containing n atoms and a 3n dimensional (3nD) space defined by a concatenation of the 3D Cartesian coordinates of the atoms. Given a standard MD trajectory, a collection of molecular conformations over time, I use principal component analysis (PCA) to identify k dominant characteristics of a trajectory and then construct a kD representation of the atomic coordinates with respect to these k characteristics. Given this new representation of a molecule, I define equations of motion and perform simulations. This simulation method allows for the efficient simulation of test molecules by reducing the storage and computational requirements of each simulation.

The essential components of my reduced simulation method include isolating k dominant features of an MD trajectory using ARPACK and defining a kD coordinate space; constructing an approximation of a potential energy surface based on the defined coordinate space; updating coordinates and velocities in the kD space based on the approximate energy surface; and analyzing the resulting information with respect to the original 3nD coordinate space.

One will generally have access to a modest portion of a molecular trajectory, thus a kD energy surface approximation should capture essential details of the underlying empirical energy surface. My research includes the implementation and analysis of approximation techniques with respect to the defined kD space. I compare my simulations to experimental data, such as data obtained from infrared (IR) spectroscopy. The IR spectrum of a molecule shows which frequencies of IR radiation are absorbed by the molecule and can be used to identify the functional groups in a molecule. This is also information that this reduced simulation method seeks to reveal. With respect to molecular mechanics, the power spectrum reveals the underlying frequencies of the molecular processes. Thus the power spectrum provides information about the dynamic behavior of atomic interactions and can be compared to experimental data, such as the IR spectrum of a molecule.

The study of protein structure and function is a driving force in research that attempts to identify a relationship between the primary structure, or sequence, of a protein and its three dimensional structure. Ideally this would allow for inferences about the three-dimensional structure of a new sequence given its similarity to a known primary structure. However, care must be taken in the evaluation of this information for some proteins are structurally similar, yet have a low percentage of matching amino acids. This is attributed to the evolution of (biological) sequences which present themselves in the form of insertions, deletions, and substitutions to a protein sequence.

Comparisons of proteins at the primary structure level begin with sequence analysis. A basic problem is to find an optimal alignment between a pair of sequences allowing for gaps. Here optimal is measured by minimizing a specified cost function. There are various dynamic programming algorithms available to find an optimal alignment. However, when the number of sequences to be aligned is greater than 2, straightforward generalization of dynamic

programming algorithms becomes cumbersome. Specifically, time and memory complexities are $O(l^k)$ and $O(2^k l^k)$, respectively, where l is the average length of the k aligned sequences.

This overall project provides opportunities for undergraduate and graduate student research across multiple disciplines. Computer science students in particular will provide a vital resource for compiling and analyzing available molecular dynamics simulation and sequence alignment data. Students will then use insight gained from utilizing developed subroutines to suggest further research questions and to support the needs of biological scientists.

Initial work on this subproject in the LONI Institute context began in May 2009 when I joined the Computer Science faculty at Southern University of Baton Rouge. The goal is to involve current graduate students in the process of molecular analysis and code development. Independent study modules will provide advanced undergraduate students opportunities to become involved in various aspects of the project.

Appendix B

LI Computational Scientists White Papers

Computational Modeling of Lung Parenchyma Tethering Small Airways

Hideki Fujioka
Center for Computational Science
Tulane University

Donald P. Gaver
Biomedical Engineering Department
Tulane University

Project Description

The delicate structure of the lung epithelium makes it susceptible to surface tension induced injury. In cases of sepsis leading to multi-organ failure, large regions of the lung can fill with fluid (become atelectic), eliminating gas-exchange to a significant fraction of the lung. Prematurely born neonates are likely to have an immature surfactant system, resulting in elevated interfacial surface tension. As a result, large regions of the lung may remain atelectic or close due to interfacial instabilities.

Patients suffering from acute lung injury (ALI) cannot breathe on their own due to the collapse and fluid occlusion of small pulmonary airways. These patients must be placed on a mechanical ventilator in order to survive. However, the microbubble flows generated during ventilation can exacerbate the existing lung injury and the mortality rate for ALI is very high (30-40%). Even for those patients who survive ALI, the tissues of the lung may be damaged due to the mechanical environment, which exposes the sensitive epithelium to abnormal physical forces that can initiate or exacerbate lung injury. This may occur with mechanical ventilation, leading to ventilator-induced lung injury (VILI).

The process of opening collapsed atelectic region involves complex fluid-structure interactions that depend on airway geometry, fluid properties and surfactant biophysical processes. Previous computational studies indicate that microbubble or liquid plug flow imparts a complex combination of normal and shear stress to the epithelial cells on airway walls. Experiments in *in vitro* systems clearly demonstrate that these flows can impart injurious mechanical stresses on airway and alveolar epithelial cells. Hydrodynamic stresses may also be transduced into injurious biological responses including the up-regulation of inflammatory pathways and altered surfactant secretion.

While most models to date have relied on single or small networks of airways, in reality pulmonary airways are surrounded by parenchyma that consists of numerous alveoli, all of which are connected to distal airways. Therefore, the dynamics of each airway and alveolus is interdependent. As such, the behavior of one component may affect all others through parenchymal tethering. For this reason, the collapse and reopening process of an airway may affect other reopening processes.

Alveoli occupy the space in the lung with maximal air space. The truncated-octahedron is a space-filling geometry, and relevant for an ideal model alveoli (Fung,YC, *J.Appl.Physiol*., 64(5):2132,1988). We have constructed 3-dimensional computational model of a truncated-octahedron alveoli referring to the model developed by (Dale, PJ, et al. *J.Biomech*. 13:865,

1980). We will analyze the deformation and stress/strain fields of the alveoli surrounding a tube when an external force is applied.

Presentation/talks

N/A

External Funding

N/A

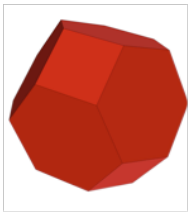


Figure 1: A Truncated-Octahedron Alveoli Model

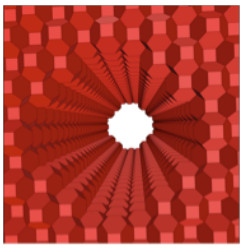


Figure 2: Alveolus Surrounding a Tube.

A computational investigation of surfactant transport during pulsatile airway reopening

Jerina E. Pillert
Biomedical Engineering Department
Tulane University

Hideki Fujioka
Center for Computational Science
Tulane University

David Halpern
Department of Mathematics
University of Alabama

Donald P. Gaver
Biomedical Engineering Department
Tulane University

Project Description

The mechanics of pulmonary airway reopening directly depend upon the efficacy of surfactant transport and adsorption to the continuously expanding and collapsing air-liquid interface. In disease states such respiratory distress syndrome (RDS) and acute respiratory distress syndrome (ARDS), occluded airways undergoing mechanical ventilation are exposed to a damaging normal-stress gradient that sweeps across the epithelial cell layer as the finger of air progress into the fluid occlusion. While it is known that this injury can be mediated by the continuous transport and adsorption of surfactants to the air-liquid interface, the dynamics between the fluid mechanics and surfactant sorption have yet to be fully understood. Since our prior experiments (*Biophysical Journal* 2008 Vol. 96) have shown that pulsatile flow enhances surfactant transport, our aim with this current study is to identify ventilation waveform parameters that optimize transport. To complete this investigation, we will computationally

simulate the transport and adsorption of surfactant surrounding a semi-infinite bubble that propagates through an occluded airway. Our computational domain consists of the fluid region ahead of and surrounding a semi-infinite finger of air. We will utilize the boundary element and finite volume methods to assess the transport of surfactant and the effect of the adsorbed surfactant on interfacial shape and its influence on the mechanical stress field; particularly the normal-stress gradient that can damage epithelial cells in ventilator-induced lung injury (VILI).

Jerina Pillet, a graduate student at Biomedical Engineering Department at Tulane University is working on this project. Hideki Fujioka helps her to develop the finite volume numerical code to solve the surfactant transport equations coupled with the boundary element CFD code. Currently, she is working on programming with the Fortran code that solves the convective diffusion equations on the air-liquid interface and the bulk liquid.

Presentation/talks

N/A

External Funding

NIH RO1 – HL81266

Computational Modeling of Fluid Dynamics and Transport in Microfluidic Mixing Devices

Katharine L. Hamlington
Biomedical Engineering Department
Tulane University

Hideki Fujioka
Center for Computational Science
Tulane University

Yuen-Yick Kwan
Center for Computational Science
Tulane University

Ricardo Cortez
Center for Computational Science
Tulane University

Donald P. Gaver
Biomedical Engineering Department
Tulane University

Project Description

Our goal is to determine the optimal geometric configuration of a passive microfluidic-mixing chamber to enhance the reactions of an analyte-antibody biosensing system. Convective

mixing is difficult at low Reynolds numbers that are characteristic of the microscale, typically requiring long length- and time-scales to allow molecular diffusion between laminar streams. Improving transport and mixing will facilitate the creation of portable inexpensive sensing devices that rapidly detect harmful biological or chemical agents. Passive mixing reliant only on pressure-driven flow through fixed geometries will reduce the complexity of the immunosensor device.

The Stokes equations are solved using the boundary element method to obtain the velocity field within the complex domains. Transport of analyte and antibody is computed with a grid-free particle strength exchange method. Our methods allow optimization of the chamber design through mathematical quantification of mixing, thus reducing the number of devices that must be fabricated for time-consuming experimental trials.

Katharine Hamlington, a graduate student at Biomedical Engineering Department at Tulane University is working on this project. Hideki Fujioka helps to parallelize her boundary element code. The code is written in Fortran90. MPI (Message Passing Interface) libraries are used for the parallelization. ScaLAPACK is used to solve a linear system. The code is able to run on the LONI linux clusters with multiple nodes.

Hideki Fujioka also developed the BEM Surface Builder, a GUI tool to design surface geometries for the boundary element CFD code. Users can design geometries, set the boundary conditions, and export a set of files to run the CFD code. It can import SVG drawing data created by other software such as Adobe Illustrator. Objective-C and Cocoa (Open Step) libraries were used to make an object-oriented code that allows to add functions easily and to provide the GUI interface on the Apple computers.

Presentation/talks

N/A

External Funding

NSF EPS-0701491.

Infrastructure for Accurate and Efficient Binding Affinity Calculations

David Mobley
Chemistry Department
University of New Orleans

Steve Rick
Chemistry Department
University of New Orleans

Shantenu Jha
Department of Computer Science

Louisiana State University

Hideki Fujioka
Center for Computational Science
Tulane University

Project Description

Accurate, reliable simulation-based tools for affinity predictions would transform the process of pharmaceutical drug discovery and enable new kinds of science. Recently, a tool called alchemical free energy calculations has shown considerable promise for predicting binding free energies from simulations. Several studies suggest these calculations could now be useful in practice in drug discovery and other applications, but for the difficulty of setting them up.

This project is to develop a pipeline to set up molecular dynamics simulations and associated alchemical free energy calculations, which will make it possible for these calculations to be more routine, and done with less expert intervention. This project involves three institutions in Louisiana State. The principal and co- investigators are David Mobley and Steve Rick at the University of New Orleans, and Shantenu Jha at Louisiana State University, and the LI Computational Scientist Hideki Fujioka at Tulane University, who helps to program computer codes.

The pipeline is developing by pursuing the following steps; Modify 'mmtools' package to expand existing protein tools to allow user interventions such as specifying protonation states for selected residues (while assigning the rest automatically by 'mcce', http://134.74.90.158/); Interface ligand building/protonation tools with protein/nucleic acid setup tools, to allow easy setup of structures and parameter files for structures, together with ligands, including assigning all hydrogens and building in missing loops; Expand the protein/molecular setup tools to be able to handle nucleic acids as easily as possible; Once the basic functionality is in place, extend these tools to set up MD simulations/free energy calculations which will utilize SAGA ("A Simple API for Grid Applications") on LONI computational resources; Test and Deploy on LONI machines, for greater throughput and resource utilization. This stage will involve test/model calculations of thermodynamic properties such as binding free energies or hydration free energies; Finish research leading to and writing of one or more papers describing the technology development and/or science conducted during the course of this project.

This project has started April 2009. As of June 12th 2009, we are working on developing "interface ligand building/protonation tools".

Presentation/talks

N/A

External Funding

N/A

Automated Data Archiving with PetaShare

Tevfik Kosar
Department of Computer Science
Louisiana State University

Gabrielle Allen
Department of Computer Science
(Louisiana State University

Sumeet Dua
Department of Computer Science
Louisiana Tech University

Frank Loeffler
Center for Computation & Technology
Louisiana State University

Erik Schnetter
Center for Computation & Technology
Louisiana State University

Hideki Fujioka
Center for Computational Science
Tulane University

Project Description

The NSF MRI PetaShare project has provided distributed storage across LONI, and is developing data management tools in collaboration with a wide range of application groups in the state. The basic hardware infrastructure has been deployed, and Kosar's research group has developed a first release of tools.

Data is the biggest current challenge in cyberinfrastructure and computational science. With the PetaShare project, and it close ties to state applications, Louisiana has a chance to make a big impact. This project would have the important consequence of providing automated mechanisms for any application using LONI to archive simulation data both individually or a collaborative group. The strategic implication of this is that we will be able to start building up data archives in the state which will serve as application drivers for a range of further projects e.g. in data mining, information science, visualization, etc.

Hideki Fujioka helps PetaShare users at Tulane University site. Currently, Tom Bishop at Tulane University uses ‘pcommand’ installed his local linux machine, the linux cluster at the center for computational science at Tulane University and LONI linux clusters. He uses a batch script to copy a set of initial configuration data from PetaShare files system to a working disk space at the computing server, and copy back the results to the PetaShare after simulations finish.

The scripts are submitted to PBS queuing system so that all the procedures are done automatically.

Presentation/talks

N/A

External Funding

N/A

Raju Gottumukkala, Ph.D.

NIMSAT Institute, University of Louisiana at Lafayette

I. Research Description

Dr. Raju Gottumukkala is a LONI Computational Scientist at the National Incident Management Systems and Advanced Technologies (NIMSAT) Institute at the University of Louisiana at Lafayette. Raju's primary research interest is in the application of cyberinfrastructure, advanced information technologies and applied mathematics to improve disaster response. NIMSAT Institute is a recently established homeland security and emergency management research facility established at the University of Louisiana at Lafayette.

The NIMSAT institute works closely with several state and national agencies including the Governor's Office of Homeland Security and Emergency Preparedness (GOHSEP) for the state of Louisiana, Department of Natural Resources (DNR), Department of Homeland Security (DHS), FEMA and local emergency management officials in improving the disaster response with advanced information technologies. The primary areas of focus of NIMSAT include cyberinfrastructure enabled tools for disaster management, Critical Infrastructure and Key Resources (CIKR), and Geospatial decision support and resource and information management. Raju currently leads various research and development efforts at the NIMSAT Institute specifically on developing a cyberinfrastructure platform for disaster management and geospatial decision support systems. During hurricanes Gustav and Ike, Raju has helped with providing research and information technology solutions for various response activities of GOHSEP and is involved in preparedness for this year's hurricane season. Raju is also collaborating with various other faculties and guides students at University of Louisiana at Lafayette in various High Performance Computing (HPC) research projects.

II. Projects

1. Parallel-GIS: A High Performance Geospatial Analysis

People Involved

Dr. Raju Gottumukkala, LONI Computational Scientist, UL Lafayette

Dr. Ramesh Kolluru, Executive Director, NIMSAT Institute, UL, Lafayette

Dr. Geoffrey Stewart, Assistant Professor, Moody school of Business, UL, Lafayette

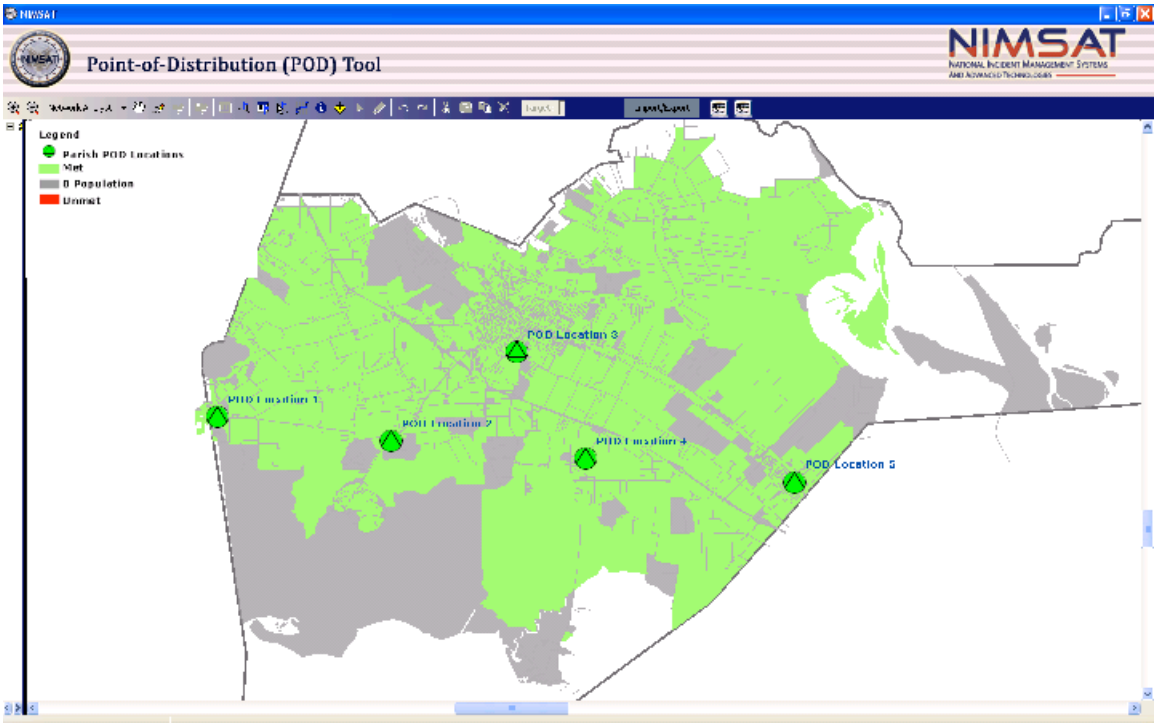
Dr. Baker Kearfott, Associate Professor, Applied Mathematics Department, UL Lafayette

Mr. Christopher Mire, Research Associate, NIMSAT Institute, UL Lafayette

Mr. Haochun Zhang, PhD Student, Applied Mathematics Department, UL Lafayette

Project Description

Geographical Information Systems (GIS) is a major decision support tool that is used in every phase of emergency management. While current GIS products like ESRI's ArcMap and Google Earth provide intuitive interfaces for emergency managers to visualize and tools to analyze spatial data, the increasing growth in size and complexity of spatiotemporal data with the availability of high resolution satellite imagery, real-time situational monitoring tools, and the demand for effective response for a given situation requires a scalable and reliable computational platform to run the spatiotemporal analysis and simulations.



This project would design and develop a scalable and reliable cyberinfrastructure framework for computationally intensive spatiotemporal analysis and decision support applications by leveraging the high-speed networks, distributed supercomputers and distributed supercomputing platform and tools on LONI and LITE. As a part of this project, we configured and installed Geographic Resource Analysis Support System (GRASS), open source geospatial analysis software to run on LONI Dell systems and parallelized various modules of GRASS for various applications. (1) POD (Points of Distributions) are temporary locations for distributing donations of food, water, ice, tarps and other supplies to people in need during emergencies. Setting up a POD is a multifaceted problem involving multiple priorities or objectives in terms of serving the people in demand, allocating POD's in close proximity to people while not affecting the local business recovery. The POD tool provides a GIS based interface and the computationally intensive spatial location analysis algorithm is an MPI algorithm that runs on LONI. This location analysis algorithm takes less than a minute versus several hours on a typical desktop system. (2) Another application that uses the GRASS GIS on LONI is the Critical Infrastructure Analysis and a Key Resources (CIKR) analysis algorithm that uses a network flow based

interdependency analysis on the nation's natural gas pipeline network and natural gas processing facilities to identify the critical interdependencies between various facilities.

2. Reliability Analysis and Resource Management in Parallel Computing Systems

People Involved

Dr. Raju Gottumukkala, LONI Computational Scientist, UL Lafayette

Dr. Box Leangsuksun, Associate Professor, Louisiana Tech University

Reliability and fault-tolerance of large-scale parallel applications is increasingly becoming important because of the growing number of hardware and software components in cluster and grid frameworks. Accurate reliability estimation improves the resource management in large-scale HPC systems and enables users and resource managers to select the appropriate level of fault tolerance for a parallel application and improves the Quality Of Service (QOS) for HPC users.

Dr. Raju Gottumukkala and Dr. Box Leangsuksun have been working on improving the reliability prediction models for applications running on High Performance Computing systems and developing reliability aware resource management algorithms to minimize the performance loss due to failures. This research is a part of the Modular Linux and Adaptive Runtime Support for High-end Computing Operating and runtime Systems (MOLAR), a multi-institution collaborative effort to provide an adaptive, reliable and efficient operating and runtime system solutions for ultra-scale high-end scientific computing on next generation supercomputing systems. This research has improved the reliability predictions models and reliability aware resource management algorithms for HPC systems.

3. Spatial Modeling of the Invasive Nutria Population

People Involved

Dr. Azmy S Ackleh, Professor, Applied Mathematics, UL Lafayette

Dr. Raju Gottumukkala, LONI Computational Scientist, UL Lafayette

Mr. Jay Monte, Masters student, Applied Mathematics, UL Lafayette

Mr. Jessie Castille, PhD student, Applied Mathematics, UL Lafayette

Project Description

This project uses LONI systems to improve the performance of a Nutria population dynamics and movement pattern simulator in order to study the Nutria's effect on the loss of Marsh lands in the Gulf of Mexico. This project was in collaboration with researchers from the USGS Wetlands Research Center. Raju has been working with Dr. Ackleh's research group for developing parallel algorithms and visualization techniques on LONI for the simulator.

The Nutria population dynamics and movement pattern is modeled using differential equations and would take several hours to days to run the simulation for a small patch of land. This LI-project would parallelize the Nutria population simulator using MPI on LONI systems. With the

LI-Scientists assistance, the initial version of the simulator code performed 3.6 faster than the sequential version. The next phase of the project would further improve the simulator performance, identify ways to visualize the simulation results and parallelize various other modules of the simulator.

Most Recent Publications:

1. D. S. Katz, G. Allen, R. Cortez, C. Cruz-Neira, R. Gottumukkala, Z. D. Greenwood, L. Guice, S. Jha, R. Kolluru, T. Kosar, L. Leger, H. Liu, C. McMahon, J. Nabrzyski, B. Rodriguez-Milla, E. Seidel, G. Speyrer, M. Stubblefield, B. Voss, and S. Whittenburg, "Louisiana: A Model for Advancing Regional e-Research through Cyberinfrastructure," Philosophical Transactions of the Royal Society A, v. 367, pp. 2459-2469, 2009.
2. Gottumukkala, N. R., R. Nassar, C.B. Leangsuksun, M. Paun. "Reliability of a system of k nodes for high performance computing applications". To appear in the December 2009 issue of the The IEEE Transactions on Reliability.

Most Recent Workshops/Presentations:

1. N. Raju Gottumukkala, Box Leangsuksun, Raja Nassar, Mihaela Paun, Dileep Sule, "Reliability Aware Optimal-K Node allocation of parallel applications in large scale HPC systems", High Availability and Performance Computing Workshop (HAPCW 2008), Denver, Colorado.
2. Raju Gottumukkala, Ramesh Kolluru, "Improving Disaster Response: NIMSAT", The 2009 gulf Coast Marine Conference
3. Raju Gottumukkala, Rusti Liner, "GIS Projects at NIMSAT Institute" The 25th Annual Remote Sensing and GIS Workshop, April 14-16 2009, Louisiana

External Funding:

1. Department of Natural Resources, "Intelligent Flood Protection Monitoring, Warning and Response System", Under Review, 2,891,000 (347K subcontract as Partner Institute)

Outreach Activities:

1. Mentoring the following students at UL, Lafayette in various LONI/ HPC Projects:
    a. Haochun Zhang, PhD Student, Department of Applied Mathematics, University of Louisiana at Lafayette
    b. Jessie Castille, PhD student, Department of Applied Mathematics, University of Louisiana at Lafayette
    c. Jay Monte, Master's student, Department of Applied Mathematics, University of Louisiana at Lafayette
2. Organized the LONI Workshop at University of Louisiana at Lafayette for 2008-2009 and gave an introductory presentation on using the TeraGrid.

2008 — 2009 Annual Report from Dr. Shizhong Yang

BioInformatics and Computational Material Research at Southern University

Faculty:

Shizhong Yang

Ebrahim S. Khosravi, Shuju Bai, Rachel Vincent-Finley, and Nigel Gwee

Graduate Student:

Kimberlee Lyles, Kiante Roberson, Charles Shropshire, Sadque Mohammed, Houman Kamran, Swetha Bodla, Murali K. Gangineni, Christopher Clayton

Undergraduate:

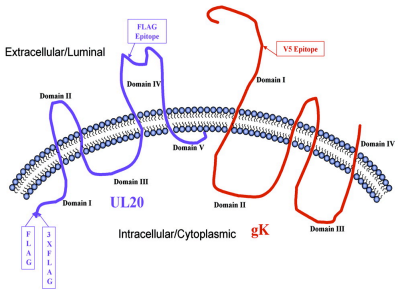
Laura Hurst, Christina Bias

Computer Science Department, Southern University

2. Project Description

(1). gK and UL20 Protein Structure and Protein-Protein Interaction (Senior Researcher)

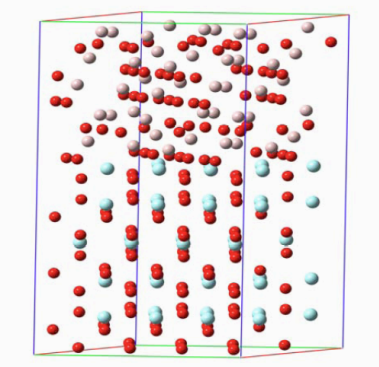
Understanding the interaction and structures of gK and UL20 could greatly facilitate predicting functional domains of each protein that may be involved in multiple functions through the virus lifecycle. In this project, Dr. Shuju Bai, the graduate student Kimberlee Lyles, Dr. E. Khosravi and Dr. Yang are collaborating with Dr. G. Kousoulas group at LSU School of Veterinary Medicine, using related computational tools to predict the gK and UL20 protein-protein and Nano-particle/Membrane Interactions. This is a two year $50,000.00 /yr LBRN project. Dr. Bai is the principal investigator and Dr. Khosravi and Dr. Kousoulas are the mentors.



Viruses gK, UL 20 and membrane

(2). A Study on New Highly Reflective Thermal Barrier Coatings (Co-PI)

Thermal barrier coatings (TBCs) are used in the hot section of rocket engines and jet engines to safeguard the engines under extreme working temperature, ~1600ºC. In this project, Southern researcher, Dr. Yang and group members is doing computer simulation on ZrO_2/Al_2O_3 interface, while Dr. Guo's research group at LSU is doing plasma spray experiment in collaborate with NASA scientists. This project was funded by NASA EPSCoR – BoR for one year with total budget $61125.00 /yr (Southern was allotted $22250.00).



Al_2O_3/ZrO_2 interface (TBC materials)

(3). *Ab Initio* and Experimental Study of A Novel Nano Ceramic Thermal Barrier Material (PI)

This is a NASA-BoR supported one year $29,905 project that combined theoretical modeling with experimental design and aiming to develop an efficient thermal barrier coating material. This project was started from May 2009.

(4). SU related proposal support (new methods, materials and algorithms developing).

Dr. Yang collaborated with SU CEES and grant office submitted several proposals doing research in high performance computing, data processing and visualization to NSF, NASA etc. as a Co-PI. (Details will be listed in part 5.)

(5). SU LONI internal project support. (Co-PI)

SU Computer Science proposed and was approved 4 projects at the end of 2009. They are reported separately.

3. Publications:

(1). Shizhong Yang, S.M. Guo, Guang-Lin Zhao, and Ebrahim Khosravi, " High infrared reflective nickel doped ZrO_2 from first principles simulation", ICCS May 2009. (international conference paper)

(2). (peer-reviewed and was accepted to be published): "Doped C_{60} study from first principles simulation", New3SC- 7, (Seventh International Conference on New Theories, Discoveries and Applications of Superconductors and Related Materials), Beijing, May 2009.

(3). Wendong Wang, Zhenjun Wang, Jinke Tang, Shizhong Yang, Hua Jin, Guang-Lin Zhao, and Qiang Li, " Seebeck coefficient and thermal conductivity in doped C_{60}", Journal of Renewable and Sustainable Energy, vol. **1**, issue 2, 23104, page1~8 (2009);

(4). G.L. Zhao, S. Yang, D. Bagayoko, J. Tang and Z.J. Wang, **"** Electronic structure of C60 semiconductors under controlled doping with B, N, and Co atoms", Diamond & Related Materials, vol. **17,** page749~752 (2008).

4. Presentations/Talks:

(1). Presentation at 2009 LAS 83rd annual conference: "First principles molecular dynamics simulation of nano gold adsorption on (0001) surface of Ruthenium", Shizhong Yang, Shuju Bai, Ebrahim Khosravi, and Guang-Lin Zhao.

(2). Invited talk: "Doped C_{60} study from first principles simulation", New3SC- 7, (Seventh International Conference on New Theories, Discoveries and Applications of Superconductors and Related Materials), Beijing, 2009.

5. External funding:

(1). Funded/approved:

Part 2. (1) ~ (3) and (5).

(2). Pending:

a. NASA pending "Novel Nano-Structured Thermal Barrier Coatings" Co-PI;
b. NSF pending "Nano Ceramic Thermal Barrier Material: Design and Fabrication" Co-PI;
c. NSF pending "Predictive Quantum Computation and Design of the Catalysts for Green Energy Applications" Co-PI;
d. NSF pending "Minority Serving Institutions Solar Energy Research Consortium" Co-PI;
e. NSF pending "Sensor Arrays and the Interpretation of Multi-scale Data Sets" Co-PI.

6. Summary:

Dr. Yang is working on three research areas: BioInformatics, Computational Material Simulation, and High Performance Computing Methods.

Besides SU faculty, I collaborated with the following faculty in the past academic year: Dr. Gus Kousoulas (LSU Vet School), Dr. Newcomer(LSU Biology), Dr. S. M. Guo(LSU Mechanical Engineering), Dr. Kun Lian (LSU CAMD), Dr. John Wefel(LSU Physics), Dr. Jinke Tang (U. of Wyoming Physics), Bingqing Wei(U. of Delaware ME).

I also organized the 2009 first LONI workshop which is sponsored by CS Deportment. More than 30 people attended the whole workshop. More than 50 students There 7 faculty and 2 graduate students running jobs on the LONI supercomputers. I am the default SU TeraGrid SU Campus Champion helping SU TeraGrid user to solve the related research issues.

All of my current research needs to use LONI and TeraGrid machines.

I visited Oak Ridge National Lab together with other 11 SU faculty. I also joined a TeraGrid workshop at UIUC and had close contact with other researchers in my research area.

My current research will support SU Computer Science, Mechanical Engineering, Physics Department, SU CEES center's funded/proposed projects. I hired and trained three graduate students working on my funded projects. They have gained skills to run jobs on LONI machines. The students working in my projects are: (1). Kiante Roberson: graduate student (1^{st} semester), Computer Science Department, minornity(black); (2). Sadque Ali Mohammed: graduate student (2^{nd} year), Computer Science Department International (India); (3). Charles alphonce shropshire: graduate student (1^{st} semester), Computer Science Department, minornity (black); (4). Kimberlee Lyles: graduate student (1^{st} semester), Computer Science Department, minornity(black).

Title: LI Computational Scientist

Name: Zhiyu Zhao

Institutional Affiliation: Department of Computer Science, College of Sciences, University of New Orleans

Roles and Specializations: Please see Project Description

Research Fields: bioinformatics algorithms (protein structure comparison, genome sequence comparison, haplotype reconstruction, gene expression data analysis, etc.); parallel and distributed algorithms

Project Description

1. A Parallel Protein Structure Alignment Tool and a Shared Feature Database for Structures in the Protein Data Bank

Role: PI

Collaborator: Dr. Christopher Summa, Department of Computer Science, University of New Orleans

Graduate Student: TBD

Description:

The research of proteomics has many biological applications. Proteomics research topics are usually related to protein sequences and structures. While both have close relationship with proteins' biological functions, structures reveal more evolutionary information than sequences do, since the structure of a protein changes more slowly in evolution than does its sequence. Also, researchers frequently find that proteins with low sequential similarity are structurally homogenous. Therefore it is particularly important to discover the structural similarity / dissimilarity among different proteins. The research of protein 3D structure similarity provides fundamental and very helpful tools for many biological research topics.

Result from protein structure determination techniques, the number of proteins discovered by biologists has increased dramatically over the last 30 years. The rapid growth of the Protein Data Bank (PDB) necessitates the development of efficient and accurate protein structure comparison and searching algorithms and automatic software tools.

We have developed a Self-Learning and Improving pairwise Protein Structure Alignment (SLIPSA) algorithm. SLIPSA is a feedback algorithm for protein structure alignment that uses a series of phases to improve the global alignment between two protein backbones. Based on a large set of proteins collected from various publications for diverse testing purposes, we have compared our algorithm with three other commonly used methods: CE, DALI and SSM. The results show that in most cases our algorithm is more accurate than those well-known methods that have been tested.

The SLIPSA algorithm is implemented with MATLAB and we have developed a web portal based on it (see http://fpsa.cs.panam.edu/). Due to the large size of the PDB and high complexity of current protein structure alignment algorithms, an alignment can be very time-consuming and computation capability of machines greatly affects alignment performance in terms of both speed and accuracy. Since our current tool is just a proof-of-concept system written with MATLAB, there is a lot of space to improve its speed performance by (1) rewriting the code with C/C++, (2) taking advantage of parallel and distributed computation power of high performance computational resources, and (3) design an efficient protein database to store as much as possible offline information to reduce the execution time used by repetitively retrieving and calculating information from original protein data files.

We are migrating our MATLAB code onto the LONI and TeraGrid by completing the following tasks: (1) rewrite all the code with serial C++ (to be completed in June 09), (2) use OpenMP or p-thread to implement a parallel version of the code, and (3) based on the parallel version, use MPI to implement a parallel and distributed version and deploy the code on Queen Bee. In the fall 09 we will also start working on the following tasks: (4) develop a protein feature database for proteins in the PDB and deploy it on Queen Bee, (5) use LONI's PetaShare data storage and management tools to provide a shared directory for protein files that we have downloaded from the PDB repository, and (6) develop a protein file and database updating program to synchronize our database and protein files with those in the PDB repository. The database will hold protein features, both sequential and structural, for proteomics related research topics. The database and all the downloaded PDB files will be shared and updated regularly so that LONI and TeraGrid users will have free access to it. The protein features in the database will be expanded upon future requests from world-wide research groups and scientific application developers in need of a comprehensive and efficient protein feature database.

We have applied for and been awarded grant by the TeraGrid Pathways Fellowship Program to support a graduate student of Dr. Summa for the fall semester to work on this project, which is likely to become the student's graduate thesis project. Drs. Summa and Zhao will be jointly supervising the student.

2. Predicting Proteins in SCOP Classification via Alignment and Threading

Role: Co-PI

Collaborator: Dr. Bin Fu, Department of Computer Science, University of Texas-Pan American

Graduate Students: TBD

Description:

A protein's function is closely linked to its 3D structure. As an increasing number of protein structures become known, the demand for algorithms to analyze 3D conformational structure increases as well. With the exponential growth in the number of newly-discovered protein structures, the view of the protein universe is constantly changing. In order to understand the functions of proteins and their relationships to each other, the classification of proteins should be updated frequently. The Structural Classification of Proteins (SCOP) database is built by labor-intensive visual inspection. On the other hand, automated classification schemes have the advantage that the view of the protein universe can be updated frequently to include newly-discovered protein structures in a timely manner.

Protein classification in the SCOP will be predicted by using 3D comparison for known 3D structures and protein threading model for known sequences with unknown 3D structures. In order to deal with the challenging computational problem in protein 3D prediction, we plan to develop a more efficient separator theory which will further improve our recently developed width bounded separator theory. It will be suitable for deriving algorithms for the protein threading problem, which is NP-hard.

The planned research work is expected to make fundamental contributions to the existing research in the field of 3D structure of proteins, while the implementations will demonstrate the practical benefits of the proposed work. Our existing web server will be improved to provide

more matured protein SCOP prediction service to the Bioinformatics community. Education will be an important part of this project. This research will bring both fundamental theoretical problem and software system development. They will be suitable for training both undergraduate and graduate students at multiple levels.

Drs. Fu and Zhao are working on this project by developing efficient computational models and algorithms to solve the proposed problems, and we have submitted a project proposal to the NIH to request support for this research. If the proposal is approved, we will provide financial support to and supervise graduate and undergraduate students to develop the proposed software system. Dr. Zhao will also be responsible for the parallel and distributed implementation of the software by taking advantage of the HPC resources on the LONI.

Publications (07/01/08 - 06/31/09)

1. Zaixin Lu, Zhiyu Zhao, Sergio Garcia, Krishnakumar Krishnaswamy, and Bin Fu, "Search Similar Protein Structures with Classification, Sequence and 3-D Alignments", to appear in the Journal of Bioinformatics and Computational Biology.

2. Huimin Chen and Zhiyu Zhao, "An Information Theoretic Viewpoint on Haplotype Reconstruction from SNP Fragments", to appear in the 3rd International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2009, China).

3. Zaixin Lu, Zhiyu Zhao, Sergio Garcia, and Bin Fu, "New algorithm and web server for finding proteins with similar 3d structures", in the Proceedings of the 2008 International Conference on Bioinformatics & Computational Biology (BIOCOMP'08, USA), pp. 674 - 680.

Presentations / Talks (07/01/08 - 06/30/09)

1. 04/14/09: Zhiyu Zhao (author & presenter), "Intermediate MATLABhttp://www.hpc.lsu.edu/training/tutorials/presentations/Intro-MATLAB-0309.pdf", Stanley Thomas Hall, Tulane University; An invited tutorial session of the LONI HPC Workshop, Spring 09, hosted by the Tulane University and open to all the LI research community, see http://www.hpc.lsu.edu//training/20090413/index.php.

2. 03/16/09: Zhiyu Zhao (author & presenter), "Introduction to LAPACK", Liberal Arts Building, UNO; A tutorial session of the LONI HPC Training, Spring 09, open to all the LI research community via UNO's Access Grid facilities, see http://www.hpc.lsu.edu/training/tutorials/index.php#spring09lapack.

3. 03/02/09: Zhiyu Zhao (author & presenter), "Introduction to MATLAB", Liberal Arts Building, UNO; A tutorial session of the LONI HPC Training, Spring 09, open to all the LI research community via UNO's Access Grid facilities, see http://www.hpc.lsu.edu/training/tutorials/index.php#spring09matlab.

4. 02/06/09: Zhiyu Zhao (author & presenter), "Introduction to the Supercomputing Resources at LONI & TeraGrid", Math Building, UNO; A presentation open to all the UNO research community as required by the chair of the Department of Computer Science, see http://www.cs.uno.edu/special/seminars.xml#Introduction%20to%20Supercomputing%20Resources%20at%20LONI%20and%20TeraGrid and http://www.cs.uno.edu/~sylvia/LONI&TeraGrid.pdf.

5. 01/29/09: Zhiyu Zhao (author & presenter), "Protein 3D Structure Alignment and Searching for Similar Structures in the Protein Data Bank", Engineering Building, UNO; A seminar talk invited by the Department of Electrical Engineering, UNO and open to all the EE faculty/staff and students, see http://www.cs.uno.edu/~sylvia/ProteinStructure.pdf.

6. 11/21/08: Zhiyu Zhao (author & presenter), "Feedback Algorithm and Web-Server for Protein Structure Alignment", CERM Building, UNO; A seminar talk invited by the Department of Computer Science and open to all the CS faculty/staff and students, see http://www.cs.uno.edu/special/seminars.xml#Feedback%20Algorithm%20and%20Web-Server%20for%20Protein%20Structure%20Alignment and http://www.cs.uno.edu/~sylvia/SLIPSA.pdf.

7. 10/31/2008: Scott Whittenburg (author) and Zhiyu Zhao (author & presenter), "Computational Research at UNO", LSU Union, LSU; A presentation required by Dr. Whittenburg (vice chancellor of research at UNO) and open to all attendees of the LI All Hands Meeting '08, see http://institute.loni.org/FirstAllHandsMeeting.php.

8. 10/28/08: Zhiyu Zhao (author & presenter), "Research on Protein 3-D Structure and Genome Sequence Related Problems", Liberal Arts Building, UNO; A talk invited by the Director of the University Honors Program and open to all the UNO honors students of fall 08, see http://www.cs.uno.edu/~sylvia/Protein&Genome.pdf.

9. 10/02/08: Zhiyu Zhao (author & presenter), "Linear Time Probabilistic Algorithms for the Singular Haplotype Reconstruction Problem from SNP Fragments", Engineering Building, UNO; A seminar talk invited by the Department of Electrical Engineering, UNO and open to all the EE faculty/staff and students, see http://www.cs.uno.edu/~sylvia/HapRec.pdf.

External Funding (07/01/08 - 06/30/09)

1. The TeraGrid Pathways Fellowship Program.

A project proposal entitled "A Parallel Protein Structure Alignment Tool and a Shared Feature Database for Structures in the Protein Data Bank" was submitted in Feb 09 and approved in Mar 09. Awarded $6,075 to support a student at UNO in the fall 09 semester to develop a parallel protein structure alignment program and a protein feature database under the supervision of the PI (Dr. Zhao), and $2,000 if the PI is going to attend the TeraGrid '09 conference in June (Note: the PI will not be able to attend the conference due to her anticipated baby delivery in June).

2. Innovations in Biomedical Computational Science and Technology (R01), NIH PAR-07-344.

A project proposal entitled "Predicting Proteins in SCOP Classification via Alignment and Threading" was submitted in Feb 09. $431,720.00 total direct and indirect costs were proposed for the entire project period of three years ($245,720 on Dr. Fu's part and $186,000 on Dr. Zhao's part to support the proposed research and the study of three graduate students (two at UTPA and one at UNO)).

Appendix C

LI Projects with Updates

Project 1. Infrastructure for Accurate and Efficient Binding Affinity Calculations

David Mobley (UNO), Steve Rick (UNO), Shantenu Jha (LSU)

Our goal is methods for predicting binding strengths, or binding free energies, between biomolecules. Accurate, reliable simulation-based tools for affinity predictions would transform the process of pharmaceutical drug discovery and enable new kinds of science.

Recently, a tool called alchemical free energy calculations has shown considerable promise for predicting binding free energies from simulations. Several studies suggest these calculations could now be useful in practice in drug discovery and other applications, but for the difficulty of setting them up [1,2].

We propose to further develop a pipeline to set up molecular dynamics simulations and associated alchemical free energy calculations, which will make it possible for these calculations to be more routine, and done with less expert intervention. We will implement tools to automate steps that now require expert intervention, removing the bottleneck and allowing widespread application of these tools using LONI computational resources. The pipeline will take an arbitrary protein of interest, begin with a protein data bank structure of the protein, assign reasonable protonation states to residues, including titratable residues, then build in missing loops and residues – those which may not be resolved in crystal structures. Another component will build and protonate ligands, either from the PDB or from names/2D structures. The pipelines will merge with docking of the ligands to the protein structures to generate candidate ligand bound orientations. The system will then be placed in solvent and prepared for alchemical free energy calculations on a molecule of interest in the system. Current simulation packages encapsulate aspects of these tools, but have trouble with ligand parameterization, and titratable residues are assigned default protonation states; missing residues and loops must be built in separately. There are standard procedures for these steps but there is currently no pipeline.

Both the Mobley and Rick labs are interested in binding of small molecules to proteins. The Mobley lab invests significant resources in studying ligand binding to model and drug binding sites [2], and the Rick lab continues studying binding of water to proteins, and its influence on ligand binding strengths [3]. The proposed pipeline will benefit these efforts, as well as other work in these labs.

Recent computational advances provide new insight into conformational flexibility of riboswitches induced by small molecules, a topic of specific interest to the Jha's group (in combination with the Aboul-ela lab). The ability to compute the free-energy of binding efficiently and accurately for distinct, yet similar small-molecules to riboswitches will be a major complement to existing research efforts and capabilities. It is worth mentioning that RNA based drug-discovery holds a special promise for the drug industry [5]. This is conceptually similar to the binding problem but may require customizations of the pipeline.

Because of the interests of all of these groups in binding, and the use of similar techniques, there is substantial overlap – both in terms of infrastructure needed, and the science being done. We believe the proposed project will also facilitate collaboration between our groups and reduce redundant efforts in the different groups. Additionally, all of the groups will benefit

from the help of a staff scientist to make these calculations take better advantage of the available LONI computational and data-management capabilities (such as Peta-share), and even extend the workflows to work seamlessly across the multiple LONI computational facilities.

This project will benefit LONI by aiding at least two LONI investigators with needs for infrastructure in this area; making these tools available also will make these simulations more accessible to others. The free energy approach is quite general (as evidenced by its applications here to diverse systems [2, 3, 5]. The pipeline proposed here also includes aspects that are common to most biomolecular simulations, so components can be adapted to benefit an even broader audience. This work also fits well with the goals of the state as a whole -- Louisiana is investigating significant resources in growing the biotechnology industry. Long-term, expansion in this area may interest the biotech/pharmaceutical industry and tie in with statewide emphasis on biotech.

We already have invested significant resources [2, 5] in these tools, so turning them into a pipeline involves linking components and filling in gaps. We anticipate that the proposed project would require 6 months of time from a qualified staff person in order to make it sufficiently general that it can be of use to others.

[1] C. Chipot et al., J. Comp. Aided. Mol. Design 19: 765-770 (2005).

[2] D. Mobley et al., J. Mol. Biol. 371: 1118-1134 (2007).

[3] L. R. Olano et al., J. Am. Chem. Soc. 126: 7991-8000 (2004).

[4] Laying the Groundwork for Drug Design Targeted at RNA, http://lbrn.lsu.edu/portal/cw_registration/presentations/Fareed_LONI_408.pdf

[5] http://www.nytimes.com/2008/11/11/science/11rna.html

Milestones:

Note that the timelines here assume 50% time on this project for 1 year.

1. (1 month) Familiarize self with mmtools package and expand existing protein tools to allow user interventions such as specifying protonation states for selected residues (while assigning the rest automatically)

2. (2 months) Expand the protein/molecular setup tools to be able to handle nucleic acids as easily as possible

3. (3 months) Interface ligand building/protonation tools with protein/nucleic acid setup tools, to allow easy setup of structures and parameter files for structures, together with ligands, including assigning all hydrogens and building in missing loops.

4. (1.5 months) Once the basic functionality is in place, extend these tools to set up MD simulations/free energy calculations which will utilize SAGA ("A Simple API for Grid Applications") on LONI computational resources.

5. (1.5 months) Test and Deploy on LONI machines, for greater throughput and resource utilization. This stage will involve test/model calculations of thermodynamic properties such as binding free energies or hydration free energies.

6. (3 months) Focus on deliverables: Finish research leading to and writing of one or more papers describing the technology development and/or science conducted during the course of this project.

Update:

Hideki recently completed the first milestone, which was familiarizing himself with the mmtools package and expanding the existing protonation tools to allow some more user input such as specifying protonation states for selected residues while assigning the rest automatically.

For technical reasons, we decided to change the order of milestones 2 and 3, so now he is working on milestone 3 (interface the protein building/ligand parameterization tools to allow easier setup of structures).

Project 2. Spatial Modeling of the Dynamics of Invasive Nutria

Azmy S. Ackleh
Department of Mathematics
University of Louisiana at Lafayette
Lafayette, LA 70504-1010

ackleh@louisiana.edu

Nutria are large beaver-like rodents, whose population is directly contributing to loss of marsh lands in the gulf coast. In order to develop new methods to restore damaged wetlands and control nutria, it is important to understand the behavior of nutria. Nutria moves from one patch to another depending on several factors including food availability. When nutria reaches high density in a particular patch it often consumes all the plants in that patch and converts it to water patch.

We have developed a MATLAB code for modeling nutria population dynamics in a 2-dimensional geographic region (see Figure 1). This code is currently used by scientists at the USGS National Wetlands Research Center to understand the impact of nutria population on wetlands. The current MATLAB code divides a given region into discrete patches. In each patch there are three difference equations that describe how the nutria population in that patch grows. The current code is extremely slow and takes on the order of one to two days to simulate a reasonable size geographical region. If we have to simulate on the order of 10,000 patches (which is a typical simulation) then one is solving 30,000 difference equations at each times step in addition to the rules that describe the movement between patches.

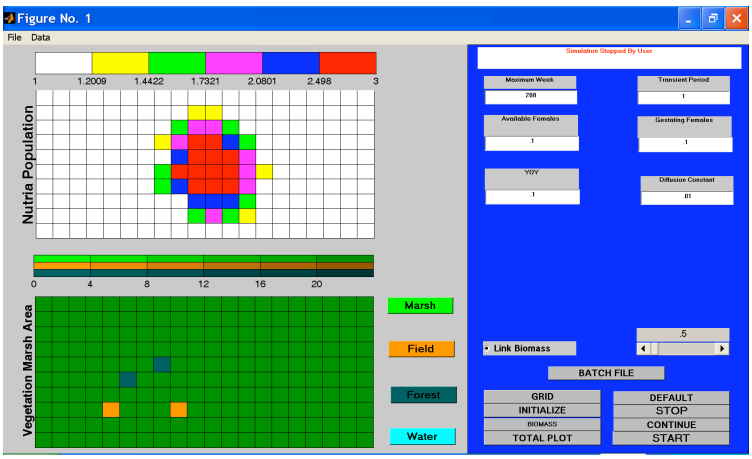


Figure 1: Current MATLAB model for nutria population dynamics

The purpose of this LONI based project is to parallelize the simulation of nutria behavior into MPI from the current MATLAB code and use LONI systems to reduce the time from the order of days to a few minutes. Recall that MATLAB is an interpreter thus it is naturally slow. My graduate student Jay Monte is working on converting the current MATLAB code to a C++ code.

However, he is not familiar of how to take a C++ code and convert into a parallel code that runs on a multi-processor machine. Thus, we request 20 hours during Fall 2008 of our LONI Computational Scientist Dr. Raju Gottumukkala to assist Jay for parallelizing this code.

Update:

People Involved: *Azmy S Ackleh, Raju Gottumukkala*

This project uses LONI to implement a parallel version of Nutria population dynamics and movement pattern simulator in order to study the Nutria's effect on the loss of Marsh lands in the Gulf of Mexico. This project was in collaboration with researchers from the USGS Wetlands Research Center.

The Nutria population dynamics and movement pattern is modeled using difference equations and the previous version that was written in Matlab would typically take several hours to days to run the simulation for a small patch of land. This LI-project would parallelize the Nutria population simulator using MPI on LONI systems, which would drastically increase the interpretation of research results to analyze the effect of Nutria on the Marsh Lands in the Gulf of Mexico.

Progress:

We have developed a parallel version of the Nutria population dynamics and movement pattern simulator using MPI. The simulator running on LONI Dell clusters performs 3.6 times faster than the sequential version. We are further improving the parallel algorithm to handle very large datasets. Also, we are investigating various approaches and tools to visualize the simulation results and improve the performance of the simulator to handle large data sets.

Milestones & Timeline:

1. 07/30/2009 Improve the current version of the simulator to handle large areas
2. 09/15/2009 Develop an approach to visualize and store simulation results from multiple scenarios to observe the Nutria's population dynamics over time.
3. 12/31/2009 Develop approaches for identifying the patterns of nutria impact on the Marsh lands using areal imagery

Project 3. Coupling LONI Institute Computational Scientists, CyberTools and Science Drivers at the Molecular Level

PI: Thomas C. Bishop (CCS, Tulane University), co-PI: Shantenu Jha (CCT, LSU), Senior Investigator: Nayong Kim(CCT, LSU),

Requested LI Computational Scientists (CSs) Time: 6 months FTE-months

Molecular dynamics(MD) simulations can now be considered a mature computational methodology. Rather than validating force fields, testing algorithms or optimizing single run parallelization the focus has shifted to interpretation and application. For this reason many simulation studies now include not one or a few simulations but whole ensembles. Replica exchange (RE) simulations also called parallel tempering is one such example. In RE-MD many copies of the same system are run simultaneously, but at different temperatures, and allowed to exchange information. This greatly enhances conformational sampling.

Several LONI institute projects utilize simulation ensembles. In case of the environmental biosensor project, Bishop's contribution is to computationally optimize the ligand-antibody interactions[1]. For this purpose we seek not only to predict structures of the antibodies, but to utilize in silico mutation analysis to help direct experimental efforts. For a given antibody structure an exhaustive study of all possible point mutations in the loop region requires over 1000 simulations. Biologic or experimental constraints reduce the set of mutations to be considered to ~100. For each ligand-antibody system a nanosecond of a replica exchange simulation with 16 replicas requires ~4days of run time on 32CPUS with Amber 8. We expect that with suitable simulation and data management tools that 100 mutations can be simulated and analyzed in less than 2months. Without suitable simulation management tools this many simulations is a user intensive task that simply cannot be accomplished.

In an NIH funded study (R01GM076356) Bishop's goal is to investigate sequence dependent variations in nucleosome stability. Since nucleosomes are the fundamental structural unit of chromatin these variations in stability potentially effect all genomic processes. Nucleosomes can be formed from any 146 basepair segment of DNA but to date all available x-ray structures of the nucleosome have utilized nearly the same 146bp sequence of DNA. In order to investigate sequence dependencies, we have developed a combination of coarse-grain sequence selection techniques and all atom molecular modeling techniques that allow us to rapidly assemble 1000s of individual nucleosomes for simulation and analysis. Each nucleosome system contains approximately 150,000 atoms and requires 4ns of simulation time to equilibrate using traditional simulation techniques [2]. This equates to 2.4 days of run time on 32 CPUs using NAMD2.6. With suitable simulation management tools we estimate that we can readily conduct 100 such simulations in the LONI environment in the course of 2 months. This is sufficient throughput to simulate a collection of 84 sequences of DNA that have been demonstrated via experiment to span the range of known nucleosome stabilities.

In both of the above cases replica exchange techniques or a more generalized exchange technique that allows for sequence exchanges as well as coordinate and velocity exchanges could be implemented to speed up the conformational search process. Even without such advances the singular obstacle to conducting the above simulations is the amount of user intervention required

to manage the simulations. To achieve the above scientific objectives requires efficient and coordinated utilization of the entire LONI system.

There have been interesting and important advances to develop a framework for an adaptive scalable framework for replica exchange simulations[3]. We propose to use and to enhance this framework in order to achieve the above scientific objectives and it is for this purpose that we are explicitly requesting LI CS time for assistance with the set-up, deployment and execution of this framework on all LONI sites/machines. Specifically, we are requesting time for: i) Assistance with integrating the applications with the framework, and ii) Assistance with the set-up, deployment and execution of this framework on all LONI sites/machines. The overall goal is optimization of LONI resources in order to achieve a lower time-to-solution for well defined scientific objectives by effectively managing distributed resources.

This effort will provide an important/critical coupling between the LI and Cybertools projects and represents collaborations between different LI partner universities, important couplings of expertise, and an excellent test bed and application scenario for important algorithmic and infrastructural developments (the framework). Our efforts will enhance currently funded efforts, form the basis of an exploratory grant and a future Cyber-enabled Discovery and Innovation proposal to be submitted in 2009. (see http://www.nsf.gov/crssprgm/cdi/)

References:

1. Identification of important residues in metal-chelate recognition by monoclonal antibodies. B. Delehanty, R.M. Jones, T.C. Bishop and D.A. Blake, Biochemistry 2003.

2. Molecular Dynamics Simulation of Nucleosomes and Free DNA. T.C.Bishop, J.Biomolec. Struct. Dyn. 2005.

3. Adaptive Distributed Replica-Exchange Simulations, A. Luckow, S. Jha, J. Kim and A. Merzky. Accepted for publication, Philosophical Transactions of the Royal Society

Update:

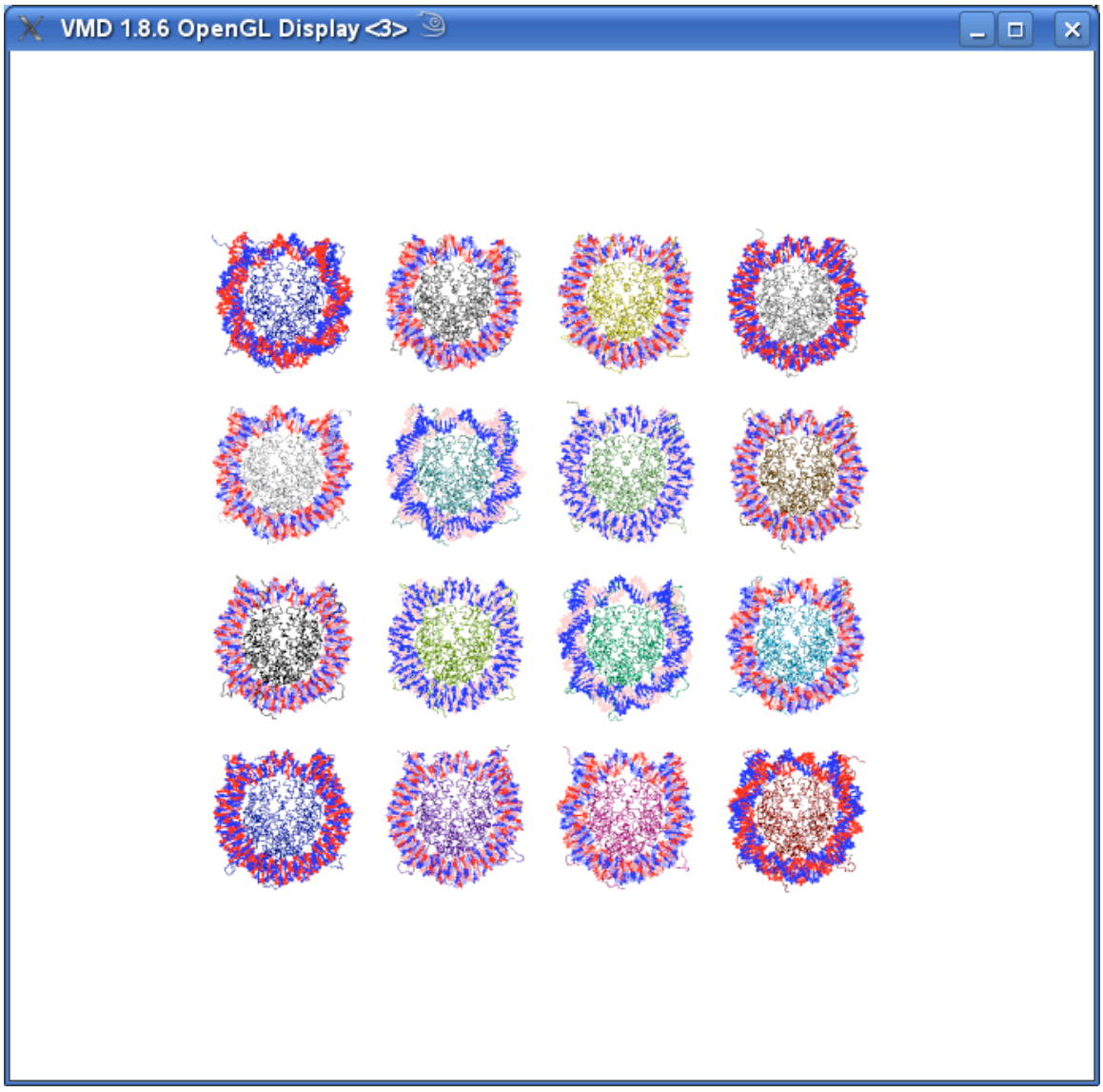
With the support of LI CSs Hideaki Fujioka we have adapted our molecular dynamics simulation workflow tools originally developed with Tevfik Kosar to utilize Petashare as part of the data management component of our computational workflow.

We were thus able to accumulate a total of 128ns of nucleosome dynamics by concurrently simulating 16 nucleosome systems. The total time to completion was approximately 10days and utilized as many as 1000 LONI processors scattered across 5 different sites. Each system consists of ~200,000 atoms and requires approximately 15hrs of run time on 64 CPUS to generate a 1ns trajectory. We have generated 8ns for each system. The combined data set is 640Gb and the total SU utilization was nearly 125,000SU.

The 16 simulations provide insight into the structure and dynamics of every possible dinucleotide step at each of the 146 possible locations in the nucleosome. The results are being prepared for presentation at "Albany 2009: The 16th Conversation June 16-20" The final day of

this international conference includes A special workshop on nucleosome positioning. Bishop is one of the co-organizers of the workshop.

The next phase will be to implement adaptive scheduling to achieve an even shorter time to completion for our computional tasks. More advanced simulation techniques, namely replica exchange MD, will also be incorporated to enhance sampling.



Project 4. Automated Data Archiving with PetaShare

PIs: Tevfik Kosar (LSU), Gabrielle Allen (LSU), Sumeet Dua (LaTech), Frank Löffler (LSU), and Erik Schnetter (LSU); LI CS: Hideki Fujioka (Tulane).

Project Description: The NSF MRI PetaShare project has provided distributed storage across LONI, and is developing data management tools in collaboration with a wide range of application groups in the state. The basic hardware infrastructure has been deployed, and Kosar's research group has developed a first release of tools.

In this proposal we request the assistance of a LONI Computational Scientist in (i) working with the LONI HPC team to deploy and test the petashare software over all LONI resources; (ii) assist in troubleshooting system problems arising as application groups are starting to use petashare; (iii) developing scripts that can integrate with the LONI queuing systems to make it easy to automatically archive simulation data with basic metadata descriptions into petashare. This project would integrate with the CyberTools NSF project, where postdoc Frank Löffler is already looking at using petashare for archiving data from large scale numerical relativity simulations.

Effort Requested and Involvement of Computational Scientist: We estimate that this would require 3 to 6 months of an FTE.

Benefit to LONI Institute: Data is the biggest current challenge in cyberinfrastructure and computational science. With the PetaShare project, and it close ties to state applications, Louisiana has a chance to make a big impact. This project would have the important consequence of providing automated mechanisms for any application using LONI to archive simulation data both individually or a collaborative group. The strategic implication of this is that we will be able to start building up data archives in the state which will serve as application drivers for a range of further projects, e.g. in data mining, information science, visualization, etc.

Project 5. Developing a High Performance Computational Biology and Material Science Lab at Southern University (HPC-BMSL)

PIs: Ebrahim Khosravi (SUBR), Shuju Bai (SUBR), Rachel Vincent-Finley (SUBR), Shizhong Yang (SUBR); LI CS: Shizhong Yang (SUBR).

Southern University and A&M College proposes to establish a High Performance Computational Biology and Material Science Lab (HPC-BMSL) focused on the high performance computing, bioinformatics, and nanomaterial simulation. The objectives of the project are, (a) developing novel high performance computation algorithms and methods to facilitate atomic level molecular dynamic simulations; predict secondary and ternary structure of proteins, protein docking, to understand life process and assist drug design; understanding and predicting the electronic, optical, magnetic, and structural properties of the selected novel electronic materials; (b) to provide an infra-structured platform for systematically mentoring and training of under-graduate, graduate students, and post-doctors at Southern University and A & M College; (c) to attract talented graduate faculties to SU and promoting and enhancing the interdisciplinary collaborating among SU campus(Computer Science, Education, Mathematics, Physics, Chemistry, Biology, EE and ME, CEES center) and with LBRN, LaSPACE, LONI supported six research universities(LSU ME, Louisiana Tech. material science, and other four campuses), and industries(pharmacy, NASA, and green energy related chemical engineering).

The four sub-projects will synergistically address complementary tasks to dramatically enhancing our fundamental knowledge and practical applications in the biophysics, biochemistry, drug design, and nano-size material science. The titles of the research subprojects are: (1). *A novel reduced coordinate space method for molecular dynamic simulations*, by R. E. V. Finley and S. Yang; (2). Developing algorithms for predicting the *Secondary and tertiary structure of proteins and modeling protein docking and interaction*, by S. Bai, E. S. Khosravi, and S. Yang; (3). *MUSIC---a LaSPACE sub-project for NASA research*, by S. Yang, E. S. Khosravi; and (4) *Research outreach*---under-represented undergraduate and graduate student training by all the members.

The first project develops and studies a reduced simulation method (RSM) that uses a kD coordinate system defined using principal component analysis (PCA) of a standard MD trajectory to explore molecular motion. A primary objection of this sub-project is to provide a tool, which allows scientists to efficiently survey molecular motion using a limited MD trajectory. The essential components of the reduced simulation method include isolating k dominant features of an MD trajectory using ARPACK and defining a kD coordinate space; constructing an approximation of a potential energy surface based on the defined coordinate space; updating coordinates and velocities in the kD space based on the approximate energy surface; and analyzing the resulting information with respect to the original 3nD coordinate space. We will apply this method to various classes of molecules to do a benchmark test. We will compare our simulations to experimental data, such as data obtained from infrared (IR) spectroscopy. The IR spectrum of a molecule shows which frequencies of IR radiation are absorbed by the molecule and can be used to identify the functional groups in a molecule. This is also information that the RSM seeks to reveal. We will compare the data by considering the Fourier transform of the velocity autocorrelation function, the power spectrum. Once we finished all the benchmark tests, we will extend and interface this method widely into bioinformatics and

nanomaterial simulation utilizing readily available LONI and TeraGrid high performance computing facilities.

The second sub-project falls into research area of computational biology. Computational algorithms will be developed to predict the secondary and ternary structure of proteins. Computational modeling will be performed to simulate protein docking and interaction. In this research, two proteins, gK and UL20 of Herpes Simplex Virus Type-1, will be used as model proteins. The research is expected to achieve efficient algorithms for predicting protein structure, facilitating drug design and combat herpes virus infections.

The third sub-project is on a proposed LaSPACE sub-project working on code design and controlling a new Helium balloon for NASA. Dr. Khosravi, Dr. Yang, two graduate, and two undergraduate students will work on the project.

The fourth project is research outreach: training graduate and undergraduate students, especially African-American students, at Southern University and A & M College, a traditionally large HBCU institution. The PI and Co-PIs have NSF supported STEM program and a proposed NSF outreach project. Currently there are 385 undergraduate students and 70 graduate students enrolled in Computer Science Department. They will be trained by intimately engaging them in the activities aimed at the attainment of the high performance computing technical objectives above through our carefully designed training programs. We expect many more students from ME/EE, Physics, Chemistry, and Biology Department to benefit from it by virtue of our track record in training students.

The PI, Dr. Khosravi, Chair of the Computer Science Department, is currently funded by Navy, Raytheon, NSF, BoR, NIH, and NGA. Two Co-PIs are supported by LONI and Computer Science Department. Two ongoing projects, which Dr. Yang is working on, are funded by LBRN and LaSPACE. Current close collaborations with CEES in SU, LSU Vet School, LSU/ME, Louisiana Tech. Material Science would generate new opportunities to attract more talented faculties and post-doctors all over the nation and the world, which without doubt fits into SU and LONI's long term development strategy.

We propose the following support from LONI: (1). 300K Queenbee CPU time for three years; (2) Dr. Rachel 6 month for research sub-project one, 3 month for student training for three years; (3) Dr. Yang's research time: 2, 3, 3 months for the first three sub-project for three years respectively, 2 months for training graduate and undergraduate students. No extra support was proposed at this time.

Update:

SUBR Project 1: Reduced basis set method development

Dr. Rachel Vincent-Finley started the initial work on this project in May 2009, at Southern University of Baton Rouge. Her expertise is in reduced basis set MD simulation. The goal is to immediately involve current graduate students in the process of molecular analysis. During the 2009 – 2010 academic year, independent study modules will provide advanced undergraduate students opportunities to become involved in various aspects of the project. When finishing the project, we would expect that our method be used to analyze the protein structure and function,

MD visualization code developed, and minority students trained for industry and academic job markets.

SUBR Project 2: gK and UL20 protein structure prediction and interaction

Kimberlee Lyles got LONI fellowship and started the research in the past Spring 2009 semester under Dr. Shuju Bai, Dr. E. Khosravi, and Dr. Shizhong Yang's instruction. She successfully setup the nano-particle/membrane model and started simulating the interaction using NAMD package at LONI machines. The results should be sent for conference presentation and journal publication when the simulation finished and analyzed. This project is supported by LBRN pilot fund. Minority International (India) graduate student Murali K. Ganginela (1st year) and undergraduate student Laura Hurst (black) were supported by this project.

SUBR Project 3: NASA-LaSPACE supported thermal barrier coating material simulation

Under the sponsor of NASA-LaSPACE, Dr. E. Khosravi and Dr. Shizhong Yang supported the following minority graduate students in the Spring 2009 semester:

(1). Kiante Roberson: graduate student (1st semester), Computer Science Department, minornity(black);

(2). Sadque Ali Mohammed: graduate student (2nd year), Computer Science Department International (India);

(3). Charles alphonce shropshire: graduate student (1st semester), Computer Science Department, minornity (black);

The following presentation and publication were supported by or related to this project:

(1). Presentation at 2009 LAS 83rd annual conference: "First principles molecular dynamics simulation of nano gold adsorption on (0001) surface of Ruthenium", Shizhong Yang, Shuju Bai, Ebrahim Khosravi, and Guang-Lin Zhao.

(2). Shizhong Yang, S.M Guo, Guang-Lin Zhao, and Ebrahim Khosravi, "High infrared reflective nickel doped ZrO2 from first principles simulation", ICCS 2009. (international conference paper).

(3). Invited talk and paper: "Doped C60 study from first principles simulation", New3SC- 7, (Seventh International Conference on New Theories, Discoveries and Applications of Superconductors and Related Materials), Beijing, May, 2009.

(4). Wendong Wang, Zhenjun Wang, Jinke Tang, Shizhong Yang, Hua Jin, Guang-Lin Zhao, and Qiang Li, "Seebeck coefficient and thermal conductivity in doped C60", Journal of Renewable and Sustainable Energy 1, 23104(2009).

(5). G.L. Zhao, S. Yang, D. Bagayoko, J. Tang and Z.J. Wang, "Electronic structure of C60 semiconductors under controlled doping with B, N, and Co atoms", Diamond & Related Materials 17, 749 (2008).

SUBR Project 4: Research outreach: HBCU minority graduate and undergraduate student training

In the Spring 2009 semester, SU sponsored LONI HPC workshop training more than 50 graduate/undergraduate students attend at least one session of the workshop. More than 30 people attended the whole workshop. In the meantime we actively included minority students in our above proposed projects. Currently there are 9 graduate/undergraduate students in the BioInformatics/Computational Material group supported by all sorts of grants in our group.

Project 6. Data Management for Disaster Management though PetaShare

PIs: Ramesh Kolluru (ULL), Tevfik Kosar (LSU), Raju Gottumukkala (ULL), Rusti Liner (ULL); LI CS: Raju Gottumukkala (ULL).

The NIMSAT Institute at the University of Louisiana at Lafayette develops disaster management applications that leverage the cyberinfrastructure resources of the LONI and TeraGrid. These disaster management applications rely extensively on urgent, reliable and secure access to potentially terabytes of heterogeneous data (in the form of geospatial text, multimedia) ranging from geospatial imagery, LIDAR data, databases of critical infrastructures, public and private infrastructure, demographics, recent and historical hazard data. PetaShare is an NSF sponsored project that provides reliable and efficient access to distributed data resources to support large-scale data generation, sharing and collaboration requirements.

The LONI Computational Scientist at ULL, Dr. Raju Gottumukkala would contribute to this project and his tasks would involve uploading NIMSAT data to distributed PetaShare storage; generating related metadata and cataloguing this data on PetaShare; resolving data format conversion issues (e.g. NetCDF, GRIB, KML, Shape files); developing scripts for interfacing the data from Petashare with hazard models (e.g ADCIRC, SLOSH, ALOHA), GIS or Google Earth based applications and disaster management applications (HURREVAC, OREMS, POD Tool). Raju would also closely work with Ms. Rusti Liner, the GIS Manager at NIMSAT in identifying data requirements and data sources to catalogue data. Dr. Ramesh Kolluru, the Director of NIMSAT would work with the industry partners and government agencies through MOU's to obtain disaster related data. Dr. Tevfik Kosar, the PI and project lead of the Petashare, would provide assistance on handling any special needs to data in terms of providing resource provisioning, security and reliability.

This project would be a part time project for one year and would take FTE of three months. The LONI Institute can significantly benefit from this project by being recognized as a platform for assisting various disaster management agencies both at the state level like GOHSEP (Governors Office of Homeland Security and Emergency Preparedness) for the state of Louisiana and has the potential to contribute to the nation through agencies like DHS and FEMA.

Updates:

People Involved: *Raju Gottumukkala, Ramesh Kolluru, Tevfik Kosar, Ismail Akturk, Rusti Liner*

Overview:

This project leverages Petashare, a data management platform on LONI for various disaster preparedness and response applications developed by the NIMSAT Institute.

Disaster management applications used for disaster preparedness and response rely extensively on urgent, reliable and secure access to potentially terabytes of heterogeneous data (in the form of geospatial text, multimedia) ranging from geospatial imagery, LIDAR data, databases of critical infrastructures, public and private infrastructure, demographics, recent and historical hazard data. This data has to be accessed by the geospatial analysis applications (like GRASS) running on LONI, Visualization applications at the LITE (Louisiana Immersive Technology

Enterprise), or has to be accessible by the decision makers on their desktops. Petashare is an NSF sponsored project that provides reliable and efficient access to distributed data resources to support large-scale data generation, sharing and collaboration requirements. This LI-project would use Petashare on LONI, as a highly available data management and storage platform for NIMSAT's disaster preparedness and response.

Progress:

Preparing for this hurricane season, We are

- Investigating mechanisms for rapid access data from the NIMSAT facility, and applications running at LITE through visualization tools like Minerva.
- Collecting and cataloguing data for hurricane preparedness and response,
- Familiarized with the Petashare interfaces to manage files and investigated mechanisms to increase security and resilience of disaster related data.

Milestones & Timeline:

4. 06/30/2009 Catalogue hurricane related data and tools
5. 07/15/2009 Deliver tools to visualize data at LITE
6. 11/01/2009 Demonstrate a sample application where the data from Petashare is access by Parallel GRASS GIS running on LONI and LITE.

Project 7. Application Profiling on LONI

PIs: Erik Schnetter (LSU), Maciej Brodowicz (LSU), Steve Brandt (LSU), and Mayank Tyagi (LSU); LI CS: unassigned.

Project Description: As we move to more complex application codes (e.g. current Cactus code black hole simulations may contain 200 modules), machines with very large numbers of cores (e.g. the Blue Waters NSF system which LSU is involved in will contain over 200,000 cores), and more complicated and diverse processors (e.g. multicore, accelerators, pipelines) there is a critical need for reliable, easy to use, and user-oriented profiling information to allow developers and users to rework or tune their codes. Through the NSF ALPACA project we are already developing application level profiling and debugging tools based on the Cactus Framework, which can be used at run time. Using these tools requires additional 3rd party software (e.g. Tau, PAPI) to be installed, tested, configured and documented on machines, and currently we are using external TeraGrid machines for much of our work because of the better set up of this software. This project would involve a LONI computational scientist to help configure profiling tools that can be used on LONI for current applications, and the scientist would also take part in porting our application level profiling scenarios to the LONI machines. In connection with a second NSF project called XiRel we are also analyzing performance data with the aim of improving our core infrastructure for numerical relativity. The computational scientist would also take part in this effort and optimize the DOE Black Oil code developed by Mayank Tyagi and Chris White in the UCOMS project which uses the PETSc solvers.

Effort Requested and Involvement of Computational Scientist: We estimate that this would require 6 months of an FTE.

Benefit to LONI Institute: This would improve the availability and use of profiling tools on LONI from the very low level to the higher application level. The involvement of a new application code (Black Oil) would ensure that the tools can really be used for applications, would improve the code base for an important statewide project, and should provide good experience to the computational scientist.

Project 8. Surface Plasmon Excitation in inhomogeneous metal-dielectric Composites

PIs: Dentcho Genov (LaTech), and Shizhong Yang (SUBR); LI CS: Shizhong Yang (SUBR).

***Background*:** The inhomogeneous metal-dielectric composites are plasmonic nanomaterials that have unique geometrical and optical properties. Under electromagnetic wave illumination these complex materials manifest energy localization in very small spatial areas (a few nanometers) and huge enhancement of the local field intensities, which correspond to excitation of localized surface plasmon (SP) modes. At critical metal concentrations, the random films are inhomogeneous and self-similar (fractal) on any length-scale. Thus, for any incident wavelength resonating clusters exist in the composite. Such broad frequency response results in anomalous optical properties including extraordinary absorption and enormous enhancement of nonlinear optical processes such as Surface Enhanced Raman Scattering (SERS), high order frequency generation, *etc.* The unique properties of the percolating films make them ideal not only for fundamental studies of light-matter interaction in disordered systems, but also for a wide range of applications in biological sensing and spectroscopy (including single molecule detection with SERS), metamaterials and surface sciences, and condensed matter physics.

Proposed research: **1.** ***Numerical methods in nanoplasmonics*:** As part of this proposal we seek to rewrite the existing FDFD codes in parallel and develop novel, highly efficient numerical methods for calculating the EM response of 2D and 3D random systems of metal nanoparticles. Additionally, we intend to use a 'memoization' method, an efficient way to do fast searches of conduction paths, to develop a new methodology which could resolve the problem in only $O(N^{3/2})$, which is to be compared to $O(N^3)$ for the standard Gauss-Seidel method (N is the number of particles). Successful development of the numerical codes will make possible simulations on the LONI supercomputers of systems with up to 10^6 and 10^4 particles in the 2D and 3D cases, respectively. This will allow for first time to study local and macroscopic response of real systems and compare with experiments. Apart from solving plasmonic nanomaterials the developed numerical codes could be effectively applied for investigation of large variety of strongly interactive, sub-wavelength ensembles of particles (not necessary metal), including dense semiconductor quantum dots systems, periodic arrays with tunable optical properties, photonic nano-circuits and optical switches. **2.** ***SP eigenproblem: localization-delocalization transition in percolating metal composites*:** Here, we intend to investigate the nature of SP eigenstates at localization-delocalization transition in 3D random media. This transition exists regardless of the dimension of the problem. For example, in the 2D case, it is manifested through a logarithmic singularity at the center of the energy band. The 3D case, however, has not been investigated yet due to computational limitations. To study the collective SP eigenproblem we intend to integrate existing parallel eigensolvers (LAPACK) to operate on the LONI machines. Due to the memory intese calculations we will look at optimizing the memory partition subroutines to take advantage of the operational memory available through the LONI infrastructure. This will allow investigation of the multi-fractal characteristics of the critical eigenstates and combined with calculation of the SP density of states will provide a complete picture of the collective SP phase transition. Consequently, the acquired data will serve as a basis for the development of a comprehensive analytical theory describing the electromagnetic response of the system at percolation. This theory may reveal new ways to enhance the local

optical response of the composites materials with direct applications in surface enhanced Raman spectroscopy, development of adaptive metal films for implementation as low-pass filters, coating materials and tunable optical media.

Impact of the proposed research: The proposed research will lead to development of novel numerical and analytical tools for solving highly complicated problems of EM interaction with complex media. Those methods will answer standing fundamental questions concerning the nature of collective electronic excitations in metal-dielectric composites. Due to the inhomogeneous nature of the problem it is crucial that very large system sizes are investigated. Such systems cannot be studied with average computational facilities and utilizing the LONI recourses will allow to traverse new regimes of operation that have been a mystery for the last 50 years. Successful realization of the project, have the potential to establish the LONI Institute as a top center for computational electromagnetism. Furthermore, the developed numerical methods will have strong practical impact on the development of adaptive composite materials for enhanced linear and nonlinear optical processes. For instance, the optimization of SERS from molecules deposited on or inside the composites could lead to spectroscopic measurements with unsurpassed sensitivity. The large range of applications, could serve as a basis to build on previous and establish new collaborations with experimental scientists within the six LONI institutions but also with other national universities including groups at UC Berkeley (Prof. X. Zhang), and Yale University (Prof. Hui Cao). The first part of this project has been included into a RCS proposal submitted on Nov. 7 to the Louisiana Board of Regents, while funding from the NSF materials division and DARPA will be sought in relation to the SP eigenproblem. The total workloads for the LI faculty and CS are 3 FTE-months per year, for total duration of the proposal of 1.5 years and expected supercomputer time allocation of 50K SUs. Also, the LI faculty will provide a PC workstation and a graduate student to work full time on the project, which will also be the subject of the student PhD thesis.

Project 9. Refinement of Integral Membrane Protein Structure Predictions

PIs: Christopher Summa (UNO), Steven Rick (UNO), and Zhiyu Zhao (UNO); LI CS: Zhiyu Zhao (UNO).

Protein Structure Prediction

The field of protein structure prediction concerns itself with the generation of models of protein structures that approximate the true, native protein structure as accurately as possible. These methods are intended to augment, or even replace, the experimental determination of a protein structure where such a determination is either highly derivative (as in the case of a protein with a close relative of known structure), or experimentally difficult (as in the case of integral membrane proteins). It has been estimated that the generation of an experimental protein structure costs, on average, between $250,000 [1] and $300,000 [2] (US). Improved methods in structure prediction, therefore, hold the promise of shifting some of the cost burden from experimentalists into (relatively) cheap computations, allowing experimentalists to focus on those structures of particular interest.

The Membrane Protein Structure Problem

It has been estimated that as much as 30% of the open reading frames of the genomes of higher eukaryotes code for proteins which span or are otherwise associated with cell membranes [3]. Despite their prevalence in biological systems, however, the scarcity of integral membrane protein structures stands in sharp contrast to the rapid accumulation of structural data for soluble proteins in the Protein Data Bank (PDB). To date roughly 58,236 X-ray or NMR derived structures of soluble proteins have been deposited in the PDB, while only ~193 structures of membrane proteins are currently known, due to inherent difficulties in membrane protein purification and crystallization. There have been a number of spectacular successes in X-ray crystallography of membrane proteins in recent years, and recent advances in crystallization techniques may well allow structural biologists to lessen the disparity in the structure database in the coming years. However, until such time as crystallization of membrane proteins becomes routine, method development in structure prediction of integral membrane proteins remains an important undertaking.

Prediction methods complement, enhance, and are enhanced by traditional methods of gaining structural information. For example, an initial model can provide a roadmap for mutagenesis experiments. The results of mutagenesis experiments can guide the building of an initial model, or suggest ways to improve upon an existing one. An experimental structure can either prove or disprove a model, can afford us suggestions on how to improve our techniques, and can provide a useful template for modeling related proteins with significant sequence homology.

Protein 3-D Structure Refinement

One of the greatest shortcomings of macromolecular energy minimization and molecular dynamics is that they generally do not preserve the native structure of proteins as observed by X-ray crystallography. This deformation of the native structure means that these methods are not generally used to refine structures produced by homology modeling techniques. In recent work [4, 5] we have shown that it is possible to improve an ensemble or near-native globular protein

structures using energy minimization techniques such that their structures are closer to native than the starting structure. A database of 75 globular proteins was used to test the ability of a variety of popular molecular mechanics force fields to maintain the native structure. Minimization from the native structure is a weak test of potential energy functions: it is complemented by a much stronger test in which the same methods are compared for their ability to attract a near-native decoy protein structure towards the native structure. Using a powerfully convergent energy minimization method, we showed that, of the traditional molecular mechanics potentials tested, only one showed a modest net improvement over a large dataset of structurally diverse proteins. A smooth, differentiable knowledge-based pairwise atomic potential performed better on this test than traditional potential functions. This method is of particular utility because of its computational efficiency relative to stochastic search methods.

We propose to test (using the LONI computational resources) whether the same or similar technique can be used for a set of membrane proteins whose crystal structures have been determined. Initial tests will focus on energy minimization both of native membrane protein structures, and on the ability to make a "perturbed" membrane protein structure (representing, for example, the output of a reasonable homology model) revert to its native configuration. Both energy minimization and molecular dynamics using replica exchange [6], will be tested using a range of potential energy functions for their ability to improve near-native decoys. This work is highly computationally intensive, and access to the LONI infrastructure would be of particular importance to the success of this project.

A key component of this work is the comparison of both the pre- and post- refined membrane protein structures to the known, native state. A robust method of comparison is essential if we are to learn the strengths and limitations of our techniques, and to determine where our methods perform well, and where they do not.

Protein 3-D Structure Alignment

3-D structures are strongly related to their biological functions [7]. Protein structures reveal more evolutionary information than protein sequences do, since the structure of a protein changes more slowly in evolution than does its sequence [8]. Also, researchers have frequently observed that proteins with low sequential similarities are structurally homogenous. Therefore it is particularly important to discover the structural similarities / dissimilarities among different proteins. The research of protein 3-D structure similarity is very helpful for many biological applications such as predicting the functions of unknown proteins from known similar protein structures, identifying protein families with common evolutionary origins, understanding the variations among different classes of proteins, and so on. Pairwise protein 3-D structure alignment attempts to compare the structural similarity between two protein backbone chains. An alignment is characterized by (1) how many positions are matched, (2) where these positions are and (3) how well they are matched. The alignment problem is non-trivial – in fact, the problem of finding the optimal global alignment between protein structures has been shown to be NP-hard[9, 10].

Introduction to SLIPSA:

SLIPSA is a Self-Learning and Improving pairwise Protein Structure Alignment algorithm developed by Drs. Bin Fu and Zhiyu Zhao's research group. It shows better accuracy when

compared with other well known algorithms such as CE [11], Dali [12] and SSM [13] (see [14, 15]). Our algorithm is implemented with Matlab and we have developed a web tool (http://fpsa.cs.uno.edu, http://fpsa.cs.panam.edu/) based on this program. SLIPSA is the foundation of our protein structure query tool which searches similar structures in the Protein Data Bank (PDB) according to a given query structure. Due to large size of PDB and high complexity of current protein structure alignment algorithms, protein structure query is very time-consuming and computation capability of machines greatly affects query performance in terms of both speed and accuracy. Since SLIPSA is a serial program written with Matlab, there is a lot of space to improve its speed performance by (1) rewriting the code with C/C++ and (2) taking advantage of parallel and distributed computation power of HPCs.

Effort Requested and Involvement of Computational Scientist

We would like to request 4 months of full time effort on the part of Dr. Sylvia Zhao. Dr. Zhao has extensive expertise in protein structure alignment, programming in Matlab and C/C++ and compiling and running code on the LONI cluster. Dr. Zhao's responsibilities will involve some programming of the parallel implementation of the SLIPSA algorithm, running energy minimization experiments and compiling and analyzing data. Dr. Summa and Dr. Rick will provide coding support for the molecular simulation code, and perform data analysis.

Benefit to LONI Institute

This proposal represents an interdisciplinary collaboration between a Computational Biologist (Dr. Summa), and Computational Physical Chemist (Dr. Rick) and a Computer Scientist (Dr. Zhao). The tools developed will be shared with LONI users once they have been validated and made "user-friendly", and should provide a important resource for Computational Structural Biology within the LONI network.

Bibliography

1. Lattman, E., The state of the Protein Structure Initiative. Proteins, 2004. 54(4): p. 611-5.

2. Service, R., Structural biology. Structural genomics, round 2. Science, 2005. 307(5715): p. 1554-8.

3. Stevens, T.J. and I.T. Arkin, Do more complex organisms have a greater proportion of membrane proteins in their genomes? Proteins, 2000. 39: p. 417-420.

4. Chopra, G., C. Summa, and M. Levitt, Chopra G, Summa CM, and Levitt M Solvent Dramatically Affects Protein Structure Refinement Proceedings of the National Academy of Sciences USA 2008 (105) 20239-20244. Proc. Natl. Acad. Sci. USA, 2008. 105: p. 20239-20244.

5. Summa, C.M., M. Levitt, and W.F. Degrado, An atomic environment potential for use in protein structure prediction. J Mol Biol, 2005. 352(4): p. 986-1001.

6. Rick, S.W., Replica exchange with dynamical scaling. Journal of Chemical Physics, 2007. 126: p. 054102.

7. Petsko, G. and D. Ringe, Protein Structure and Function 2004: New Science Press.

8. Eidhammer, I., I. Jonassen, and W.R. Taylor, Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis. . 2004: John Wiley and Sons.

9. Godzik, A., The structural alignment between two proteins: Is there a unique answer? Protein Science, 1996. 5: p. 1325–1338.

10. Lathrop, R.H., The protein threading problem with sequence amino acid interaction preferences is np-complete. Protein Engineering, 1994. 7: p. 1059-1068.

11. Shindyalov, I.N. and P.E. Bourne, Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. Protein Engineering, 1998. 11: p. 739-747.

12. Holm, L. and C. Sander, Protein structure comparison by alignment of distance matrices. Journal of Molecular Biology, 1993. 233: p. 123-138.

13. Krissinel, E. and K. Henrick, Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. Acta Crystallographica, 2004. D60: p. 2256-2268.

14. Zhao, Z. and B. Fu. A flexible algorithm for pairwise protein structure alignment. in Proceedings International Conference on Bioinformatics and Computational Biology. 2007.

15. Zhao, Z., et al., Feedback algorithm and web-server for protein structure alignment. Journal of Computational Biology, 2008. 15(3): p. 505-524

Project 10. Parallel-GIS: A High Performance Open Source Geospatial Analysis

PIs: Ramesh Kolluru (ULL), Baker Kearfott (ULL), Raju Gottumukkala (ULL); LI CS: Raju Gottumukkala (ULL).

Geographic Resource Analysis Support System (GRASS) is multipurpose open source GIS software for geospatial data analysis, modeling, management and visualization. Various modules of GRASS are currently being utilized in multiple areas of science such as Geography, Sociology, Ecology, Remote Sensing, Urban-Planning, Geostatistics, Geophysics and Hydrology. GRASS would be a versatile tool to better understand the impact of disasters on the people, community and assets. The various aspects of disaster management efforts can be significantly improved by better interlinking the workflow of geospatial data, geospatial analysis modules in GRASS, natural disaster prediction models and logistics modules for planning.

Dr. Ramesh Kolluru and Dr. Baker Kearfott are working on a project through the Governor's Information Technology Initiative (ITI) for University Of Louisiana at Lafayette. The major objective of this project is to deploy GRASS on LONI and LITE, parallelize certain modules of GRASS that can be applied to disaster management (e.g raster modules like LIDAR data processing, satellite image processing algorithms, vector modules like road network analysis), and develop optimization/Operations Research algorithms/libraries portable with GRASS for supporting emergency management and planning during disasters.

Dr. Raju Gottumukkala would take primary responsibility in this project through understanding various GRASS modules, deploying GRASS on LONI and LITE, interlink various GRASS modules (including data and applications) with external models like weather, storm surge or plume models, train and assist students with their research and projects in parallelizing certain modules of GRASS that can be applied to disaster management, and help Dr. Kearfott and Dr. Kolluru with developing MPI based OR algorithms library that can be integrated with GRASS modules. There are two student's who are currently working on this project. Jeevan Gogineni a master's student from Computer Science department and Zhang Haochun, a PhD student from Math department both would be working with Raju on this project.

We request 10 hours per week of the LONI Computational Scientist, Dr. Raju Gottumukkala for the specified tasks on this project.

Updates:

People Involved: Raju Gottumukkala, Ramesh Kolluru, Baker R. Kearfott

The major objective of this project is to provide a High-Performance GIS on LONI and LITE through GRASS (Geographic Resources Analysis Support System) a multipurpose open source GIS software for geospatial data analysis, modeling, management and visualization. High-Performance GIS improves the disaster management by analyzing data at a fine granular level and improve (e.g raster modules like LIDAR data processing, image processing algorithm, and vector modules like road network analysis algorithms).

Various projects at NIMSAT currently leverage the open source and parallelizable aspects of GRASS on LONI. (1) The Point of Distributions (PODs) application uses a parallelized version of the shortest path finding module from GRASS for determining the proximity of POD

locations. The POD application also uses a parallel version of the site selection algorithm to find the optimal POD site. These modules have been implemented and tested on the LONI Dell systems. (2) The Critical Infrastructure Analysis and Key Resources (CIKR) project extends some of the GRASS modules to perform analysis on the nation's natural gas pipeline network and natural gas facilities to identify critical pipelines and facilities.

Appendix D

LI Graduate Fellows, 2008-2009

Christopher Clayton / Kimberlee Lyles, SUBR (Fall 2008 and Spring 2009, respectively)

Research:

During the Fall 2008 semester, Mr. Christopher Clayton was supported by the LONI Institute fellowship working on the LBRN supported virus gK and UL20 secondary structure prediction. He used hidden Markov model and neuro network methods. His contributions are: (1). Tested all the available codes and compared gK and UL20 results with experimental data. (2). Hidden Markov is better than Neuro Network method in gK virus structure prediction. (3). Successfully predicted signal peptides for both gK and UL20.

However, in Spring 2009 semester, Mr, Clayton got sick, and Ms. Kimberlee Lyles continued his work on (1). gK and UL20 homolog alignment and functional analysis; (2). Molecular dynamics (MD) to study the interaction between gK and UL20. Ms. Lyles did gK and UL20 sequence analysis and successfully setup the NAMD model for simple nano-particle interaction inside cell membrane. She is doing some test runs of the MD at LONI supercomputers and analyzing the states of nano-particle inside the membrane. At the same time, she is setting up the gK and UL20/membrane model for the MD simulation. The results will be submitted to a conference soon.

Publications and Presentations of the research:

1. Presentation at 2009 LAS 83rd annual conference: "First principles molecular dynamics simulation of nano gold adsorption on (0001) surface of Ruthenium", Shizhong Yang, Shuju Bai, Ebrahim Khosravi, and Guang-Lin Zhao.

2. Shizhong Yang, S.M Guo, Guang-Lin Zhao, and Ebrahim Khosravi, " High infrared reflective nickel doped ZrO_2 from first principles simulation", ICCS 2009. (international conference paper)

3. Invited talk and paper: "Doped C_{60} study from first principles simulation", New3SC-7, (**S**eventh International Conference on New Theories, Discoveries and Applications of Superconductors and Related Materials), Beijing, May, 2009.

4. Wendong Wang, Zhenjun Wang, Jinke Tang, Shizhong Yang, Hua Jin, Guang-Lin Zhao, and Qiang Li, " Seebeck coefficient and thermal conductivity in doped C_{60}", Journal of Renewable and Sustainable Energy **1**, 23104(2009);

5. G.L. Zhao, S. Yang, D. Bagayoko, J. Tang and Z.J. Wang, **"** Electronic structure of C60 semiconductors under controlled doping with B, N, and Co atoms", Diamond & Related Materials **17,** 749 (2008).

Use of LONI and HPC resources:

They have used LONI machines up to more than 1000 cpu to simulate BioInformatics and materials.

Collaborations:

They are in close contact with LSU Vet School (Gus), Biology (Newcomer), Mechanical Engineering (S. Guo), LSU CAMD (K. Lian and P. Zhou), and Physics (Wefel and Guzik).

Jeremy Dewar, Tulane

Research:

Mr. Dewar has been looking into a specific approach to solving systems of nonlinear hyperbolic conversation laws. His research has been with examining the performance of a smoothness indicator for central-upwind schemes. Solving the Riemann problem can be broken down to solving self-sharpening discontinuities (shock waves) and discontinuities that will not self-sharpen (contact waves). A smoothness indicator can discern between these two waves. This allows for a smooth solution approach to the contact waves (this will better preserve the structure of the discontinuity that cannot sharpen itself), and a limited solution approach to the shockwaves (keeping such waves from causing oscillations without an 'intelligent' Riemann problem solver).

The smoothness indicator he has researched is specific to the Euler gas dynamics equations, but can be generalized to other systems of equations by looking at either the velocity or the pressure of the system. This indicator has been extended to a 2D grid and gives results similar to well-known results in 1D.

Publications and Presentations of the research:

There has not been a publication about this project, yet. This project was given to him as a 1st year graduate student with the goal of writing a paper on the results. They intent to have public presentations about his work. Some preliminary results may be discussed at ICOSAHOM '09.

Use of LONI and HPC resources, and collaborations:

Mr. Dewar corresponded with Dr. Fujioka, LI computational scientist, about MPI and Open MP implementations. He was a great help in extending the project to multiple processors.

A. Murat Eren, UNO

Research:

LI Grad Fellow Murat Eren has made excellent progress in this past year. He has significantly impacted the efforts of fellow graduate students in Dr. Winters-Hilt's group as well as significantly improved his advisor's grant prospects with his nice results and excellent presentation of those results. Murat is an amateur photographer, so it is no coincidence that he has a good eye for clear presentation of results or visualization of data-features. Murat is a highly skilled programmer that has contributed to the design of the Pardus Operating system (for the Turkish Army), this aided him greatly in establishing a real-time pattern recognition feedback interface between an experimental apparatus (a nanopore detector) and my AI-based feature extraction and classification methods (1). He is skilled at creating visualization software -- in the past year he created visualization software for my channel current feature extraction and classification software (2). Not to be left out of the algorithmic development side of the effort, Murat has independently developed a kernel-based clustering procedure (unfortunately already published in IEEE in 2002), and he did this in just the past month. Murat and Dr. Winters-Hilt are now using a novel variant of his kernel-based clustering procedure as a preprocessing step prior to using an SVM-based clustering scheme that I've explored separately – Murat is now helping Dr. Winters-Hilt to re-submit two papers in this regard (3,4), as well as prepare a manuscript for the exciting results anticipated for the above hybrid clustering approach (5). Mr. Murat's numerous areas of expertise have had a significant impact on the work of fellow Ph.D. students Zuliang Jiang and Carl Baribault, as well as Dr. Winters-Hilt's grant and patent filing efforts (6,7,8). Murat has helped Carl, in particular, to make use of the LONI facilities for his HMM-based gene structure identification work, and it's in this latter effort that the group has been making the most use of the LONI infrastructure.

Publications of the research:

1. Eren AM, Amin I, Alba A, Morales E, Stoyanov A, and Winters-Hilt S. Pattern Recognition Informed Feedback for Nanopore Detector Cheminformatics. Submitted to BMC Biotechnology.

2. Eren AM & Stephen Winters-Hilt. A Visualization Tool for Nanopore Experiments. Submitted to MCBIOS Proceedings for BMC Bioinformatics.

3. Winters-Hilt S, Eren AM, and Armond Jr. K. Distributed SVM Learning and Support Vector Reduction. Re-submission planned to BMC Bioinformatics.

4. Winters-Hilt S, Eren AM, and Merat S. Unsupervised clustering using supervised support vector machines. Re-submission planned BMC Bioinformatics.

5. Winters-Hilt S and Eren AM. SVM-based clustering with kernel-clustering for kernel-tuning and seed cluster-region identifications.

6. Winters-Hilt S and Jiang Z. An Efficient Self-Tuning Explicit and Adaptive HMM with Duration Algorithm. Accepted by IEEE Transactions on Signal Processing, June 2009. (http://www.cs.uno.edu/~winters/ESTEAHMMD_preprint.pdf)

Presentations of the research:

1. A. Murat Eren & Stephen Winters-Hilt. Pattern recognition-informed sampling for nanopore biosensing. MidSouth Computational Biology and Bioinformatics Society (MCBIOS), Starkville, MS, Feb. 20-21, 2009.

2. A.Murat Eren & Stephen Winters-Hilt. A Visualization Tool for Nanopore Experiments. MidSouth Computational Biology and Bioinformatics Society (MCBIOS), Starkville, MS, Feb. 20-21, 2009.

3. Joshua Morrison, A. Murat Eren, and Stephen Winters-Hilt. Machine Learning Web Interfaces for Bioinformatics & Cheminformatics. MidSouth Computational Biology and Bioinformatics Society (MCBIOS), Starkville, MS, Feb. 20-21, 2009.

4. Amanda Alba, Eric Morales, A. Murat Eren, Joshua Morrison and Stephen Winters-Hilt. Nanopore-transduction based study of individual molecular binding events. MidSouth Computational Biology and Bioinformatics Society (MCBIOS), Starkville, MS, Feb. 20-21, 2009.

Patents or Licensing Agreements:

1. Winters-Hilt S and Zhang J. An efficient implementation for HMM with duration. PATENT pending, UNO filing, October 2008.

2. Winters-Hilt, S., Pattern Recognition Informed (PRI) Nanopore Detection for Sample Boosting, Nanomanipulation, and Device Stabilization; and PRI Device Stabilization Methods in General. PATENT pending, UNO filing, August 2008.

Use of LONI and HPC resources:

As mentioned, via collaboration with fellow PhD researchers Carl Baribault and Zuliang Jiang, Murat is helping on gene-structure identification projects that use significant computational time – with a rapidly growing demand. The use of LONI facilities would be significant, and rapidly-growing, on the basis of the gene-structure identification work alone, but now Murat is also engaging in SVM self-tuning methods (in our clustering collaboration) as well as with large dataset processing requiring SVM-chunking, both of which will also entail extensive use of LONI's resources.

Collaborations:

We have not had extensive collaboration with LONI members outside the group in our first year, but in anticipation of students bridging between the AI-side of the Winters-Hilt group and the computer forensics work of fellow LONI PI's Golden Richard and Vassil Roussev, we expect to see greater involvement with the LONI PI's at UNO (the last student to do this type work, Brian Roux, just graduated, and worked partly with Golden's group and partly with Winters-Hilt's).

John Jack, LA Tech

Research:

Title for his PhD dissertation: "Discrete Nondeterministic Modeling of Biochemical Networks":

The ideas expressed in this work pertain to biochemical modeling. They explore their technique, the Nondeterministic Waiting Time algorithm, for modeling molecular signaling cascades. This algorithm builds on earlier work from the lab of Dr. Andrei Paun, the advisor for Mr. Jack's dissertation. They discuss several important extensions including: (i) a heap with special maintenance functions for sorting reaction waiting times, (ii) a nondeterminstic component for handling reaction competition, and (iii) a memory enhancement allowing slower reactions to compete with faster reactions.

Several example systems provide comparisons between modeling with systems of ordinary differential equations, the Gillespie Algorithm, and our Nondeterministic Waiting Time algorithm. Their algorithm has a unique ability to exhibit behavior similar to the solutions to systems of ordinary differential equations for certain models and parameter choices, but it also has the nondeterministic component which yields results similar stochastic methods (e.g., the Gillespie Algorithm).

They also investigate the Fas-mediated apoptotic signaling cascade. Fas signaling has important implications in the research of cancer, autoimmune and neurodegenerative disorders. They provide an exhaustive account of results from the Nondeterminstic Waiting Time algorithm in comparison to solutions to the system of ordinary differential equations described by another modeling group. Their work with the Fas pathway led them to explore a new model, focusing on the effects of HIV-1 proteins on the Fas signaling cascade. There is extensive information in the literature on the effects of the HIV-1 proteins on this pathway. The model described in their work represents the first attempt ever made in modeling Fas-induced apoptosis in lantently infected T cells.

There are several extensions for the Fas and the HIV models. Calcium signaling would be an interesting avenue to investigate, building on some recent results reported in the literature. We also suggest a new direction for the Nondeterministic Waiting Time algorithm exploring parallelization options.

Publications of the research:

J. Jack and A. Paun, ``Discrete Modeling of Biochemical Signaling with Memory Enhancement," LNBI Transactions on Computational Systems Biology 2009, 14 pp. [accepted].

J. Jack, A. Paun, Simulation of Signaling Pathways through discrete methods, JALC, accepted 2009.

J. Jack, A. Paun, F A. Rodriguez-Paton, Discrete nondeterministic modeling of the FAS pathway, Int. Journal of Foundations of Computer Science, vol. 19 (October 2008), no. 5, pp. 1147-1162.

J. Jack, A. Paun, A. Rodriguez-Paton, Effects of HIV-1 Proteins on the Fas-Mediated Apoptotic Signaling Cascade: A Computational Study of Latent CD4+ T Cell Activation,

accepted at Ninth Workshop on Molecular Computation, WMC9, Edinburgh (UK) July 28-31, 2008, 20pp.

Presentations of the research:

Mr. Jack gave a presentation at the EPA's National Center for Computational Toxicology (NCCT) on the research involving the Nondeterministic Waiting Time (NWT) algorithm. The talk was given in May 2009 and was one hour in length.

Dr. Andrei Paun (Jack's dissertation advisor) is presenting various aspects of their research at "Descriptional Complexity of Formal Systems" in Magdeburg, Germany. He is an invited speaker at the conference which takes place July 6th - 9th.

Use of LONI and HPC resources:

The Nondeterministic Waiting Time algorithm is written in C. Using MPI, they are able to run simultaneous biochemical simulations for model fitting. Instead of running each simulation separately (serial), they can explore multiple kinetic rates on multiple nodes, assisting in the fitting of our models to the biochemical information available in the literature.

The models they fit involved HIV-1-infected T cells and cancer cells. They investigated the Fas-mediated signaling cascade, which is one of the pathways responsible for apoptosis. The parallelization also allows them to explore the different cellular evolutions due to the nondeterminism of the NWT algorithm.

Collaborations:

Via his advisor, Dr. Paun, they had interactions with LSU, specifically with Drs. Thomas Klei, Hilary Thompson, Bill Wischusen, Konstantin Gus Kousoulas, and Doan H. Nguyen.

Other achievements:

Mr. Jack was able to complete the degree requirements for a Ph.D. in Computational Analysis and Modeling by May 2009. Without the funding from LONI, it would have been difficult for him to accomplish this goal.

Also, he has been granted a three-four year post-doctoral position at the EPA's National Center for Computational Toxicology. At this post, he will continue his research into modeling molecular signaling cascades. The project he will be a part of is called virtual Liver -- modeling the effects of toxic elements on the cellular and intracellular interactions of the human liver.

Jijun Lao, LSU

Research:

During the last year using large-scale massively parallel molecular dynamics simulations on LONI computer clusters, they carried out various investigations aimed at revealing the fundamentals of the effects of materials microstructure on mechanical properties and structural stability of metallic bulk, thin films and nanowire structures. In particular, their molecular dynamics simulation studies of Pd nanowires indicate that depending on the wire diameter the surface stress can cause Pd nanowire to undergo spontaneous structural reorientation or phase transformation. Under tensile loading and unloading Pd nanowires transform reversibly between the two crystallographic orientations exhibiting pseudoelastic behavior characterized by fully recoverable strains of up to 50%. The temperature-dependence of the pseudoelastic behavior enables the shape memory effect in Pd nanowires. These novel properties can greatly impact nanowires usage in a large class of nanodevices including sensors, actuators and transducers.

Mr. Lao is also involved in two research projects: one dealing with MD simulation studies of interfacial strain induced formation of metallic nanotubes and nanocoils and the other one dealing with the investigation of the pore nucleation and growth in lipid bilayer membranes in the presence of dimethylsulfoxide. Using the preliminary results we obtained in the first project we have already prepared a first draft of a paper that we plan to submit for publication to Physical Review Letters.

Publications and Presentations of the research:

1. J. Lao and D. Moldovan, "Surface stress induced structural transformations and pseudoelastic effects in palladium nanowires" Appl. Phys. Lett. 93, 093108, 2008

2. J. Lao and D. Moldovan, "Molecular dynamics simulation study of pseudoelastic effects in palladium nanowires" The Fourth International Conference on Multiscale Materials Modeling, Tallahassee, Florida, October 27-31, 2008.

3. J. Lao and D. Moldovan, "Interfacial strain induced self-rolling of Aluminum nanotubes" In preparation, plan to submit for publication to Physical Review Letters

Use of LONI and HPC resources:

The MD simulations for both projects are performed on LONI systems on which we currently have a 500,000 hours allocation quota.

Philip Schexnayder, Jin-Feng Chen, ULL (Spring 2009)

Research:

Philip Schexnayder: Despite his short tenure as a graduate student, Philip has made an impressive progress in the development of new computational tools for the rhythmic analysis of echolocation and communication signals of marine mammals (particularly, sperm and beaked whales). The algorithm provides detection of a particular species in a continuous stream of broadband acoustic data and a robust (to low Signal-to-Noise-Ratio) method for association of rhythmic frequencies with individuals. The method has been successfully applied to passive experimental acoustic recordings collected by the Littoral Acoustic Demonstration Center (LADC) in the Gulf of Mexico. Multi-channel data analysis would benefit from a use of parallel processing techniques and the LONI HPC resources. As the next step, we are planning to adapt the code for the LONI environment.

During the Spring 2009 semester, Philip weekly participated in supercomputing training conducted by the OU Supercomputing Center for Education & Research through the UL AceessGrid. He is also registered to attend the LONI training in Baton Rouge in the summer.

Jin-Feng Chen: Jin Feng Chen's research focuses on developing a methodology for the construction of 3D Highway models that can be used as engineering analysis tools for highway infrastructure system. This involves collecting image and video data collected from a driver's perspective and to extract information pertinent to the roadway, including the road surface, the shoulder areas, guardrails, traffic control devices, and all roadside elements. The information is then integrated into a 3D environment that will give engineers new tools to examine and identify highway features such as the degree of curvature, super elevation, sight distance, and pavement edge lines. This tool will enable highway engineers to design and evaluate highway infrastructures from new perspectives, which are not feasible with the currently available technologies. As a part of this research, Jin Feng was involved in the development of a center based Treemap algorithm, 3D Treemap algorithm and a clustering algorithm to represent the clustered results on a browser.

Publications and Presentations of the research:

1. Juliette W. Ioup, George E. Ioup, Lisa A. Pflag, Arslan M. Tashmukhambetov, Christopher O. Tiemann, Alan Berstein, Natalia Sidorovskaia, Philip Schnenayder et al., "Localization to verify the identification of individual sperm whales using click properties," The Journal of the Acoustical Society of America, 125(4, pt.2 of 2), April 2009, p. 2616 (published abstract)

2. Natalia Sidorovskaia, Philip Schexnayder, et al., "Rhythmic analysis of sperm whale broadband acoustic signals," The Journal of the Acoustical Society of America, 125(4, pt.2 of 2), April 2009, p. 2738 (published abstract)

3. S. Chu, J. Chen, Z. Wu, V. Raghavan, H. Chu. "A Treemap-based Result Interface for Search Engine Users", 12th International Conference on Human-Computer, Interaction (HCI 2007), Volume 8, July 2007.

4. Philip Schnenayder, Physics Department seminar, April 2009.

5. Philip Schnenayder, Oral presentation at the 157th meeting of the Acoustical Society of America, Portland, Oregon, May 22 2009: “Rhythmic analysis of sperm whale broadband acoustic signals”.

Patent or licensing agreements:

Shixian Chu, Jinfeng Chen, Zonghuan Wu, Chee-Hung Henry Chu, Vijay Raghavan, “Method and Apparatus for Information Visualized Expression and Visualized Human Computer Interactive Expression Thereof”, PCT/CN2008/000168

Use of LONI and HPC resources:

Philip Schnenayder attended training sessions offered by LONI and LSU HPC.

Other achievements:

Philip Schnenayder was awarded partial financial support from the Graduate Student Organization to attend the conference. He also received the Acoustical Society of America student’s travel grant.

Appendix E. Published Publications

Available online at www.sciencedirect.com
ScienceDirect
Diamond & Related Materials 17 (2008) 749–752
DIAMOND AND RELATED MATERIALS
www.elsevier.com/locate/diamond

Electronic structure of C_{60} semiconductors under controlled doping with B, N, and Co atoms

G.L. Zhao [a,*], S. Yang [a], D. Bagayoko [a], J. Tang [b], Z.J. Wang [b]

[a] Physics Department and High Performance Computing Laboratory, Southern University and A & M College, Baton Rouge, LA 70813 USA
[b] Physics Department, University of New Orleans, New Orleans, LA 70148 USA

Available online 23 December 2007

Abstract

We present our recent studies of *ab initio* density functional theory (DFT) calculations of the electronic structures of several selected n- and p-type doped C_{60} semiconductors. A super-cell approach was used. We performed a series of *ab initio* density functional computations to systematically study the changes of the electronic structure of C_{60} semiconductors doped with boron, nitrogen and cobalt atoms. We found that boron and cobalt doped, face-centered cubic (FCC) C_{60} solids have the electronic structures of n-type semiconductors. Nitrogen doped FCC C_{60} solid has an electronic structure similar to those of a p-type semiconductor, with shallow impurity energy levels near the top of the valence bands of the host material.

Keywords: Fullerenes; Simulation; n-type and p-type doping; Electronic properties

1. Introduction

The unique properties of C_{60} materials present some new opportunities for technology applications [1,2]. Especially, the very low Debye temperature (Θ_D=70 K) and the low thermal conductivity of C_{60} bulk semiconductors present an opportunity for constructing new thermoelectric materials with a high *figure-of-merit ZT*, [3] where $ZT = \frac{S^2\sigma T}{\kappa}$; S is the thermoelectric power (or Seebeck coefficient); σ is the electrical conductivity; T is the temperature; and κ is the thermal conductivity. Controlled doping of C_{60} semiconductors presents an effective method to tune the electronic properties of the material. However, to identify suitable doping elements and doping concentrations in C_{60} bulk semiconductors, for achieving a high thermoelectric *figure-of-merit*, is a challenging task in the exploration of the new C_{60} fullerene based thermoelectric materials. Understanding the electronic structure of the doped materials is another major task in the research. On the other hand, *ab initio* density functional calculations are effective methods to reveal the electronic structure of the materials under controlled doping. In this work, we performed the *ab initio* density functional calculations to study the electronic structure of C_{60} semiconductors doped with B, N, and Co atoms in the interstitial sites of the parent material in FCC lattice. Miyamoto et al. [4] studied the B-doped C_{60} (BC_{59}) with one carbon atom substituted by a boron atom. Ching et al. calculated the electronic energy bands, the density of states (DOS), and the optical properties of C_{60}, K_3C_{60}, and K_6C_{60} solids in FCC lattice [5–7]. For more than a decade, most of the research on the interstitially doped C_{60} solids focused on high doping concentrations and for the superconducting properties. Gu et al. performed AC susceptibility measurements of Sn doped C_{60} superconductor [8]. Saito and Oshiyama [9] calculated electronic structures of Ca_3C_{60} and Ca_5C_{60} solids and Kortan et al. [10] found by measurements that the high concentration Ca doped C_{60} solid experienced FCC → BCC (body-centered cubic) → SC(simple cubic) phase transformations for the Ca concentration from Ca_3C_{60} → Ca_4C_{60} → Ca_5C_{60}. By doping Rb and Tl alloy to C_{60} solids, Iqbal et al. [11] increased the Tc to 45 °K. Umemoto and Saito [12] calculated body-centered-orthorhombic fulleride

* Corresponding author.
E-mail address: zhao@phys.subr.edu (G.L. Zhao).

0925-9635/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.diamond.2007.12.033

Ba_4C_{60} using local density approximation(LDA), they explained the observed lattice-constant differences, $a>b>c$. Korenivski and Rao [13] evaluated the BCC phase Ba_6C_{60} using SQUID measurements and found the upper critical field H_{c2} to be approximately 2 T. Alkali metal-doped C_{60} solids at high concentrations (with the alkali atoms in the interstitial sites of bulk C_{60} solids) have been intensively investigated (see Ref. [14–22] and Ref. [23] for a review). We are aware of no detailed report of experiment or theoretical work on the electronic properties of C_{60} semiconductors with low B, N and Co doping concentrations. In this brief report, we present the results of utilizing an *ab initio* plane-wave pseudopotential method to calculate the changes of the electronic structure of the above three doped C_{60} semiconductors.

2. Computation methods

We implemented the first-principle density functional calculations using the projector augmented wave (PAW) method, taking the relativistic effects into account [24,25]. The exchange-correlation interaction was described by the generalized gradient approximation (GGA). The Vienna *ab initio* simulation package (VASP) [26–29] was used in these calculations. The 2s and 2p electron states of C, B and N atoms were described as valence states, whereas for Co atom, the 3d and 4s states were treated as valences. The core electron states were treated as those of free atoms in a frozen core approximation. We used a super-cell approach that includes 60 carbon atoms and one doping atom (1:60 doping concentration) as well as 240 carbon atoms and one doping atom (1:240 doping concentration) in the comparative calculations. All the atomic coordinates and unit-cell volumes were relaxed in the *ab initio* DFT calculations. We implemented spin-polarized electron density calculations. With the plane-wave energy cutoff at 450 eV, the calculated total energies converged to the order of about 0.01 meV. The residue forces on atoms were less than 10 meV/Å. In the super-cell method, we used a $4\times4\times4$ and $1\times1\times1$ Monkhost grids in the k space sampling for the 1:60 and 1:240 doping concentrations, respectively. The Bader charge [30] was calculated for both the dopant atoms and the host C_{60} atoms.

3. Results and discussions

The calculated results of B, N, and Co-doped C_{60} semiconductors with 1:60 concentration are summarized in Table 1. The calculated total and partial electron density of states (DOS) are presented in Figs. 1–3, where the Fermi level (E_F) was set at 0.0 eV. As can be seen from Table 1 and Figs. 1–3, nitrogen doped C_{60} solid is a p-type semiconductor, while boron and cobalt doped C_{60} solids are n-type semiconductors. There are small expansions of +0.27 % (for B and N) and +0.415% (for Co) of the unit-cell volumes in the three doping cases at the tetrahedral site of FCC C_{60} solids. The boron and nitrogen doped C_{60} have no net magnetic moment. Interestingly, the net magnetic moment of cobalt doped C_{60} solid is 3 μ_B, which is nearly the same as that of a free cobalt atom.

Table 1
Summary results of B, N, and Co-doped (1:60 concentration) C60 semiconductors

Dopant	Dopant site	Type	$\Delta V/V$	$M_B(\mu_B)$	ΔQ (e)
B	Tetra	n	+0.27%	0.0	+0.29
N	Tetra	p	+0.27%	0.0	−0.26
Co	Tetra	n	+0.415%	3.0	+0.415

The positive and negative signs of ΔQ mean losing and gaining electrons in |e|. The positive and negative signs of $\Delta V/V$ denote expansion and contraction of unit-cell volume, respectively. M_B is the magnetic moment of the system in μ_B.

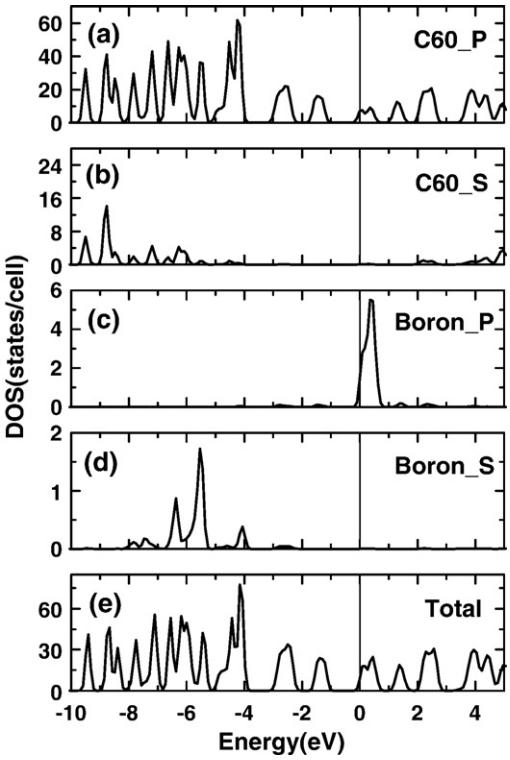


Fig. 1. The partial and total densities of states of B-doped C_{60} solid. The Fermi energy is set at 0.0 eV.

There is a small charge transfer of about +0.29|e| from the dopant B atom to the carbon atoms of C_{60}. Similarly, the charge transfer from the dopant Co atom to the carbon atoms of C_{60} is about +0.415|e|. The charge transfer from the dopant N atom to the carbon atoms of C_{60} is negative, i.e. −0.26|e|, which is consistent with the character of p-type doped semiconductors. Different from the substitutional doping in n- or p-type silicon semiconductors, the dopants in C_{60} solids occupy interstitial sites of the FCC lattice because of the high stability of the C_{60} fullerene structure and a weak interaction between them. The charge transfer properties can be understood from the relative weak electron affinities of B atom (at 0.28 eV) and Co atom (at 0.66 eV) in comparison with that of carbon atom (at 1.26 eV) [31]. The charge transfers also agree qualitatively with the octet rule which suggest carbon losing electrons to nitrogen but not to boron and cobalt. The results are also consistent with the partial DOS of B and N-doped C_{60} solid in Figs. 1 and 2. The boron p-states form the impurity electron states that merge with the electron states of the conduction band edge of host C_{60} material

around the Fermi level, as in Fig. 1. Miyamoto et al. [4] calculated the substitutionally doped BC_{59} using a local density approximation (LDA). From the calculated results of DOS, they found that the BC_{59} is a hole doped fullerene. The DOS of dopant boron is situated about 0.2 eV above the valence band of C_{59}. The results reported here for C_{60} solids interstitially doped with B are different from those of Miyamoto et al. We find the DOS of the dopant boron to be close to the bottom of the conduction band of the system and the doped material is an n-type semiconductor.

The nitrogen p-states form a distinct peak structure that is located above the top of the valence band of host C_{60} solid, as in Fig. 2. Therefore, for interstitial doping, nitrogen doped FCC C_{60} solid has an electronic structure similar to that of a p-type semiconductor, with shallow impurity energy levels near the top of the valence bands of the host material.

In all of the three doped C_{60} semiconductors considered, the total energy is lower for the dopants at the tetrahedral site than for other sites such as the octahedral sites. Consequently, in this article, we report the results of the dopant B, N, and Co atoms at the tetrahedral site of C_{60} host material. From the partial DOS in Figs. 1 and 2, we can see that C_{60} also has a contribution to the DOS near the Fermi level. This is due to the hybridization of the C 2p state with states of the corresponding dopant atoms. Fig. 3 shows the spin-polarized electron density of states for Co-doped C_{60} semiconductors. The spin up and down DOS of Co-doped C_{60} solid have noticeably different structures near the Fermi level. For the cases of lower doping concentrations at 1:240, we only observed that the density of states due to the dopant atoms

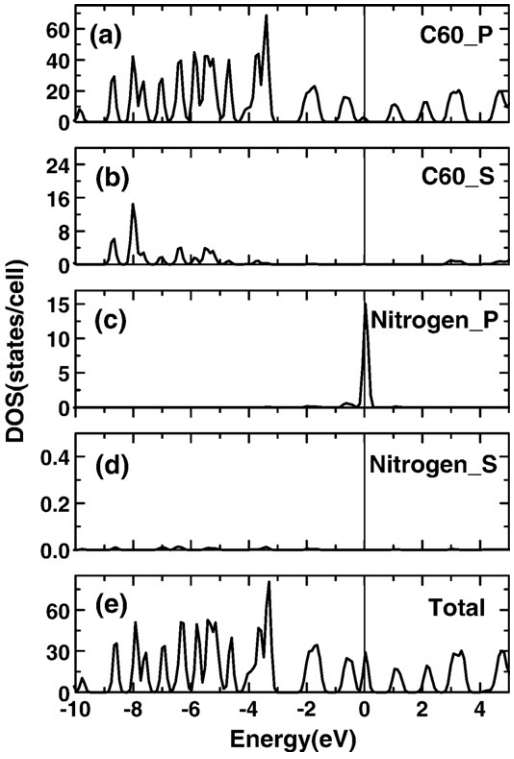


Fig. 2. The partial and total densities of states of N-doped C_{60}. The Fermi level is at 0.0 eV.

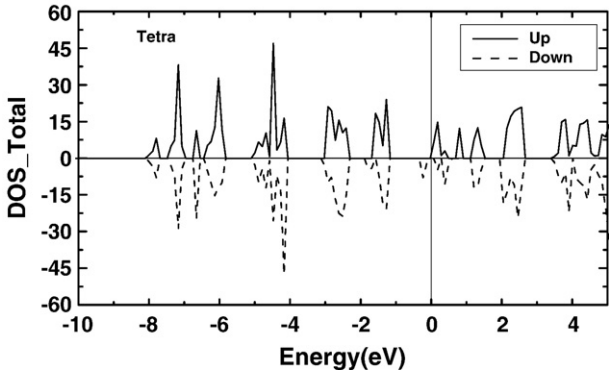


Fig. 3. The spin up and down density of states of Co-doped C_{60} solid. The Fermi energy is at 0.0 eV.

(B, N, or Co) decreased while the general structure of the DOS remained nearly the same.

4. Conclusions

In summary, we utilized the *ab initio* DFT method to calculate the electronic structures of B, N, and Co-doped C_{60} semiconductors. Both the B and Co-doped C_{60} solids are n-type semiconductors. The N-doped C_{60} solid is a p-type semiconductor. The calculated electronic properties of B, N, and Co-doped C_{60} semiconductors will facilitate the exploration of the new materials for the near future applications.

Acknowledgement

This work was funded in part by the US National Science Foundation (NSF Award No. CCF-0508245), by the Department of the Navy, Office of Naval Research (Grant No. N00014-05-1-0009), and by a Capacity Building grant through the CITI office at Southern University and A & M College.

References

[1] H.W. Kroto, J.R. Heath, S.C. O'Brien, R.F. Curl, R.E. Smalley, Nature 318 (1985) 162.
[2] W. Kratschmer, L.D. Lamb, K. Fostiropoulos, D.R. Hoffman, Nature 347 (1991) 354.
[3] R.C. Yu, N. Tea, M.B. Salamon, D. Lorens, R. Malhotra, Phys. Rev. Lett. 68 (1992) 2050.
[4] Y. Miyamoto, N. Hamada, A. Oshiyama, S. Saito, Phys. Rev. B 46 (1992) 1749.
[5] W.Y. Ching, M.Z. Huang, Y.N. Xu, W.G. Harter, F.T. Chan, Phys. Rev. Lett. 67 (1991) 2045.
[6] Y.N. Xu, M.Z. Huang, W.Y. Ching, Phys. Rev. B 44 (1991-I) 13171.
[7] Y.N. Xu, M.Z. Huang, W.Y. Ching, Phys. Rev. B 46 (1992-I) 4241.
[8] Z. Gu, J. Qian, Z. Jin, X. Zhou, S. Feng, W. Zhou, X. Zhu, Solid State Comm. 82 (1992) 167.
[9] S. Saito, A. Oshiyama, Solid State Comm. 83 (1992) 107.
[10] A.R. Kortan, N. Kopylov, S. Glarum, E.M. Gyorgy, A.P. Ramirez, R.M. Fleming, F.A. Thiel, R.C. Haddon, Nature 355 (1992) 529.
[11] Z. Iqbal, R.H. Baughman, B.L. Ramakrishna, S. Khare, N.S. Murthy, H.J. Bornemann, D.E. Morris, Science 254 (1991) 826.
[12] K. Umemoto, S. Saito, Phys. Rev. B 61 (2000) 14204.
[13] V. Korenivski, K.V. Rao, J. Supercond. 8 (1995) 67.
[14] S. Satpathy, V.P. Antropov, O.K. Andersen, O. Jepsen, O. Gunnarsson, A.I. Liechtenstein, Phys. Rev. B 46 (1992) 1773.

[15] G. Dresselhaus, M.S. Dresselhaus, P.C. Eklund, Phys. Rev. B 45 (1992) 6923.
[16] R.C. Haddon, Pure and Appl.Chem. 65 (1993) 11.
[17] L.D. Rotter, Z. Schlesinger, J.P. McCauley Jr, N. Coustel, J.E. Fischer, A.B. Smith III, Nature 355 (1992) 532.
[18] Z. Gedik, S. Ciraci, Phys. Rev. B 45 (1992) 8213.
[19] S.C. Erwin, W.E. Pickett, Science 254 (1991) 600.
[20] A.F. Hebard, M.J. Rosseinsky, R.C. Haddon, D.W. Murphy, S.H. Glarum, T.T.M. Palstra, A.P. Ramirez, A.R. Kortan, Nature 350 (1991) 600.
[21] C.B. Zhu, Z.J. Xu, J.M. Yan, Sci. China Series B-Chem. 40 (1997) 72.
[22] L.Y. Zheng, Y.N. Chiu, S.T. Lai, R.D. Letelier, J. Mol. Struct.: Theochem 712 (2004) 149.
[23] M.S. Dresselhaus, D. Dresselhaus, P.C. Eklund, Science of Fullerenes and Carbon Nanotubes, Academic Press, New York, 1996 Ch. 12.
[24] G. Kresse, D. Joubert, Phys. Rev. B 59 (1999) 1758.
[25] P.E. Blöchl, Phys. Rev. B 50 (1994) 17953.
[26] G. Kresse, J. Hafner, Phys. Rev. B 47 (1993) 558.
[27] G. Kresse, J. Furthmüller, Comp. Mater. Sci. 6 (1996) 15.
[28] G. Kresse, J. Furthmüller, Phys. Rev. B 54 (1996) 11169.
[29] VASP 2003 manual, see website: http://cms.mpi.univie.ac.at/vasp/.
[30] R.F.W. Bader, Atoms in Molecules—A Quantum Theory, Oxford University Press, Oxford, 1990.
[31] CRC Handbook of Chemistry and Physics70th Edition, CRC Press, Boca Raton, Florida, 1989–1990.

Extraction of Human DNA Replication Timing Patterns from Discrete Microarray Data

Anindya Dutta[1], Neerja Karnani[1], Ankit Malhotra[1], Gabriel Robins[2], and Christopher M. Taylor[1,2,3,4]

1-Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908
2-Department of Computer Science, University of Virginia, Charlottesville, VA 22904
3-Department of Computer Science, University of New Orleans, New Orleans, LA 70148
4-Research Institute for Children, Children's Hospital, New Orleans, LA 70118
{adq8,nk6t,am3cp,gr3e}@virginia.edu,{taylor}@cs.uno.edu

Abstract. Effective reproduction is essential for the survival and proliferation of any organism, from the birth of new offspring to the reproduction of individual cells. Each portion of a cell's DNA must be copied exactly once during the replication phase of its cell cycle to ensure viability. In humans, this is achieved by a complex pattern of replication origins and terminations along the chromosomes until the final product is realized. DNA Tiling Microarrays are utilized to assay discrete pools of DNA replicated during different parts of the replication phase. We present a generalized framework for analyzing this discrete timing data to recover a relatively continuous profile of the DNA replication timing. This approach can be used to assay DNA replication timing over a variety of human cell lines or extended to other organisms.

Key words: human replication timing, DNA tiling microarrays

1 Introduction and Related Work

DNA replication is a crucial step in the life cycle of a cell as faithful reproduction of the genetic material is essential for viability of daughter cells [1]. In higher eukaryotes, this process is carried out via the firing of numerous origins of replication along the chromosomes in order to replicate the DNA in a reasonable amount of time. The replication forks emanating from these origins work in parallel to replicate the entire genome, producing a complex schedule of DNA replication timing.

The replication time of individual areas of the genome is of interest for a variety of reasons including the influence of chromatin structure, transcriptional activity, and the possibility of allelic variation in replication timing [2]. Hence, replication timing has been studied in a number of model organisms including

Saccharomyces cerevisiae [3, 4], *Schizosaccharomyces pombe* [5], and *Drosophila melanogaster* [6, 7]. Microarray technology has played a major role in many of the studies of DNA replication timing [8–10], and more recent studies have extended these techniques to human cell lines [11–16].

One important method used for high resolution studies of DNA replication timing is the isolation of discrete pools of DNA replicated during different parts of S-phase, followed by their hybridization to genome tiling microarrays. We have adopted this method for our work in human DNA replication timing, and developed algorithms to analyze such data effectively and efficiently. In this paper, we present algorithms and techniques for recovery of a relatively continuous profile of DNA replication timing from these discrete pools of replicated DNA.

2 Methods

2.1 Data Collection

The starting point for our analysis is a set of discrete pools of DNA replicated during different parts of S-phase that have each been hybridized to a tiling microarray. In order to harvest enough DNA for the arrays, biologists synchronize a population of cells at the entry point of S-phase. The cells are then released together into the replication phase. Labeling methods are used to isolate the portions of the DNA which replicate during each part of S-phase. This synchronization and release can introduce non-trivial 'synchronization error' whereby each cell c of the population moves with some delay Δ_c with respect to the actual time of release t. Hence, the time at which cell c begins its replication phase is not t, as desired, but actually $t + \Delta_c$. In the case of application and removal of drugs to achieve synchronization, the delay for each individual cell amounts to the amount of time it takes for the cell to recover after the drug has been removed. This can be viewed as a stochastic process.

The magnitude of synchronization error present with a given technique must be accounted for when designing the length of the labeling periods to be used in the experiment. In general, the labeling periods should be made at least twice as long as the expected synchronization error from the chosen technique. We have used instantiations of this experimental design to investigate replication timing in chromosomes 21 and 22 [15], and the ENCODE regions[1] [16]. Below, we present a generalized framework for analyzing this type of experimental data.

Labeling periods begin with the start of S-phase, denoted as 0 hour. The length of each labeling period, L, provides a delicate balance between temporal resolution and resistance to synchronization error. Larger values of L decrease temporal resolution as all DNA replicating within a single time period appears on the same array. However, as smaller values of L approach the expected synchronization error, noise introduced across time period boundaries increases.

The first time period is 0-to-L hours, followed by L-to-2L hours, 2L-to-3L hours, and so on. The length of S-phase in the cell line under consideration guides

[1] The ENCODE regions [12] comprise approximately 1% of the human genome.

the choice of how many time periods to assay. Each time period is labeled and hybridized to its own array set, so the cost of the experiment increases linearly with the number of time periods. For this reason, it is sometimes desirable to use less time periods than would cover the full duration of S-phase since the amount of replicated DNA tends to fall dramatically near the end of the replication phase [15].

2.2 Time of Replication of 50% (TR50) of a Locus

The ultimate goal of a DNA replication timing study is to identify, with as much precision as possible, the actual time during S-phase that a given locus replicated. Under ideal circumstances, a given probe on the array set will display signal[2] in a single time period, with no signal across the other time periods. In this case, the locus in question has replicated sometime during the time period that displays signal. However, this scenario is rare due to synchronization error, tiling array artifacts such as cross-hybridization, and allelic variation in replication timing. The approach we take is to compute a value called the Time of Replication of 50% (TR50) for each probe in the array set. This TR50 value is a linear interpolation of the time at which the cumulative signal across all time periods for the probe passes the 50% point. We denote the signal for probe p in time period X-to-Y as $(\text{X-to-Y})_p$.

The steps to compute the TR50 value for probe p are as follows:

Normalization - Remove baseline signal present across all time periods

```
// Find the minimum signal value of all of the time periods
Minimum = min((0-to-L)_p, (L-to-2L)_p, ...);

// Subtract the minimum signal value from each of the time periods
(0-to-L)_p = (0-to-L)_p - Minimum;
(L-to-2L)_p = (L-to-2L)_p - Minimum;
...
```

Linear Interpolation - Calculate the TR50 value

```
Total = (0-to-L)_p + (L-to-2L)_p + ...; // Sum signal of all time periods

If (Total == 0) // Skip probes with 0 total signal
{ skip this probe; }

// Find the point at which 50% cumulative signal is passed
X = 0; // Start at the first time period
Cum = (0-to-L)_p; // Start with the first time period's signal
While (Cum < (Total / 2.0)) // Check for 50% of total signal
```

[2] Signal for a probe on the array is maintained as a general concept throughout the paper in order to be applicable to both arrays that have only perfect match (PM) probes and arrays that pair a mis-match (MM) probe with each PM probe. In the latter case any negative signals created by the MM probe having higher intensity than the PM probe are truncated to 0. No negative signals are allowed on the array.

```
{
        X = X + L; // Move to the next time period
        Cum = Cum + (X-to-X+L)_p; // Add next time period's signal
}

// Perform linear interpolation (X is the beginning of the time period
// where the cumulative signal surpassed 50% of the total signal)
TR50 = X + L * ((0.5 * Total - (Cum - (X-to-X+L)_p)) / (X-to-X+L)_p);
```

2.3 Temporal Specificity and Allelic Variation

The TR50 value provides an estimate of the time when the majority of replication occurs for a given locus in cases where all alleles at the locus replicate synchronously. This is called temporally specific replication (TSR). However, it has been well documented that different alleles at a given locus can replicate asynchronously [17–26]. This phenomenon, which we denote as temporally non-specific replication (TNSR), can produce a misleading result for the TR50 value. With TNSR, the TR50 value gives the average replication time over all alleles, which can produce a value at a time when no allele was being replicated. For this reason, it is important to identify and separate TNSR probes from TSR probes, which we do via our Temporal Specificity Algorithm.

Many normal cell lines are diploid in nature, having two copies of each chromosomal locus. However, HeLa cells, which we have used in some of our work, typically exhibit three copies of each chromosomal locus [16]. Tetraploidy, having four homologous sets of chromosomes, is common in plants and appears in some insects, amphibians, and reptiles [27]. We have generalized our Temporal Specificity Algorithm for application to cell lines that exhibit N copies of each chromosomal locus. Though there can be exceptions to the general ploidy in any given cell line, N should be set to the most prevalent occurrence of copy number in the cell line. For cell lines that exhibit more than one very common copy number, the larger value should be chosen for N. Having N larger than the actual copy number will perform more accurate classification than when the value of N is less than the actual copy number.

The steps of the Temporal Specificity Algorithm are as follows:

Normalization - This step is the same as in the TR50 calculation[3]

```
Total = (0-to-L)_p + (L-to-2L)_p + ...; // Sum signal of all time periods

If (Total == 0) // Skip probes with 0 total signal
{ skip this probe; }

// Find the maximum sum of all sets of two adjacent time periods
Maxsum = max((0-to-L)_p+(L-to-2L)_p, (L-to-2L)_p+(2L-to-3L)_p, ...);

// Find the maximum signal value of all of the time periods
```

[3] In practice the TR50 calculation and Temporal Specificity Algorithm are computed together, but they are presented separately here for clarity.

```
Maximum = max((0-to-L)_p, (L-to-2L)_p, ...);

// Find the maximum sum of all sets of two adjacent time periods that
// does not include the maximum signal value in either time period
Maxsumnot = 0;
X = L; // start X at the beginning of the 2nd time period
While ((X-to-X+L)_p exists)
{
        If (((X-L-to-X)_p < Maximum) and ((X-to-X+L)_p < Maximum))
        { // Neither time period includes the maximum signal
                Maxsumnot = max(Maxsumnot, (X-L-to-X)_p + (X-to-X+L)_p);
        }
}

If (Maxsum > (1 - 1/N) * Total) // Are all alleles replicating together?
{ classify probe as TSR; }
Else If (Maxsumnot >= (1/N) * Total) // Is at least one allele separate?
{ classify probe as TNSR; }
Else // Isolated signal is not strong enough to represent an allele.
{ classify probe as TSR; }
```

This classification scheme might seem arcane at first because it has been evolved over a number of attempts to classify the probes correctly. The final algorithm was arrived at after a thorough combinatorial analysis of the possible positions of replicating alleles with respect to time periods and their boundaries. We elucidate the reasoning behind each part of the algorithm in detail below.

Our original attempts to classify probes focused on the signal of each time period individually. However, due to the presence of synchronization error in the population, loci that replicate near the boundary of two adjacent time periods can contribute significant signal to both. This causes such loci to appear to undergo TNSR, even though the alleles may actually replicate together near the boundary. To address this issue we adopted the strategy of summing adjacent time periods. The sum of two adjacent time periods gives a view of the replication that occurs in either time period or on the boundary between them.

The first step of the classification algorithm is to determine if there is strong evidence that all alleles replicated together. The candidate set of adjacent time periods is selected by finding the maximum sum of signal for any set of two adjacent time periods. If this sum exceeds (1 - $1/N$) of the total signal, we classify the probe as TSR. This implies that less than $1/N$ of the total signal is contained in the other time periods. With N alleles at the locus, each individual allele is expected to contribute $1/N$ of the total signal across all time periods.

The second step is only performed if the first step failed to yield strong evidence for all alleles replicating together. In the second step, we look for evidence that at least one allele is replicating apart from a time period with the maximum signal value. We already know (since the first step failed) that at least $1/N$ of the signal is isolated from the two adjacent time periods that contained the maximum sum. The objective here is to determine if the signal that is isolated from the maximum signal value is concentrated enough to represent at least one allele.

To test this, we find the maximum sum of two adjacent time periods that does not include a time period with the maximum signal. Note that the maximum signal does not have to appear in one of the two time periods that contributed to the sum in the first step. Hence this test is subtly unrelated to the first. If this sum is at least $1/N$ of the total signal, then there is evidence for at least one allele replicating apart from the majority of signal. Namely, the evidence is for an allele to be replicating in one of the two time periods that produced this sum or on the boundary between them. In this case the probe is classified as TNSR.

Lastly, if the second test fails to yield evidence for an allele replicating apart from the majority of signal, then we consider the remaining scattered signal to be due to array artifacts and classify the probe as TSR.

2.4 Segregation of Temporally Specific and Temporally Non-specific Area

The probe data computed by the TR50 and Temporal Specificity Algorithm is very noisy due to cross-hybridization and other microarray artifacts. To address this, we take advantage of the fact that the replication mechanism provides us with spatial locality for replicated segments. As a replication fork proceeds, it causes adjacent loci on the chromosomes to replicate at similar times until the fork stalls or meets DNA that has already been replicated.

We pass a sliding window over each chromosomal sequence in order to generate broad regions of replication. The first task is to segregate TSR regions from TNSR regions. The size of the window used should be chosen to match the expected size of a replication fragment. Replication fragments will vary in size based on the length of time that the responsible replication fork operated in the given time period, so this parameter should be chosen based on what the typical expected size for a replication fragment is. A variety of strategies can be used to estimate this [16], but this is also a tunable parameter. Larger window sizes will attenuate noise in the data more; however, if the window size becomes larger than the replication fragments, then multiple of them can get merged. Smaller window sizes will suffer more from the noise inherent in the microarray data. This parameter can be increased until noise is attenuated at an acceptable level.

We also require a minimum probe density to generate intervals. If too few probes fall into a window, then such an area will not be classified for lack of tiling data. This is another tunable parameter. Setting a higher density decreases coverage of the generated intervals but increases confidence in the classification.

The sliding window is placed at the beginning of each chromosome to start segregation of the regions. As the window moves from probe to probe, the minimum probe density is tested for and when this density is exceeded a TSR interval or TNSR interval is begun based on whether there are more TSR or TNSR probes in the window. In the event of a tie, the window begins when the next probe is reached (which will break the tie). The current interval is ended when the probe density drops below the minimum level or when the TSR to TNSR probe ratio changes direction (in which case a new interval is started).

2.5 TR50 Smoothing

The TR50 values provide a noisy view of the replication timing pattern. In order to get a more continuous estimate of the replication profile, a locally weighted least squares (lowess) smoothing [28] is performed on the set of TSR probes. The smoother is set to consider all probes within the same window size used for the segregation above. Only TR50 values for the TSR probes are used because TR50 values for TNSR probes are unreliable as discussed above.

3 Results

We used the methods described in Section 2 to analyze two technical replicates and one biological replicate of the HeLa cell line (human) using Affymetrix ENCODE tiling arrays [16]. In this section, we report results pertinent to the methods themselves. Throughout this section, Replicate 1 (Rep1) and Replicate 2 (Rep2) refer to two technical replicates (the same biological sample hybridized to two sets of arrays) and Replicate 3 (Rep3) refers to the biological replicate.

Computation on the individual probes (Sections 2.2 and 2.3) performs a normalization step for probes that have no time period with 0 signal. The percentage of probes normalized during this process is shown in Figure 1.

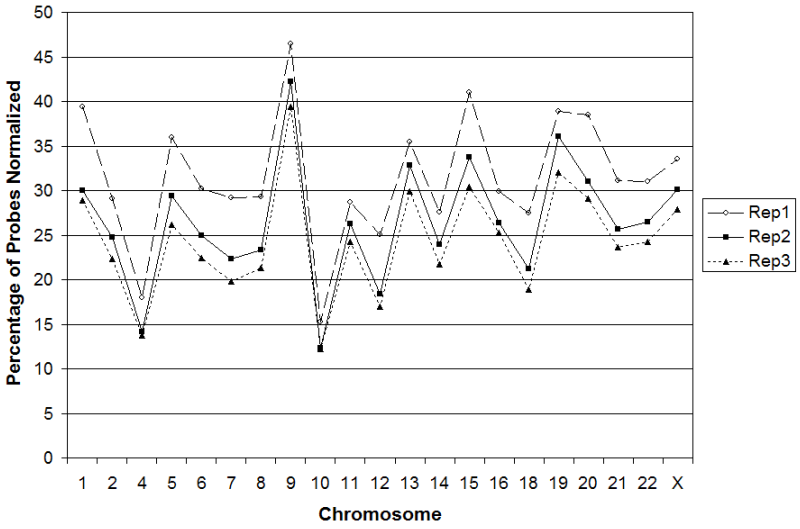


Fig. 1. Percentage of probes normalized for each replicate.

This graph plots the percentage of probes on the array where every time period had positive signal. Such probes were normalized by subtracting the minimum signal from all time periods, in order to remove baseline signal for the

given probe. The three replicates all exhibit the same trend for each chromosome, indicating that the process is indeed removing signal from array artifacts, instead of removing variations in signal between different replicates.

Figure 2 shows the percentage of probes classified as TSR for each replicate on each chromosome. In this case, Rep2 and Rep3 show the same general trend for each chromosome, while Rep1 has a more varied pattern. This underscores the importance of processing the probes through windows in the next steps since two technical replicates (Rep1 and Rep2) show varied results at the probe level.

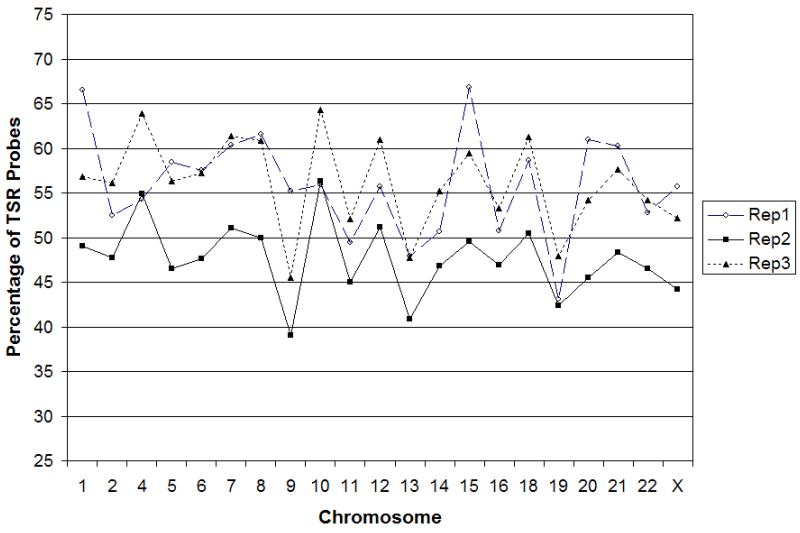


Fig. 2. Percentage of probes classified as temporally specific (TSR) per replicate.

The segregation and TR50 smoothing was done in a window of size 60,000 base pairs. This was chosen based on a profiling calculation of the expected size of replication fragments in the experiment under consideration [16].

Figure 3 shows the TR50 data for a region on chromosome 21 with the smoothed TR50 curve overlaid. Segregation of TSR regions from TNSR regions was performed with a minimum probe density of 25%. This required at least 600 probes to fall inside of the sliding 60,000 base pair window in order to generate intervals (each probe tiles 25 base pairs). The segregation intervals are shown in Figure 3 above the TR50 data.

The TR50 data at the probe level is quite noisy, but a pattern can be seen in the data where tightly grouped probes produce darker areas in the graph. The smoothed TR50 curve follows these trends closely. There is a late replicating domain (broad peak) in the graph which is surrounded by early replicating DNA. These domains have proven to be quite interesting, as the broad peak of late replication is associated with low gene density, low transcriptional activity, and a high level of repressive histone marks [2]. Further, the early replicating

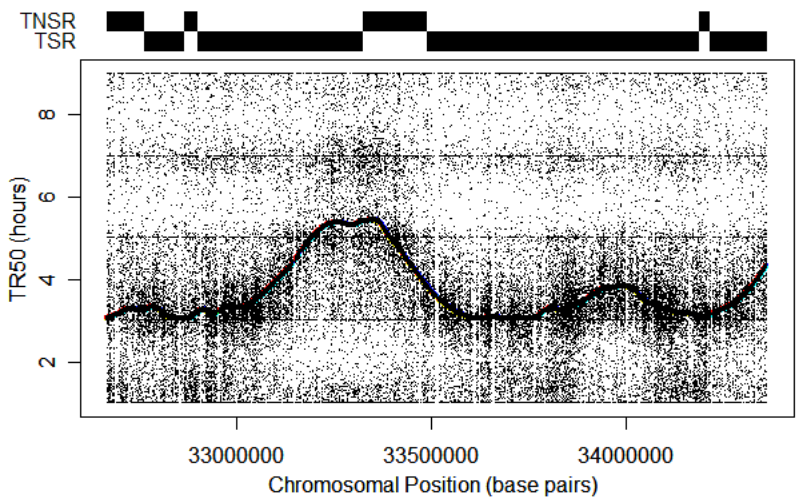


Fig. 3. Replication profile over a region of chromosome 21. Each dot in the graph corresponds to the TR50 value of a single probe. The smoothed TR50 curve is overlaid. The segregation of TSR and TNSR regions is shown above the graph.

domains surrounding this broad peak are associated with high gene density, high transcriptional activity, and high levels of activating histone marks. The troughs in the replication curve allude to possible sites of replication origin, while the peaks could be indicative of sites of fork termination. Notice from the segregation at the top of the figure that there is a large section of TNSR on the right half of the broad peak of late replication. This area is associated with high levels of both repressive and activating histone marks. We have found that on average roughly 20% of the ENCODE regions undergo TNSR [16].

All of the algorithms and techniques that we have presented to generate the replication timing profile run in linear time with respect to the size of the tiling array set used. This efficiency is achieved by using incrementally updating sliding windows. The linear runtime will allow for the general methods presented here to be utilized for whole genome analysis with moderate computational resource requirements. The replication timing profile constructed (displayed in Figure 3) produces a relatively continuous view of the replication timing in addition to identifying TNSR regions where inter-allelic variation of replication timing occurs.

4 Conclusion

We have presented a generalized framework and algorithms for analyzing a common type of DNA replication timing assay using tiling arrays. We have also discussed techniques for choosing parameters for analysis of a given replication timing array set. This approach overcomes the noise present in such tiling array

data to reconstruct a relatively continuous replication timing profile and identify areas of temporally non-specific replication. The algorithms developed have linear time complexity in the size of the tiling array set so that the approach can be used for whole genome analysis in a variety of organisms requiring only a moderate expenditure of computational resources. Lastly, we have discussed an example of the framework being applied to a set of DNA replication timing data over a small portion of the human genome. In the future, we intend to utilize this approach to analyze replication timing over the full human genome.

5 Acknowledgements and Data Availability

This work was supported by funding from the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH). Raw microarray data for the timing experiments discussed is available in ArrayExpress with accession number E-MEXP-708. Segregation and TR50 data are available as tracks in the UCSC genome browser:

http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=encodeUvaDnaRepSeg
http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=encodeUvaDnaRepTr50

References

1. Bell, S.P., Dutta, A.: DNA Replication in Eukaryotic Cells. Annu. Rev. Biochem. 2002;71:333–74.
2. ENCODE Project Consortium: Identification and Analysis of Functional Elements in 1447(7146):799–816
3. Raghuraman, M.K., Winzeler, E.A., Collingwood, D., Hunt, S., Wodicka, L., Conway, A., Lockhart, D.J., Davis, R.W., Brewer, B.J., Fangman, W.L.: Replication Dynamics of the Yeast Genome. Science. 2001 Oct 5;294(5540):115–21.
4. Lucas, I.A., Raghuraman, M.K.: The Dynamics of Chromosome Replication in Yeast. Curr. Top. Dev. Biol. 2003;55:1–73.
5. Kim, S.M., Huberman, J.A.: Regulation of Replication Timing in Fission Yeast. EMBO J. 20001 Nov 1;20(21):6115–26.
6. Schubeler, D. Scalzo, D., Kooperberg, C., van Steensel, B., Delrow, J., Groudine, M.: Genome-wide DNA Replication Profile for Drosophila Melanogaster: a Link Between Transcription and Replication Timing. Nat. Genet. 2002 Nov;32(3):438–42.
7. MacAlpine, D.M., Rodriguez, H.K., Bell, S.P.: Coordination of Replication and Transcription Along a Drosophila Chromosome. Genes. Dev. 2004 Dec 15; 18(24):3094–105.
8. McCune, H.J., Donaldson, A.D.: DNA Replication: Telling Time with Microarrays. Genome Biol. 2003;4(2):204. Epub 2003 Jan 30.
9. MacAlpine, D.M., Bell, S.P.: A Genomic View of Eukaryotic DNA Replication. Chromosome Res. 2005;13:309–26.
10. Donaldson, A.D.: Shaping Time: Chromatin Structure and the DNA Replication Programme. Trends Genet. 2005 Aug;21(8):444–9.
11. Woodfine, K., Fiegler, H., Beare, D.M., Collins, J.E., McCann, O.T., Young, B.D., Debernardi, S., Mott, R. Dunham, I., Carter, N.P.: Replication Timing of the Human Genome. Hum. Mol. Genet. 2004 Jan 15;13(2):191–202.

12. ENCODE Project Consortium: The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004 Oct 22;306(5696):636–40.
13. White, E.J., Emanuelsson, O., Scalzo, D., Royce, T., Kosak, S., Oakeley, E.J., Weissman, S., Gerstein, M., Groudine, M., Snyder, M., Schubeler, D.: DNA Replication-timing Analysis of Human Chromosome 22 at High Resolution and Different Developmental States. Proc. natl. Acad. Sci. U.S.A. 2004 Dec 21;101(51):17771–6.
14. Woodfine, K., Beare, D.M., Ichimura, K., Debernardi, S., Mungall, A.J., Fiegler, H., Collins, V.P., Carter, N.P., Dunham, I.: Replication Timing of Human Chromosome 6. Cell Cycle. 2005 Jan;4(1):172–6.
15. Jeon, Y., Bekiranov, S., Karnani, N., Kapranov, P., Ghosh, S., MacAlpine, D., Lee, C., Hwang, D.S., Gingeras, T.R., Dutta, A.: Temporal Profile of Replication of Human Chromosomes. Proc. Natl. Acad. Sci. U.S.A. 2005 May 3;102(18):6419–24.
16. Karnani, N., Taylor, C., Malhotra, A., Dutta, A.: Pan-S Replication Patterns and Chromosomal Domains Defined by Genome-tiling Arrays of ENCODE Genomic Areas. Genome Res. 2007 Jun;17(6):865–76.
17. Kitsberg, D., Selig, S., Brandeis, M., Simon, I., Keshet, I., Driscoll, D.J., Nicholls, R.D., Cedar, H.: Allele-specific Replication Timing of Imprinted Gene Regions. Nature. 1993 Jul 29;364(6436):459–63.
18. Knoll, J.H., Cheng, S.D., Lalande, M.: Allele Specificity of DNA Replication Timing in the Angelman/Prader-Willi Syndrome Imprinted Chromosomal Region. Nat. Genet. 1994 Jan;6(1):41–6.
19. Boggs, B.A., Chinault, A.C.: Analysis of Replication Timing Properties of Human X-chromosomal Loci by Fluorescence in Situ Hybridization. Proc. Natl. Acad. Sci. U.S.A. 1994 Jun 21;91(13):6083–7.
20. Carothers, A.D., Bickmore, W.A.: Models of DNA Replication Timing in Interphase Nuclei: An Exercise in Inferring Process from State. Biometrics. 1995 Jun;51(2):750–5.
21. Bickmore, W.A., Carothers, A.D.: Factors Affecting the Timing and Imprinting of Replication on a Mammalian Chromosome. J. Cell. Sci. 1995 Aug;108 (Pt 8):2801–9.
22. Kawame, H., Gartler, S.M., Hansen, R.S.: Allele-specific Replication Timing in Imprinted Domains: Absence of Asynchrony at Several Loci. Hum. Mol. Genet. 1995 Dec;4(12):2287–93.
23. Greally, J.M., Starr, D.J., Hwang, S., Song, L., Jaarola, M., Zemel, S.: The Mouse H19 Locus Mediates a Transition Between Imprinted and Non-imprinted DNA Replication Patterns. Hum. Mol. Genet. 1998 Jan;7(1):91–5.
24. Simon, I., Tenzen, T., Reubinoff, B.E., Hillman, D., McCarrey, J.R., Cedar, H.: Asynchronous Replication of Imprinted Genes is Established in the Gametes and Maintained During Development. Nature. 1999 Oct 28;401(6756):929–32.
25. Mostoslavsky, R., Singh, N., Tenzen, T., Goldmit, M., Gabay, C., Elizur, S., Qi, P., Reubinoff, B.E., Chess, A., Cedar, H., Bergman, Y.: Asynchronous Replication and Allelic Exclusion in the Immune System. Nature. 2001 Nov 8;414(6860):221–5.
26. Kagotani, K., Takebayashi, S., Kohda, A., Taguchi, H., Paulsen, M., Walter, J., Reik, W., Okumura, K.: Replication Timing Properties within the Mouse Distal Chromosome 7 Imprinting Cluster. Biosci. Biotechnol. Biochem. 2002 May;66(5):1046–51.
27. Griffiths, A.J., Wessler, S.R., Lewontin, R.C., Carroll, S.B.: Introduction to Genetic Analysis. W.H. Freeman. 2007 Feb 16;0-7167-6887-9.
28. Jacoby, W.G.: Statistical Graphics for Univariate and Bivariate Data. Sage Publications Inc. 1997 Feb 24;0-7619-0083-7.

ARTICLES

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

The ENCODE Project Consortium*

We report the generation and analysis of functional data from multiple, diverse experiments performed on a targeted 1% of the human genome as part of the pilot phase of the ENCODE Project. These data have been further integrated and augmented by a number of evolutionary and computational analyses. Together, our results advance the collective knowledge about human genome function in several major areas. First, our studies provide convincing evidence that the genome is pervasively transcribed, such that the majority of its bases can be found in primary transcripts, including non-protein-coding transcripts, and those that extensively overlap one another. Second, systematic examination of transcriptional regulation has yielded new understanding about transcription start sites, including their relationship to specific regulatory sequences and features of chromatin accessibility and histone modification. Third, a more sophisticated view of chromatin structure has emerged, including its inter-relationship with DNA replication and transcriptional regulation. Finally, integration of these new sources of information, in particular with respect to mammalian evolution based on inter- and intra-species sequence comparisons, has yielded new mechanistic and evolutionary insights concerning the functional landscape of the human genome. Together, these studies are defining a path for pursuit of a more comprehensive characterization of human genome function.

The human genome is an elegant but cryptic store of information. The roughly three billion bases encode, either directly or indirectly, the instructions for synthesizing nearly all the molecules that form each human cell, tissue and organ. Sequencing the human genome[1–3] provided highly accurate DNA sequences for each of the 24 chromosomes. However, at present, we have an incomplete understanding of the protein-coding portions of the genome, and markedly less understanding of both non-protein-coding transcripts and genomic elements that temporally and spatially regulate gene expression. To understand the human genome, and by extension the biological processes it orchestrates and the ways in which its defects can give rise to disease, we need a more transparent view of the information it encodes.

The molecular mechanisms by which genomic information directs the synthesis of different biomolecules has been the focus of much of molecular biology research over the last three decades. Previous studies have typically concentrated on individual genes, with the resulting general principles then providing insights into transcription, chromatin remodelling, messenger RNA splicing, DNA replication and numerous other genomic processes. Although many such principles seem valid as additional genes are investigated, they generally have not provided genome-wide insights about biological function.

The first genome-wide analyses that shed light on human genome function made use of observing the actions of evolution. The ever-growing set of vertebrate genome sequences[4–8] is providing increasing power to reveal the genomic regions that have been most and least acted on by the forces of evolution. However, although these studies convincingly indicate the presence of numerous genomic regions under strong evolutionary constraint, they have less power in identifying the precise bases that are constrained and provide little, if any, insight into why those bases are biologically important. Furthermore, although we have good models for how protein-coding regions evolve, our present understanding about the evolution of other functional genomic regions is poorly developed. Experimental studies that augment what we learn from evolutionary analyses are key for solidifying our insights regarding genome function.

The Encyclopedia of DNA Elements (ENCODE) Project[9] aims to provide a more biologically informative representation of the human genome by using high-throughput methods to identify and catalogue the functional elements encoded. In its pilot phase, 35 groups provided more than 200 experimental and computational data sets that examined in unprecedented detail a targeted 29,998 kilobases (kb) of the human genome. These roughly 30 Mb—equivalent to ~1% of the human genome—are sufficiently large and diverse to allow for rigorous pilot testing of multiple experimental and computational methods. These 30 Mb are divided among 44 genomic regions; approximately 15 Mb reside in 14 regions for which there is already substantial biological knowledge, whereas the other 15 Mb reside in 30 regions chosen by a stratified random-sampling method (see http://www.genome.gov/10506161). The highlights of our findings to date include:

- The human genome is pervasively transcribed, such that the majority of its bases are associated with at least one primary transcript and many transcripts link distal regions to established protein-coding loci.
- Many novel non-protein-coding transcripts have been identified, with many of these overlapping protein-coding loci and others located in regions of the genome previously thought to be transcriptionally silent.
- Numerous previously unrecognized transcription start sites have been identified, many of which show chromatin structure and sequence-specific protein-binding properties similar to well-understood promoters.

*A list of authors and their affiliations appears at the end of the paper.

- Regulatory sequences that surround transcription start sites are symmetrically distributed, with no bias towards upstream regions.
- Chromatin accessibility and histone modification patterns are highly predictive of both the presence and activity of transcription start sites.
- Distal DNaseI hypersensitive sites have characteristic histone modification patterns that reliably distinguish them from promoters; some of these distal sites show marks consistent with insulator function.
- DNA replication timing is correlated with chromatin structure.
- A total of 5% of the bases in the genome can be confidently identified as being under evolutionary constraint in mammals; for approximately 60% of these constrained bases, there is evidence of function on the basis of the results of the experimental assays performed to date.
- Although there is general overlap between genomic regions identified as functional by experimental assays and those under evolutionary constraint, not all bases within these experimentally defined regions show evidence of constraint.
- Different functional elements vary greatly in their sequence variability across the human population and in their likelihood of residing within a structurally variable region of the genome.
- Surprisingly, many functional elements are seemingly unconstrained across mammalian evolution. This suggests the possibility of a large pool of neutral elements that are biochemically active but provide no specific benefit to the organism. This pool may serve as a 'warehouse' for natural selection, potentially acting as the source of lineage-specific elements and functionally conserved but non-orthologous elements between species.

Below, we first provide an overview of the experimental techniques used for our studies, after which we describe the insights gained from analysing and integrating the generated data sets. We conclude with a perspective of what we have learned to date about this 1% of the human genome and what we believe the prospects are for a broader and deeper investigation of the functional elements in the human genome. To aid the reader, Box 1 provides a glossary for many of the abbreviations used throughout this paper.

Experimental techniques

Table 1 (expanded in Supplementary Information section 1.1) lists the major experimental techniques used for the studies reported here, relevant acronyms, and references reporting the generated data sets. These data sets reflect over 400 million experimental data points (603 million data points if one includes comparative sequencing bases). In describing the major results and initial conclusions, we seek to distinguish 'biochemical function' from 'biological role'. Biochemical function reflects the direct behaviour of a molecule(s), whereas biological role is used to describe the consequence(s) of this function for the organism. Genome-analysis techniques nearly always focus on biochemical function but not necessarily on biological role. This is because the former is more amenable to large-scale data-generation methods, whereas the latter is more difficult to assay on a large scale.

The ENCODE pilot project aimed to establish redundancy with respect to the findings represented by different data sets. In some instances, this involved the intentional use of different assays that were based on a similar technique, whereas in other situations, different techniques assayed the same biochemical function. Such redundancy has allowed methods to be compared and consensus data sets to be generated, much of which is discussed in companion papers, such as the ChIP-chip platform comparison[10,11]. All ENCODE data have been released after verification but before this publication, as befits a 'community resource' project (see http://www.wellcome.ac.uk/doc_wtd003208.html). Verification is defined as when the experiment is reproducibly confirmed (see Supplementary Information section 1.2). The main portal for ENCODE data is provided by the UCSC Genome Browser (http://genome.ucsc.edu/ENCODE/); this is

Box 1 | Frequently used abbreviations in this paper

AR Ancient repeat: a repeat that was inserted into the early mammalian lineage and has since become dormant; the majority of ancient repeats are thought to be neutrally evolving.
CAGE tag A short sequence from the 5′ end of a transcript
CDS Coding sequence: a region of a cDNA or genome that encodes proteins
ChIP-chip Chromatin immunoprecipitation followed by detection of the products using a genomic tiling array
CNV Copy number variants: regions of the genome that have large duplications in some individuals in the human population
CS Constrained sequence: a genomic region associated with evidence of negative selection (that is, rejection of mutations relative to neutral regions)
DHS DNaseI hypersensitive site: a region of the genome showing a sharply different sensitivity to DNaseI compared with its immediate locale
EST Expressed sequence tag: a short sequence of a cDNA indicative of expression at this point
FAIRE Formaldehyde-assisted isolation of regulatory elements: a method to assay open chromatin using formaldehyde crosslinking followed by detection of the products using a genomic tiling array
FDR False discovery rate: a statistical method for setting thresholds on statistical tests to correct for multiple testing
GENCODE Integrated annotation of existing cDNA and protein resources to define transcripts with both manual review and experimental testing procedures
GSC Genome structure correction: a method to adapt statistical tests to make fewer assumptions about the distribution of features on the genome sequence. This provides a conservative correction to standard tests
HMM Hidden Markov model: a machine-learning technique that can establish optimal parameters for a given model to explain the observed data
Indel An insertion or deletion; two sequences often show a length difference within alignments, but it is not always clear whether this reflects a previous insertion or a deletion
PET A short sequence that contains both the 5′ and 3′ ends of a transcript
RACE Rapid amplification of cDNA ends: a technique for amplifying cDNA sequences between a known internal position in a transcript and its 5′ end
RFBR Regulatory factor binding region: a genomic region found by a ChIP-chip assay to be bound by a protein factor
RFBR-Seqsp Regulatory factor binding regions that are from sequence-specific binding factors
RT–PCR Reverse transcriptase polymerase chain reaction: a technique for amplifying a specific region of a transcript
RxFrag Fragment of a RACE reaction: a genomic region found to be present in a RACE product by an unbiased tiling-array assay
SNP Single nucleotide polymorphism: a single base pair change between two individuals in the human population
STAGE Sequence tag analysis of genomic enrichment: a method similar to ChIP-chip for detecting protein factor binding regions but using extensive short sequence determination rather than genomic tiling arrays
SVM Support vector machine: a machine-learning technique that can establish an optimal classifier on the basis of labelled training data
TR50 A measure of replication timing corresponding to the time in the cell cycle when 50% of the cells have replicated their DNA at a specific genomic position
TSS Transcription start site
TxFrag Fragment of a transcript: a genomic region found to be present in a transcript by an unbiased tiling-array assay
Un.TxFrag A TxFrag that is not associated with any other functional annotation
UTR Untranslated region: part of a cDNA either at the 5′ or 3′ end that does not encode a protein sequence

augmented by multiple other websites (see Supplementary Information section 1.1).

A common feature of genomic analyses is the need to assess the significance of the co-occurrence of features or of other statistical tests. One confounding factor is the heterogeneity of the genome, which can produce uninteresting correlations of variables distributed across the genome. We have developed and used a statistical framework that mitigates many of these hidden correlations by adjusting the appropriate null distribution of the test statistics. We term this correction procedure genome structure correction (GSC) (see Supplementary Information section 1.3).

In the next five sections, we detail the various biological insights of the pilot phase of the ENCODE Project.

Transcription

Overview. RNA transcripts are involved in many cellular functions, either directly as biologically active molecules or indirectly by encoding other active molecules. In the conventional view of genome organization, sets of RNA transcripts (for example, messenger RNAs) are encoded by distinct loci, with each usually dedicated to a single biological role (for example, encoding a specific protein). However, this picture has substantially grown in complexity in recent years[12]. Other forms of RNA molecules (such as small nucleolar RNAs and micro (mi)RNAs) are known to exist, and often these are encoded by regions that intercalate with protein-coding genes. These observations are consistent with the well-known discrepancy between the levels of observable mRNAs and large structural RNAs compared with the total RNA in a cell, suggesting that there are numerous RNA species yet to be classified[13–15]. In addition, studies of specific loci have indicated the presence of RNA transcripts that have a role in chromatin maintenance and other regulatory control. We sought to assay and analyse transcription comprehensively across the 44 ENCODE regions in an effort to understand the repertoire of encoded RNA molecules.

Transcript maps. We used three methods to identify transcripts emanating from the ENCODE regions: hybridization of RNA (either total or polyA-selected) to unbiased tiling arrays (see Supplementary Information section 2.1), tag sequencing of cap-selected RNA at the 5′ or joint 5′/3′ ends (see Supplementary Information sections 2.2 and S2.3), and integrated annotation of available complementary DNA and EST sequences involving computational, manual, and experimental approaches[16] (see Supplementary Information section 2.4). We abbreviate the regions identified by unbiased tiling arrays as TxFrags, the cap-selected RNAs as CAGE or PET tags (see Box 1), and the integrated annotation as GENCODE transcripts. When a TxFrag does not overlap a GENCODE annotation, we call it an Un.TxFrag. Validation of these various studies is described in papers reporting these data sets[17] (see Supplementary Information sections 2.1.4 and 2.1.5).

These methods recapitulate previous findings, but provide enhanced resolution owing to the larger number of tissues sampled and the integration of results across the three approaches (see Table 2). To begin with, our studies show that 14.7% of the bases represented in the unbiased tiling arrays are transcribed in at least one tissue sample. Consistent with previous work[14,15], many (63%) TxFrags reside outside of GENCODE annotations, both in intronic (40.9%) and intergenic (22.6%) regions. GENCODE annotations are richer than the more-conservative RefSeq or Ensembl annotations, with 2,608 transcripts clustered into 487 loci, leading to an average of 5.4 transcripts per locus. Finally, extensive testing of predicted protein-coding sequences outside of GENCODE annotations was positive in only 2% of cases[16], suggesting that GENCODE annotations cover nearly all protein-coding sequences. The GENCODE annotations are categorized both by likely function (mainly, the presence of an open reading frame) and by classification evidence (for example, transcripts based solely on ESTs are distinguished from other scenarios); this classification is not strongly correlated with expression levels (see Supplementary Information sections 2.4.2 and 2.4.3).

Analyses of more biological samples have allowed a richer description of the transcription specificity (see Fig. 1 and Supplementary Information section 2.5). We found that 40% of TxFrags are present in only one sample, whereas only 2% are present in all samples. Although exon-containing TxFrags are more likely (74%) to be expressed in more than one sample, 45% of unannotated TxFrags are also expressed in multiple samples. GENCODE annotations of separate loci often (42%) overlap with respect to their genomic coordinates, in particular on opposite strands (33% of loci). Further analysis of GENCODE-annotated sequences with respect to the positions of open reading frames revealed that some component exons do not have the expected synonymous versus non-synonymous substitution patterns of protein-coding sequence (see Supplement Information section 2.6) and some have deletions incompatible with

Table 1 | Summary of types of experimental techniques used in ENCODE

Feature class	Experimental technique(s)	Abbreviations	References	Number of experimental data points
Transcription	Tiling array, integrated annotation	TxFrag, RxFrag, GENCODE	117 118 19 119	63,348,656
5′ ends of transcripts*	Tag sequencing	PET, CAGE	121 13	864,964
Histone modifications	Tiling array	Histone nomenclature†, RFBR	46	4,401,291
Chromatin‡ structure	QT-PCR, tiling array	DHS, FAIRE	42 43 44 122	15,318,324
Sequence-specific factors	Tiling array, tag sequencing, promoter assays	STAGE, ChIP-Chip, ChIP-PET, RFBR	41,52 11,120 123 81 34,51 124 49 33 40	324,846,018
Replication	Tiling array	TR50	59 75	14,735,740
Computational analysis	Computational methods	CCI, RFBR cluster	80 125 10 16 126 127	NA
Comparative sequence analysis*	Genomic sequencing, multi-sequence alignments, computational analyses	CS	87 86 26	NA
Polymorphisms*	Resequencing, copy number variation	CNV	103 128	NA

* Not all data generated by the ENCODE Project.
† Histone code nomenclature follows the Brno nomenclature as described in ref. 129.
‡Also contains histone modification.

Table 2 | Bases detected in processed transcripts either as a GENCODE exon, a TxFrag, or as either a GENCODE exon or a TxFrag

	GENCODE exon	TxFrag	Either GENCODE exon or TxFrag
Total detectable transcripts (bases)	1,776,157 (5.9%)	1,369,611 (4.6%)	2,519,280 (8.4%)
Transcripts detected in tiled regions of arrays (bases)	1,447,192 (9.8%)	1,369,611 (9.3%)	2,163,303 (14.7%)

Percentages are of total bases in ENCODE in the first row and bases tiled in arrays in the second row.

protein structure[18]. Such exons are on average less expressed (25% versus 87% by RT–PCR; see Supplementary Information section 2.7) than exons involved in more than one transcript (see Supplementary Information section 2.4.3), but when expressed have a tissue distribution comparable to well-established genes.

Critical questions are raised by the presence of a large amount of unannotated transcription with respect to how the corresponding sequences are organized in the genome—do these reflect longer transcripts that include known loci, do they link known loci, or are they completely separate from known loci? We further investigated these issues using both computational and new experimental techniques.

Unannotated transcription. Consistent with previous findings, the Un.TxFrags did not show evidence of encoding proteins (see Supplementary Information section 2.8). One might expect Un.TxFrags to be linked within transcripts that exhibit coordinated expression and have similar conservation profiles across species. To test this, we clustered Un.TxFrags using two methods. The first method[19] used expression levels in 11 cell lines or conditions, dinucleotide composition, location relative to annotated genes, and evolutionary conservation profiles to cluster TxFrags (both unannotated and annotated). By this method, 14% of Un.TxFrags could be assigned to annotated loci, and 21% could be clustered into 200 novel loci (with an average of ~7 TxFrags per locus). We experimentally examined these novel loci to study the connectivity of transcripts amongst Un.TxFrags and between Un.TxFrags and known exons. Overall, about 40% of the connections (18 out of 46) were validated by RT–PCR. The second clustering method involved analysing a time course (0, 2, 8 and 32 h) of expression changes in human HL60 cells following retinoic-acid stimulation. There is a coordinated program of expression changes from annotated loci, which can be shown by plotting Pearson correlation values of the expression levels of exons inside annotated loci versus unrelated exons (see Supplementary Information section 2.8.2). Similarly, there is coordinated expression of nearby Un.TxFrags, albeit lower, though still significantly different from randomized sets. Both clustering methods indicate that there is coordinated behaviour of many Un.TxFrags, consistent with them residing in connected transcripts.

Transcript connectivity. We used a combination of RACE and tiling arrays[20] to investigate the diversity of transcripts emanating from protein-coding loci. Analogous to TxFrags, we refer to transcripts detected using RACE followed by hybridization to tiling arrays as RxFrags. We performed RACE to examine 399 protein-coding loci (those loci found entirely in ENCODE regions) using RNA derived from 12 tissues, and were able to unambiguously detect 4,573 RxFrags for 359 loci (see Supplementary Information section 2.9). Almost half of these RxFrags (2,324) do not overlap a GENCODE exon, and most (90%) loci have at least one novel RxFrag, which often extends a considerable distance beyond the 5′ end of the locus. Figure 2 shows the distribution of distances between these new RACE-detected ends and the previously annotated TSS of each locus. The average distance of the extensions is between 50 kb and 100 kb, with many extensions (>20%) being more than 200 kb. Consistent with the known presence of overlapping genes in the human genome, our findings reveal evidence for an overlapping gene at 224 loci, with transcripts from 180 of these loci (~50% of the RACE-positive loci) appearing to have incorporated at least one exon from an upstream gene.

To characterize further the 5′ RxFrag extensions, we performed RT–PCR followed by cloning and sequencing for 550 of the 5′ RxFrags (including the 261 longest extensions identified for each locus). The approach of mapping RACE products using microarrays is a combination method previously described and validated in several studies[14,17,20]. Hybridization of the RT–PCR products to tiling arrays confirmed connectivity in almost 60% of the cases. Sequenced clones confirmed transcript extensions. Longer extensions were harder to clone and sequence, but 5 out of 18 RT–PCR-positive extensions over 100 kb were verified by sequencing (see Supplementary Information section 2.9.7 and ref. 17). The detection of numerous RxFrag extensions coupled with evidence of considerable intronic transcription indicates that protein-coding loci are more transcriptionally complex than previously thought. Instead of the traditional view that many genes have one or more alternative transcripts that code for alternative proteins, our data suggest that a given gene may both encode multiple protein products and produce other transcripts that include sequences from both strands and from neighbouring loci (often without encoding a different protein). Figure 3 illustrates such a case, in which a new fusion transcript is expressed in the small intestine, and consists of at least three coding exons from the *ATP5O* gene and at least two coding exons from the *DONSON*

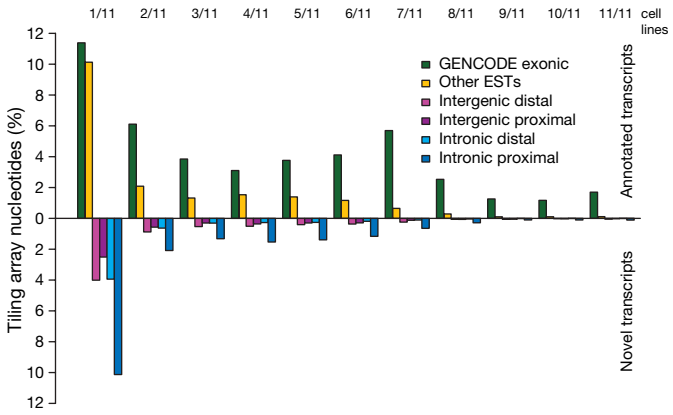


Figure 1 | Annotated and unannotated TxFrags detected in different cell lines. The proportion of different types of transcripts detected in the indicated number of cell lines (from 1/11 at the far left to 11/11 at the far right) is shown. The data for annotated and unannotated TxFrags are indicated separately, and also split into different categories based on GENCODE classification: exonic, intergenic (proximal being within 5 kb of a gene and distal being otherwise), intronic (proximal being within 5 kb of an intron and distal being otherwise), and matching other ESTs not used in the GENCODE annotation (principally because they were unspliced). The *y* axis indicates the per cent of tiling array nucleotides present in that class for that number of samples (combination of cell lines and tissues).

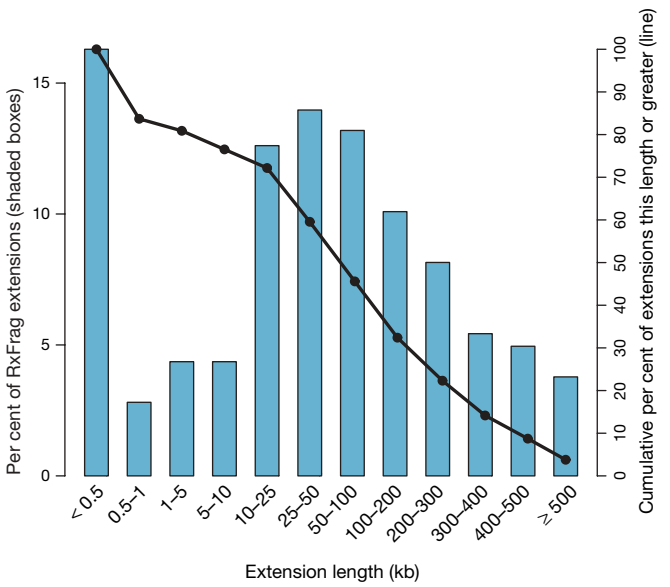


Figure 2 | Length of genomic extensions to GENCODE-annotated genes on the basis of RACE experiments followed by array hybridizations (RxFrags). The indicated bars reflect the frequency of extension lengths among different length classes. The solid line shows the cumulative frequency of extensions of that length or greater. Most of the extensions are greater than 50 kb from the annotated gene (see text for details).

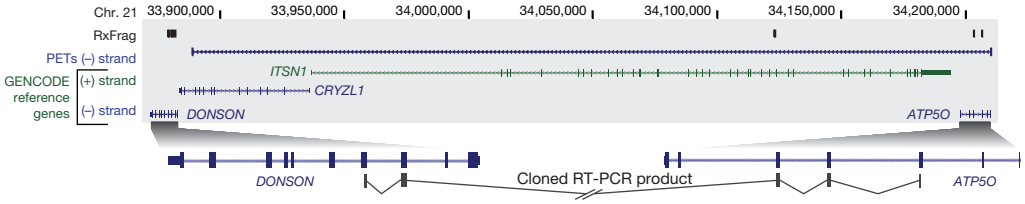


Figure 3 | Overview of RACE experiments showing a gene fusion. Transcripts emanating from the region between the *DONSON* and *ATP5O* genes. A 330-kb interval of human chromosome 21 (within ENm005) is shown, which contains four annotated genes: *DONSON*, *CRYZL1*, *ITSN1* and *ATP5O*. The 5′ RACE products generated from small intestine RNA and detected by tiling-array analyses (RxFrags) are shown along the top. Along the bottom is shown the placement of a cloned and sequenced RT–PCR product that has two exons from the *DONSON* gene followed by three exons from the *ATP5O* gene; these sequences are separated by a 300 kb intron in the genome. A PET tag shows the termini of a transcript consistent with this RT–PCR product.

gene, with no evidence of sequences from two intervening protein-coding genes (*ITSN1* and *CRYZL1*).

Pseudogenes. Pseudogenes, reviewed in refs 21 and 22, are generally considered non-functional copies of genes, are sometimes transcribed and often complicate analysis of transcription owing to close sequence similarity to functional genes. We used various computational methods to identify 201 pseudogenes (124 processed and 77 non-processed) in the ENCODE regions (see Supplementary Information section 2.10 and ref. 23). Tiling-array analysis of 189 of these revealed that 56% overlapped at least one TxFrag. However, possible cross-hybridization between the pseudogenes and their corresponding parent genes may have confounded such analyses. To assess better the extent of pseudogene transcription, 160 pseudogenes (111 processed and 49 non-processed) were examined for expression using RACE/tiling-array analysis (see Supplementary Information section 2.9.2). Transcripts were detected for 14 pseudogenes (8 processed and 6 non-processed) in at least one of the 12 tested RNA sources, the majority (9) being in testis (see ref. 23). Additionally, there was evidence for the transcription of 25 pseudogenes on the basis of their proximity (within 100 bp of a pseudogene end) to CAGE tags (8), PETs (2), or cDNAs/ESTs (21). Overall, we estimate that at least 19% of the pseudogenes in the ENCODE regions are transcribed, which is consistent with previous estimates[24,25].

Non-protein-coding RNA. Non-protein-coding RNAs (ncRNAs) include structural RNAs (for example, transfer RNAs, ribosomal RNAs, and small nuclear RNAs) and more recently discovered regulatory RNAs (for example, miRNAs). There are only 8 well-characterized ncRNA genes within the ENCODE regions (*U70, ACA36, ACA56, mir-192, mir-194-2, mir-196, mir-483* and *H19*), whereas representatives of other classes, (for example, box C/D snoRNAs, tRNAs, and functional snRNAs) seem to be completely absent in the ENCODE regions. Tiling-array data provided evidence for transcription in at least one of the assayed RNA samples for all of these ncRNAs, with the exception of mir-483 (expression of mir-483 might be specific to fetal liver, which was not tested). There is also evidence for the transcription of 6 out of 8 pseudogenes of ncRNAs (mainly snoRNA-derived). Similar to the analysis of protein-pseudogenes, the hybridization results could also originate from the known snoRNA gene elsewhere in the genome.

Many known ncRNAs are characterized by a well-defined RNA secondary structure. We applied two *de novo* ncRNA prediction algorithms—EvoFold and RNAz—to predict structured ncRNAs (as well as functional structures in mRNAs) using the multi-species sequence alignments (see below, Supplementary Information section 2.11 and ref. 26). Using a sensitivity threshold capable of detecting all known miRNAs and snoRNAs, we identified 4,986 and 3,707 candidate ncRNA loci with EvoFold and RNAz, respectively. Only 268 loci (5% and 7%, respectively) were found with both programs, representing a 1.6-fold enrichment over that expected by chance; the lack of more extensive overlap is due to the two programs having optimal sensitivity at different levels of GC content and conservation. We experimentally examined 50 of these targets using RACE/tiling-array analysis for brain and testis tissues (see Supplementary Information sections 2.11 and 2.9.3); the predictions were validated at a 56%, 65%, and 63% rate for Evofold, RNAz and dual predictions, respectively.

Primary transcripts. The detection of numerous unannotated transcripts coupled with increasing knowledge of the general complexity of transcription prompted us to examine the extent of primary (that is, unspliced) transcripts across the ENCODE regions. Three data sources provide insight about these primary transcripts: the GENCODE annotation, PETs, and RxFrag extensions. Figure 4 summarizes the fraction of bases in the ENCODE regions that overlap transcripts identified by these technologies. Remarkably, 93% of bases are represented in a primary transcript identified by at least two independent observations (but potentially using the same technology); this figure is reduced to 74% in the case of primary transcripts detected by at least two different technologies. These increased spans are not mainly due to cell line rearrangements because they were present in multiple tissue experiments that confirmed the spans (see Supplementary Information section 2.12). These estimates assume that the presence of PETs or RxFrags defining the terminal ends of a transcript imply that the entire intervening DNA is transcribed and then processed. Other mechanisms, thought to be unlikely in the human genome, such as *trans*-splicing or polymerase jumping would also produce these long termini and potentially should be reconsidered in more detail.

Previous studies have suggested a similar broad amount of transcription across the human[14,15] and mouse[27] genomes. Our studies confirm these results, and have investigated the genesis of these transcripts in greater detail, confirming the presence of substantial intragenic and intergenic transcription. At the same time, many of the resulting transcripts are neither traditional protein-coding

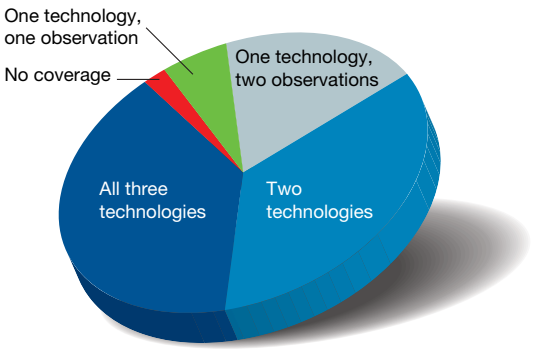


Figure 4 | Coverage of primary transcripts across ENCODE regions. Three different technologies (integrated annotation from GENCODE, RACE-array experiments (RxFrags) and PET tags) were used to assess the presence of a nucleotide in a primary transcript. Use of these technologies provided the opportunity to have multiple observations of each finding. The proportion of genomic bases detected in the ENCODE regions associated with each of the following scenarios is depicted: detected by all three technologies, by two of the three technologies, by one technology but with multiple observations, and by one technology with only one observation. Also indicated are genomic bases without any detectable coverage of primary transcripts.

transcripts nor easily explained as structural non-coding RNAs. Other studies have noted complex transcription around specific loci or chimaeric-gene structures (for example refs 28–30), but these have often been considered exceptions; our data show that complex intercalated transcription is common at many loci. The results presented in the next section show extensive amounts of regulatory factors around novel TSSs, which is consistent with this extensive transcription. The biological relevance of these unannotated transcripts remains unanswered by these studies. Evolutionary information (detailed below) is mixed in this regard; for example, it indicates that unannotated transcripts show weaker evolutionary conservation than many other annotated features. As with other ENCODE-detected elements, it is difficult to identify clear biological roles for the majority of these transcripts; such experiments are challenging to perform on a large scale and, furthermore, it seems likely that many of the corresponding biochemical events may be evolutionarily neutral (see below).

Regulation of transcription

Overview. A significant challenge in biology is to identify the transcriptional regulatory elements that control the expression of each transcript and to understand how the function of these elements is coordinated to execute complex cellular processes. A simple, commonplace view of transcriptional regulation involves five types of *cis*-acting regulatory sequences—promoters, enhancers, silencers, insulators and locus control regions[31]. Overall, transcriptional regulation involves the interplay of multiple components, whereby the availability of specific transcription factors and the accessibility of specific genomic regions determine whether a transcript is generated[31]. However, the current view of transcriptional regulation is known to be overly simplified, with many details remaining to be established. For example, the consensus sequences of transcription factor binding sites (typically 6 to 10 bases) have relatively little information content and are present numerous times in the genome, with the great majority of these not participating in transcriptional regulation. Does chromatin structure then determine whether such a sequence has a regulatory role? Are there complex inter-factor interactions that integrate the signals from multiple sites? How are signals from different distal regulatory elements coupled without affecting all neighbouring genes? Meanwhile, our understanding of the repertoire of transcriptional events is becoming more complex, with an increasing appreciation of alternative TSSs[32,33] and the presence of non-coding[27,34] and anti-sense transcripts[35,36].

To better understand transcriptional regulation, we sought to begin cataloguing the regulatory elements residing within the 44 ENCODE regions. For this pilot project, we mainly focused on the binding of regulatory proteins and chromatin structure involved in transcriptional regulation. We analysed over 150 data sets, mainly from ChIP-chip[37–39], ChIP-PET and STAGE[40,41] studies (see Supplementary Information section 3.1 and 3.2). These methods use chromatin immunoprecipitation with specific antibodies to enrich for DNA in physical contact with the targeted epitope. This enriched DNA can then be analysed using either microarrays (ChIP-chip) or high-throughput sequencing (ChIP-PET and STAGE). The assays included 18 sequence-specific transcription factors and components of the general transcription machinery (for example, RNA polymerase II (Pol II), TAF1 and TFIIB/GTF2B). In addition, we tested more than 600 potential promoter fragments for transcriptional activity by transient-transfection reporter assays that used 16 human cell lines[33]. We also examined chromatin structure by studying the ENCODE regions for DNaseI sensitivity (by quantitative PCR[42] and tiling arrays[43,44], see Supplementary Information section 3.3), histone composition[45], histone modifications (using ChIP-chip assays)[37,46], and histone displacement (using FAIRE, see Supplementary Information section 3.4). Below, we detail these analyses, starting with the efforts to define and classify the 5′ ends of transcripts with respect to their associated regulatory signals. Following that are summaries of generated data about sequence-specific transcription factor binding and clusters of regulatory elements. Finally, we describe how this information can be integrated to make predictions about transcriptional regulation.

Transcription start site catalogue. We analysed two data sets to catalogue TSSs in the ENCODE regions: the 5′ ends of GENCODE-annotated transcripts and the combined results of two 5′-end-capture technologies—CAGE and PET-tagging. The initial results suggested the potential presence of 16,051 unique TSSs. However, in many cases, multiple TSSs resided within a single small segment (up to ~200 bases); this was due to some promoters containing TSSs with many very close precise initiation sites[47]. To normalize for this effect, we grouped TSSs that were 60 or fewer bases apart into a single cluster, and in each case considered the most frequent CAGE or PET tag (or the 5′-most TSS in the case of TSSs identified only from GENCODE data) as representative of that cluster for downstream analyses.

The above effort yielded 7,157 TSS clusters in the ENCODE regions. We classified these TSSs into three categories: known (present at the end of GENCODE-defined transcripts), novel (supported by other evidence) and unsupported. The novel TSSs were further subdivided on the basis of the nature of the supporting evidence (see Table 3 and Supplementary Information section 3.5), with all four of the resulting subtypes showing significant overlap with experimental evidence using the GSC statistic. Although there is a larger relative proportion of singleton tags in the novel category, when analysis is restricted to only singleton tags, the novel TSSs continue to have highly significant overlap with supporting evidence (see Supplementary Information section 3.5.1).

Correlating genomic features with chromatin structure and transcription factor binding. By measuring relative sensitivity to DNaseI digestion (see Supplementary Information section 3.3), we identified DNaseI hypersensitive sites throughout the ENCODE regions. DHSs and TSSs both reflect genomic regions thought to be enriched for regulatory information and many DHSs reside at or near TSSs. We partitioned DHSs into those within 2.5 kb of a TSS (958; 46.5%) and the remaining ones, which were classified as distal (1,102; 53.5%). We then cross-analysed the TSSs and DHSs with data sets relating to histone modifications, chromatin accessibility and sequence-specific transcription factor binding by summarizing these signals in aggregate relative to the distance from TSSs or DHSs. Figure 5 shows representative profiles of specific histone modifications, Pol II and selected transcription factor binding for the different categories of TSSs. Further profiles and statistical analysis of these studies can be found in Supplementary Information 3.6.

In the case of the three TSS categories (known, novel and unsupported), known and novel TSSs are both associated with similar signals for multiple factors (ranging from histone modifications through DNaseI accessibility), whereas unsupported TSSs are not.

Table 3 | Different categories of TSSs defined on the basis of support from different transcript-survey methods

Category	Transcript survey method	Number of TSS clusters (non-redundant)*	*P* value†	Singleton clusters‡ (%)
Known	GENCODE 5′ ends	1,730	2×10^{-70}	25 (74 overall)
Novel	GENCODE sense exons	1,437	6×10^{-39}	64
	GENCODE antisense exons	521	3×10^{-8}	65
	Unbiased transcription survey	639	7×10^{-63}	71
	CpG island	164	4×10^{-90}	60
Unsupported	None	2,666	-	83.4

* Number of TSS clusters with this support, excluding TSSs from higher categories.
† Probability of overlap between the transcript support and the PET/CAGE tags, as calculated by the Genome Structure Correction statistic (see Supplementary Information section 1.3).
‡ Per cent of clusters with only one tag. For the 'known' category this was calculated as the per cent of GENCODE 5′ ends with tag support (25%) or overall (74%).

The enrichments seen with chromatin modifications and sequence-specific factors, along with the significant clustering of this evidence, indicate that the novel TSSs do not reflect false positives and probably use the same biological machinery as other promoters. Sequence-specific transcription factors show a marked increase in binding across the broad region that encompasses each TSS. This increase is notably symmetric, with binding equally likely upstream or downstream of a TSS (see Supplementary Information section 3.7 for an explanation of why this symmetrical signal is not an artefact of the analysis of the signals). Furthermore, there is enrichment of SMARCC1 binding (a member of the SWI/SNF chromatin-modifying complex), which persists across a broader extent than other factors. The broad signals with this factor indicate that the ChIP-chip results reflect both specific enrichment at the TSS and broader enrichments across ~5-kb regions (this is not due to technical issues, see Supplementary Information section 3.8).

We selected 577 GENCODE-defined TSSs at the 5′ ends of a protein-coding transcript with over 3 exons, to assess expression status. Each transcript was classified as: (1) 'active' (gene on) or 'inactive' (gene off) on the basis of the unbiased transcript surveys, and (2) residing near a 'CpG island' or not ('non-CpG island') (see Supplementary Information section 3.17). As expected, the aggregate signal of histone modifications is mainly attributable to active TSSs (Fig. 5), in particular those near CpG islands. Pronounced doublet peaks at the TSS can be seen with these large signals (similar to previous work in yeast[48]) owing to the chromatin accessibility at the TSS. Many of the histone marks and Pol II signals are now clearly asymmetrical, with a persistent level of Pol II into the genic region, as expected. However, the sequence-specific factors remain largely symmetrically distributed. TSSs near CpG islands show a broader distribution of histone marks than those not near CpG islands (see Supplementary Information section 3.6). The binding of some transcription factors (E2F1, E2F4 and MYC) is extensive in the case of active genes, and is lower (or absent) in the case of inactive genes.

Chromatin signature of distal elements. Distal DHSs show characteristic patterns of histone modification that are the inverse of TSSs, with high H3K4me1 accompanied by lower levels of H3K4Me3 and H3Ac (Fig. 5). Many factors with high occupancy at TSSs (for example, E2F4) show little enrichment at distal DHSs, whereas other factors (for example, MYC) are enriched at both TSSs and distal DHSs[49]. A particularly interesting observation is the relative enrichment of the insulator-associated factor CTCF[50] at both distal DHSs and TSSs; this contrasts with SWI/SNF components SMARCC2 and SMARCC1, which are TSS-centric. Such differential

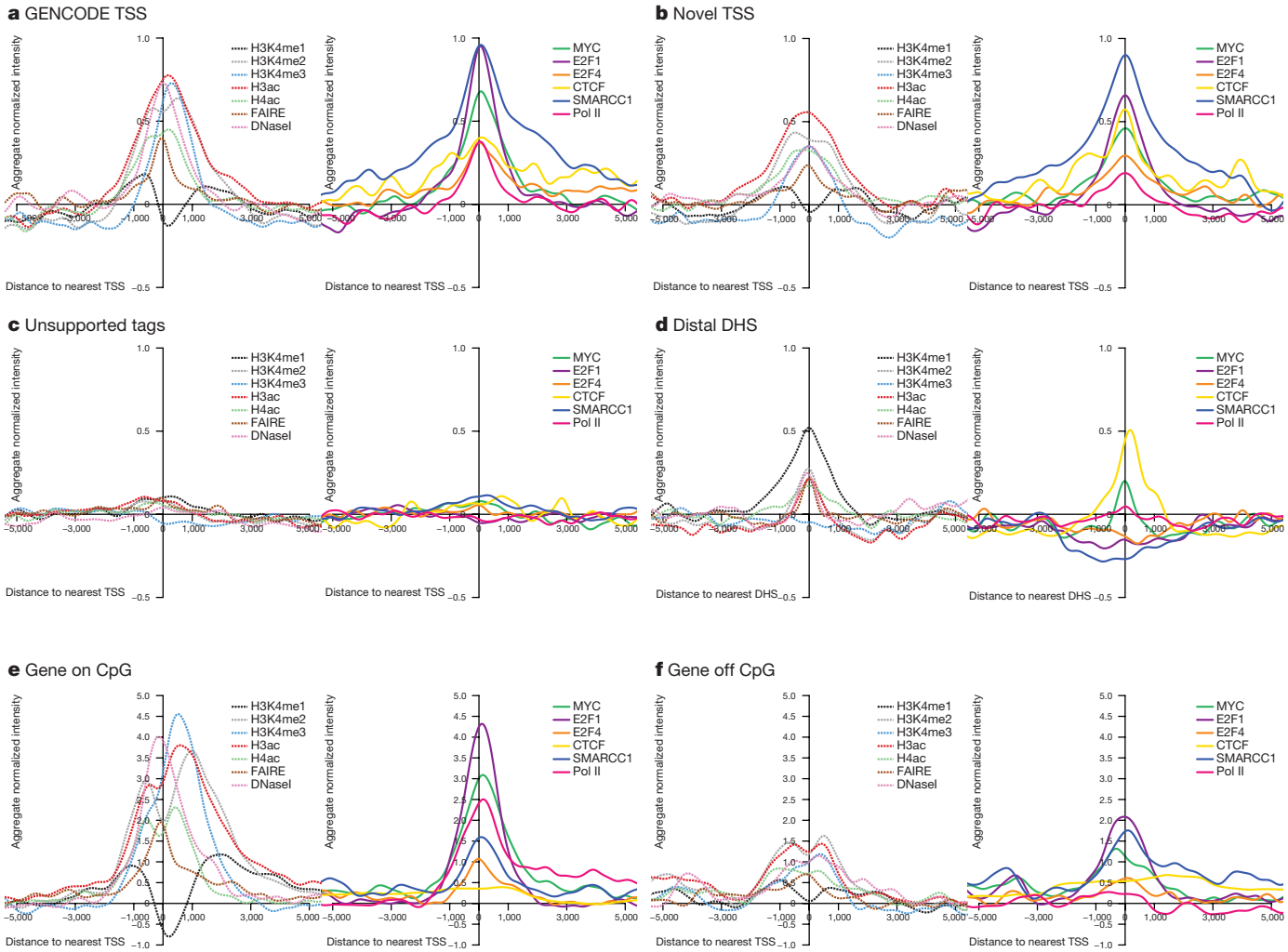


Figure 5 | Aggregate signals of tiling-array experiments from either ChIP-chip or chromatin structure assays, represented for different classes of TSSs and DHS. For each plot, the signal was first normalized with a mean of 0 and standard deviation of 1, and then the normalized scores were summed at each position for that class of TSS or DHS and smoothed using a kernel density method (see Supplementary Information section 3.6). For each class of sites there are two adjacent plots. The left plot depicts the data for general factors: FAIRE and DNaseI sensitivity as assays of chromatin accessibility and H3K4me1, H3K4me2, H3K4me3, H3ac and H4ac histone modifications (as indicated); the right plot shows the data for additional factors, namely MYC, E2F1, E2F4, CTCF, SMARCC1 and Pol II. The columns provide data for the different classes of TSS or DHS (unsmoothed data and statistical analysis shown in Supplementary Information section 3.6).

behaviour of sequence-specific factors points to distinct biological differences, mediated by transcription factors, between distal regulatory sites and TSSs.

Unbiased maps of sequence-specific regulatory factor binding. The previous section focused on specific positions defined by TSSs or DHSs. We then analysed sequence-specific transcription factor binding data in an unbiased fashion. We refer to regions with enriched binding of regulatory factors as RFBRs. RFBRs were identified on the basis of ChIP-chip data in two ways: first, each investigator developed and used their own analysis method(s) to define high-enrichment regions, and second (and independently), a stringent false discovery rate (FDR) method was applied to analyse all data using three cut-offs (1%, 5% and 10%). The laboratory-specific and FDR-based methods were highly correlated, particularly for regions with strong signals[10,11]. For consistency, we used the results obtained with the FDR-based method (see Supplementary Information section 3.10). These RFBRs can be used to find sequence motifs (see Supplementary Information section S3.11).

RFBRs are associated with the 5′ ends of transcripts. The distribution of RFBRs is non-random (see ref. 10) and correlates with the positions of TSSs. We examined the distribution of specific RFBRs relative to the known TSSs. Different transcription factors and histone modifications vary with respect to their association with TSSs (Fig. 6; see Supplementary Information section 3.12 for modelling of random expectation). Factors for which binding sites are most enriched at the 5′ ends of genes include histone modifications, TAF1 and RNA Pol II with a hypo-phosphorylated carboxy-terminal domain[51]—confirming previous expectations. Surprisingly, we found that E2F1, a sequence-specific factor that regulates the expression of many genes at the G1 to S transition[52], is also tightly associated with TSSs[52]; this association is as strong as that of TAF1, the well-known TATA box-binding protein associated factor 1 (ref. 53). These results suggest that E2F1 has a more general role in transcription than previously suspected, similar to that for MYC[54–56]. In contrast, the large-scale assays did not support the promoter binding that was found in smaller-scale studies (for example, on SIRT1 and SPI1 (PU1)).

Integration of data on sequence-specific factors. We expect that regulatory information is not dispersed independently across the genome, but rather is clustered into distinct regions[57]. We refer to regions that contain multiple regulatory elements as 'regulatory clusters'. We sought to predict the location of regulatory clusters by cross-integrating data generated using all transcription factor and histone modification assays, including results falling below an arbitrary threshold in individual experiments. Specifically, we used four complementary methods to integrate the data from 129 ChIP-chip data sets (see Supplementary Information section 3.13 and ref. 58. These four methods detect different classes of regulatory clusters and as a whole identified 1,393 clusters. Of these, 344 were identified by all four methods, with another 500 found by three methods (see Supplementary Information section 3.13.5). 67% of the 344 regulatory clusters identified by all four methods (or 65% of the full set of 1,393) reside within 2.5 kb of a known or novel TSS (as defined above; see Table 3 and Supplementary Information section 3.14 for a breakdown by category). Restricting this analysis to previously annotated TSSs (for example, RefSeq or Ensembl) reveals that roughly 25% of the regulatory clusters are close to a previously identified TSS. These results suggest that many of the regulatory clusters identified by integrating the ChIP-chip data sets are undiscovered promoters or are somehow associated with transcription in another fashion. To test these possibilities, sets of 126 and 28 non-GENCODE-based regulatory clusters were tested for promoter activity (see Supplementary Information section 3.15) and by RACE, respectively. These studies revealed that 24.6% of the 126 tested regulatory clusters had promoter activity and that 78.6% of the 28 regulatory clusters analysed by RACE yielded products consistent with a TSS[58]. The ChIP-chip data sets were generated on a mixture of cell lines, predominantly HeLa and GM06990, and were different from the CAGE/PET data, meaning that tissue specificity contributes to the presence of unique TSSs and regulatory clusters. The large increase in promoter proximal regulatory clusters identified by including the additional novel TSSs coupled with the positive promoter and RACE assays suggests that most of the regulatory regions identifiable by these clustering methods represent bona fide promoters (see Supplementary Information 3.16). Although the regulatory factor assays were more biased towards regions associated with promoters, many of the sites from these experiments would have previously been described as distal to promoters. This suggests that commonplace use of RefSeq- or Ensembl-based gene definition to define promoter proximity will dramatically overestimate the number of distal sites.

Predicting TSSs and transcriptional activity on the basis of chromatin structure. The strong association between TSSs and both histone modifications and DHSs prompted us to investigate whether the location and activity of TSSs could be predicted solely on the basis of chromatin structure information. We trained a support vector machine (SVM) by using histone modification data anchored around DHSs to discriminate between DHSs near TSSs and those distant from TSSs. We used a selected 2,573 DHSs, split roughly between TSS-proximal DHSs and TSS-distal DHSs, as a training set. The SVM performed well, with an accuracy of 83% (see Supplementary Information section 3.17). Using this SVM, we then predicted new TSSs using information about DHSs and histone modifications—of 110 high-scoring predicted TSSs, 81 resided within 2.5 kb of a novel TSS. As expected, these show a significant overlap to the novel TSS groups (defined above) but without a strong bias towards any particular category (see Supplementary Information section 3.17.1.5).

To investigate the relationship between chromatin structure and gene expression, we examined transcript levels in two cell lines using a transcript-tiling array. We compared this transcript data with the results of ChIP-chip experiments that measured histone modifications across the ENCODE regions. From this, we developed a variety of predictors of expression status using chromatin modifications as variables; these were derived using both decision trees and SVMs (see Supplementary Information section 3.17). The best of these correctly predicts expression status (transcribed versus non-transcribed) in 91% of cases. This success rate did not decrease dramatically when the predicting algorithm incorporated the results from one cell line to predict the expression status of another cell line. Interestingly, despite

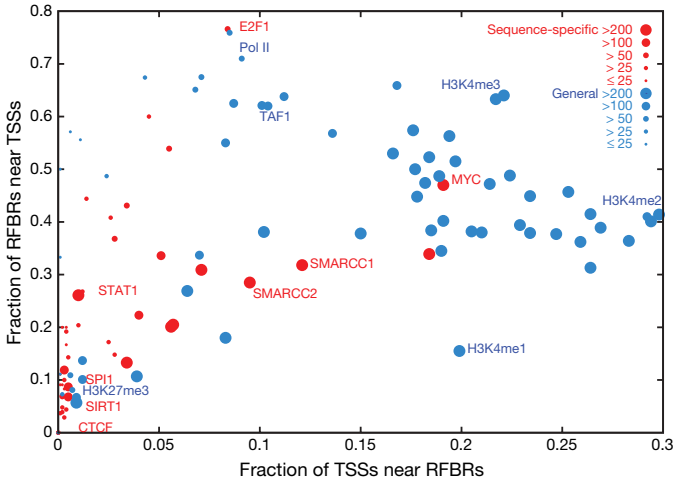


Figure 6 | Distribution of RFBRs relative to GENCODE TSSs. Different RFBRs from sequence-specific factors (red) or general factors (blue) are plotted showing their relative distribution near TSSs. The *x* axis indicates the proportion of TSSs close (within 2.5 kb) to the specified factor. The *y* axis indicates the proportion of RFBRs close to TSSs. The size of the circle provides an indication of the number of RFBRs for each factor. A handful of representative factors are labelled.

the striking difference in histone modification enrichments in TSSs residing near versus those more distal to CpG islands (see Fig. 5 and Supplementary Information section 3.6), including information about the proximity to CpG islands did not improve the predictors. This suggests that despite the marked differences in histone modifications among these TSS classes, a single predictor can be made, using the interactions between the different histone modification levels.

In summary, we have integrated many data sets to provide a more complete view of regulatory information, both around specific sites (TSSs and DHSs) and in an unbiased manner. From analysing multiple data sets, we find 4,491 known and novel TSSs in the ENCODE regions, almost tenfold more than the number of established genes. This large number of TSSs might explain the extensive transcription described above; it also begins to change our perspective about regulatory information—without such a large TSS catalogue, many of the regulatory clusters would have been classified as residing distal to promoters. In addition to this revelation about the abundance of promoter-proximal regulatory elements, we also identified a considerable number of putative distal regulatory elements, particularly on the basis of the presence of DHSs. Our study of distal regulatory elements was probably most hindered by the paucity of data generated using distal-element-associated transcription factors; nevertheless, we clearly detected a set of distal-DHS-associated segments bound by CTCF or MYC. Finally, we showed that information about chromatin structure alone could be used to make effective predictions about both the location and activity of TSSs.

Replication

Overview. DNA replication must be carefully coordinated, both across the genome and with respect to development. On a larger scale, early replication in S phase is broadly correlated with gene density and transcriptional activity[59–66]; however, this relationship is not universal, as some actively transcribed genes replicate late and vice versa[61,64–68]. Importantly, the relationship between transcription and DNA replication emerges only when the signal of transcription is averaged over a large window (>100 kb)[63], suggesting that larger-scale chromosomal architecture may be more important than the activity of specific genes[69].

The ENCODE Project provided a unique opportunity to examine whether individual histone modifications on human chromatin can be correlated with the time of replication and whether such correlations support the general relationship of active, open chromatin with early replication. Our studies also tested whether segments showing interallelic variation in the time of replication have two different types of histone modifications consistent with an interallelic variation in chromatin state.

DNA replication data set. We mapped replication timing across the ENCODE regions by analysing Brd-U-labelled fractions from synchronized HeLa cells (collected at 2 h intervals throughout S phase) on tiling arrays (see Supplementary Information section 4.1). Although the HeLa cell line has a considerably altered karyotype, correlation of these data with other cell line data (see below) suggests the results are relevant to other cell types. The results are expressed as the time at which 50% of any given genomic position is replicated (TR50), with higher values signifying later replication times. In addition to the five 'activating' histone marks, we also correlated the TR50 with H3K27me3, a modification associated with polycomb-mediated transcriptional repression[70–74]. To provide a consistent comparison framework, the histone data were smoothed to 100-kb resolution, and then correlated with the TR50 data by a sliding window correlation analysis (see Supplementary Information section 4.2). The continuous profiles of the activating marks, histone H3K4 mono-, di-, and tri-methylation and histone H3 and H4 acetylation, are generally anti-correlated with the TR50 signal (Fig. 7a and Supplementary Information section 4.3). In contrast, H3K27me3 marks show a predominantly positive correlation with late-replicating segments (Fig. 7a; see Supplementary Information section 4.3 for additional analysis).

Although most genomic regions replicate in a temporally specific window in S phase, other regions demonstrate an atypical pattern of replication (Pan-S) where replication signals are seen in multiple parts of S phase. We have suggested that such a pattern of replication stems from interallelic variation in the chromatin structure[59,75]. If one allele is in active chromatin and the other in repressed chromatin, both types of modified histones are expected to be enriched in the Pan-S segments. An ENCODE region was classified as non-specific (or Pan-S) regions when >60% of the probes in a 10-kb window

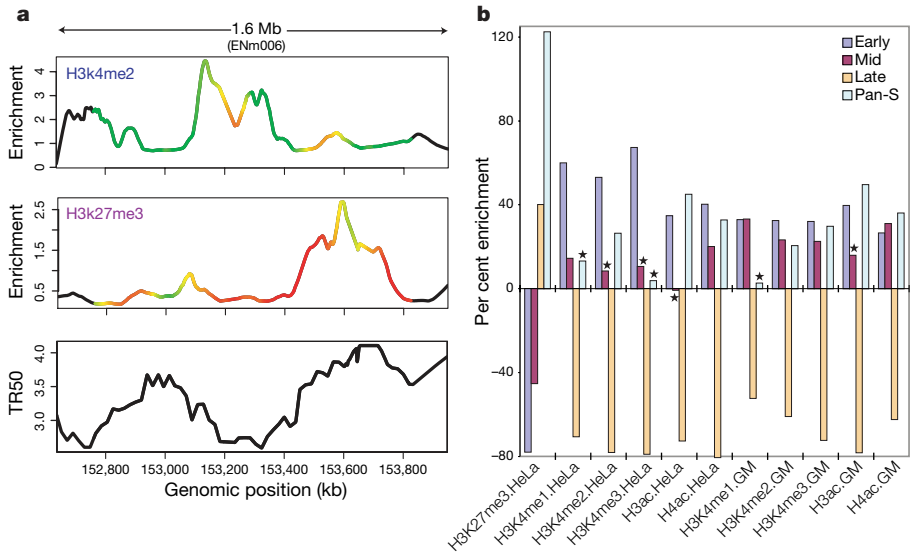


Figure 7 | Correlation between replication timing and histone modifications. a, Comparison of two histone modifications (H3K4me2 and H3K27me3), plotted as enrichment ratio from the Chip-chip experiments and the time for 50% of the DNA to replicate (TR50), indicated for ENCODE region ENm006. The colours on the curves reflect the correlation strength in a sliding 250-kb window. **b**, Differing levels of histone modification for different TR50 partitions. The amounts of enrichment or depletion of different histone modifications in various cell lines are depicted (indicated along the bottom as 'histone mark.cell line'; GM = GM06990). Asterisks indicate enrichments/depletions that are not significant on the basis of multiple tests. Each set has four partitions on the basis of replication timing: early, mid, late and Pan-S.

replicated in multiple intervals in S phase. The remaining regions were sub-classified into early-, mid- or late-replicating based on the average TR50 of the temporally specific probes within a 10-kb window[75]. For regions of each class of replication timing, we determined the relative enrichment of various histone modification peaks in HeLa cells (Fig. 7b; Supplementary Information section 4.4). The correlations of activating and repressing histone modification peaks with TR50 are confirmed by this analysis (Fig. 7b). Intriguingly, the Pan-S segments are unique in being enriched for both activating (H3K4me2, H3ac and H4ac) and repressing (H3K27me3) histones, consistent with the suggestion that the Pan-S replication pattern arises from interallelic variation in chromatin structure and time of replication[75]. This observation is also consistent with the Pan-S replication pattern seen for the H19/IGF2 locus, a known imprinted region with differential epigenetic modifications across the two alleles[76].

The extensive rearrangements in the genome of HeLa cells led us to ask whether the detected correlations between TR50 and chromatin state are seen with other cell lines. The histone modification data with GM06990 cells allowed us to test whether the time of replication of genomic segments in HeLa cells correlated with the chromatin state in GM06990 cells. Early- and late-replicating segments in HeLa cells are enriched and depleted, respectively, for activating marks in GM06990 cells (Fig. 7b). Thus, despite the presence of genomic rearrangements (see Supplementary Information section 2.12), the TR50 and chromatin state in HeLa cells are not far from a constitutive baseline also seen with a cell line from a different lineage. The enrichment of multiple activating histone modifications and the depletion of a repressive modification from segments that replicate early in S phase extends previous work in the field at a level of detail and scale not attempted before in mammalian cells. The duality of histone modification patterns in Pan-S areas of the HeLa genome, and the concordance of chromatin marks and replication time across two disparate cell lines (HeLa and GM06990) confirm the coordination of histone modifications with replication in the human genome.

Chromatin architecture and genomic domains

Overview. The packaging of genomic DNA into chromatin is intimately connected with the control of gene expression and other chromosomal processes. We next examined chromatin structure over a larger scale to ascertain its relation to transcription and other processes. Large domains (50 to >200 kb) of generalized DNaseI sensitivity have been detected around developmentally regulated gene clusters[77], prompting speculation that the genome is organized into 'open' and 'closed' chromatin territories that represent higher-order functional domains. We explored how different chromatin features, particularly histone modifications, correlate with chromatin structure, both over short and long distances.

Chromatin accessibility and histone modifications. We used histone modification studies and DNaseI sensitivity data sets (introduced above) to examine general chromatin accessibility without focusing on the specific DHS sites (see Supplementary Information sections 3.1, 3.3 and 3.4). A fundamental difficulty in analysing continuous data across large genomic regions is determining the appropriate scale for analysis (for example, 2 kb, 5 kb, 20 kb, and so on). To address this problem, we developed an approach based on wavelet analysis, a mathematical tool pioneered in the field of signal processing that has recently been applied to continuous-value genomic analyses. Wavelet analysis provides a means for consistently transforming continuous signals into different scales, enabling the correlation of different phenomena independently at differing scales in a consistent manner.

Global correlations of chromatin accessibility and histone modifications. We computed the regional correlation between DNaseI sensitivity and each histone modification at multiple scales using a wavelet approach (Fig. 8 and Supplementary Information section 4.2). To make quantitative comparisons between different histone modifications, we computed histograms of correlation values between DNaseI sensitivity and each histone modification at several scales and then tested these for significance at specific scales. Figure 8c shows the distribution of correlation values at a 16-kb scale, which is considerably larger than individual *cis*-acting regulatory elements. At this scale, H3K4me2, H3K4me3 and H3ac show similarly high correlation. However, they are significantly distinguished from H3K4me1 and H4ac modifications ($P < 1.5 \times 10^{-33}$; see Supplementary Information section 4.5), which show lower correlation with DNaseI sensitivity. These results suggest that larger-scale relationships between chromatin accessibility and histone modifications are dominated by sub-regions in which higher average DNaseI sensitivity is accompanied by high levels of H3K4me2, H3K4me3 and H3ac modifications.

Local correlations of chromatin accessibility and histone modifications. Narrowing to a scale of ~2 kb revealed a more complex situation, in which H3K4me2 is the histone modification that is best correlated with DNaseI sensitivity. However, there is no clear combination of marks that correlate with DNaseI sensitivity in a way that is analogous to that seen at a larger scale (see Supplementary Information section 4.3). One explanation for the increased

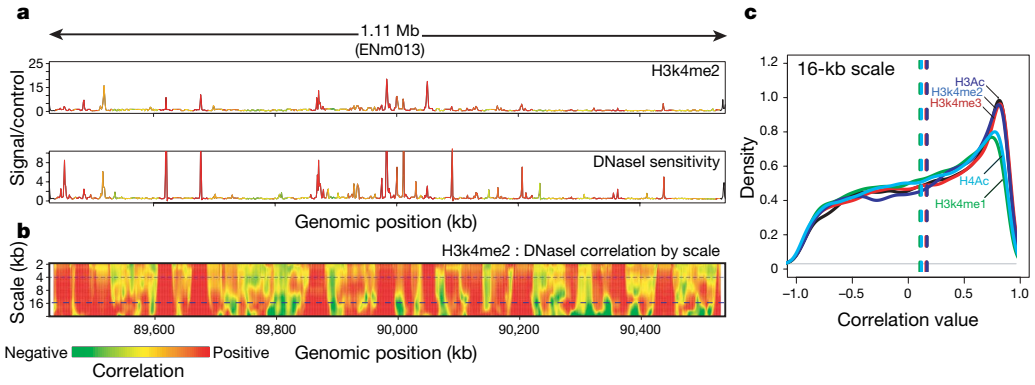


Figure 8 | Wavelet correlations of histone marks and DNaseI sensitivity. As an example, correlations between DNaseI sensitivity and H3K4me2 (both in the GM06990 cell line) over a 1.1-Mb region on chromosome 7 (ENCODE region ENm013) are shown. **a**, The relationship between histone modification H3K4me2 (upper plot) and DNaseI sensitivity (lower plot) is shown for ENCODE region ENm013. The curves are coloured with the strength of the local correlation at the 4-kb scale (top dashed line in panel **b**). **b**, The same data as in **a** are represented as a wavelet correlation. The *y* axis shows the differing scales decomposed by the wavelet analysis from large to small scale (in kb); the colour at each point in the heatmap represents the level of correlation at the given scale, measured in a 20 kb window centred at the given position. **c**, Distribution of correlation values at the 16 kb scale between the indicated histone marks. The *y* axis is the density of these correlation values across ENCODE; all modifications show a peak at a positive-correlation value.

complexity at smaller scales is that there is a mixture of different classes of accessible chromatin regions, each having a different pattern of histone modifications. To examine this, we computed the degree to which local peaks in histone methylation or acetylation occur at DHSs (see Supplementary Information section 4.5.1). We found that 84%, 91% and 93% of significant peaks in H3K4 mono-, di- and tri-methylation, respectively, and 93% and 81% of significant peaks in H3ac and H4ac acetylation, respectively, coincided with DHSs (see Supplementary Information section 4.5). Conversely, a proportion of DHSs seemed not to be associated with significant peaks in H3K4 mono-, di- or tri-methylation (37%, 29% and 47%, respectively), nor with peaks in H3 or H4 acetylation (both 57%). Because only a limited number of histone modification marks were assayed, the possibility remains that some DHSs harbour other histone modifications. The absence of a more complete concordance between DHSs and peaks in histone acetylation is surprising given the widely accepted notion that histone acetylation has a central role in mediating chromatin accessibility by disrupting higher-order chromatin folding.

DNA structure at DHSs. The observation that distinctive hydroxyl radical cleavage patterns are associated with specific DNA structures[78] prompted us to investigate whether DHS subclasses differed with respect to their local DNA structure. Conversely, because different DNA sequences can give rise to similar hydroxyl radical cleavage patterns[79], genomic regions that adopt a particular local structure do not necessarily have the same nucleotide sequence. Using a Gibbs sampling algorithm on hydroxyl radical cleavage patterns of 3,150 DHSs[80], we discovered an 8-base segment with a conserved cleavage signature (CORCS; see Supplementary Information section 4.6). The underlying DNA sequences that give rise to this pattern have little primary sequence similarity despite this similar structural pattern. Furthermore, this structural element is strongly enriched in promoter-proximal DHSs (11.3-fold enrichment compared to the rest of the ENCODE regions) relative to promoter-distal DHSs (1.5-fold enrichment); this element is enriched 10.9-fold in CpG islands, but is higher still (26.4-fold) in CpG islands that overlap a DHS.

Large-scale domains in the ENCODE regions. The presence of extensive correlations seen between histone modifications, DNaseI sensitivity, replication, transcript density and protein factor binding led us to investigate whether all these features are organized systematically across the genome. To test this, we performed an unsupervised training of a two-state HMM with inputs from these different features (see Supplementary Information section 4.7 and ref. 81). No other information except for the experimental variables was used for the HMM training routines. We consistently found that one state ('active') generally corresponded to domains with high levels of H3ac and RNA transcription, low levels of H3K27me3 marks, and early replication timing, whereas the other state ('repressed') reflected domains with low H3ac and RNA, high H3K27me3, and late replication (see Fig. 9). In total, we identified 70 active regions spanning 11.4 Mb and 82 inactive regions spanning 17.8 Mb (median size 136 kb versus 104 kb respectively). The active domains are markedly enriched for GENCODE TSSs, CpG islands and Alu repetitive elements ($P < 0.0001$ for each), whereas repressed regions are significantly enriched for LINE1 and LTR transposons ($P < 0.001$). Taken together, these results demonstrate remarkable concordance between ENCODE functional data types and provide a view of higher-order functional domains defined by a broader range of factors at a markedly higher resolution than was previously available[82].

Evolutionary constraint and population variability

Overview. Functional genomic sequences can also be identified by examining evolutionary changes across multiple extant species and within the human population. Indeed, such studies complement experimental assays that identify specific functional elements[83–85]. Evolutionary constraint (that is, the rejection of mutations at a particular location) can be measured by either (i) comparing observed substitutions to neutral rates calculated from multi-sequence alignments[86–88], or (ii) determining the presence and frequency of intra-species polymorphisms. Importantly, both approaches are indifferent to any specific function that the constrained sequence might confer.

Previous studies comparing the human, mouse, rat and dog genomes examined bulk evolutionary properties of all nucleotides in the genome, and provided little insight about the precise positions of constrained bases. Interestingly, these studies indicated that the

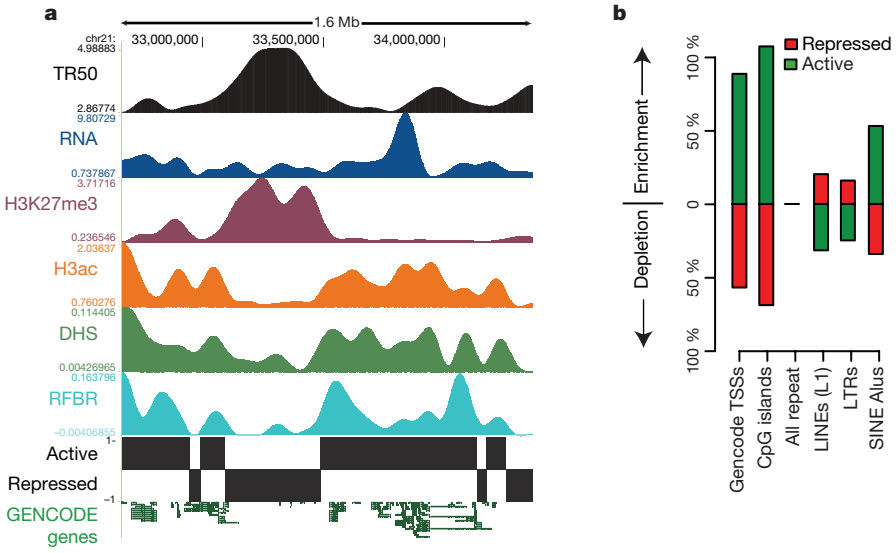


Figure 9 | Higher-order functional domains in the genome. The general concordance of multiple data types is shown for an illustrative ENCODE region (ENm005). **a**, Domains were determined by simultaneous HMM segmentation of replication time (TR50; black), bulk RNA transcription (blue), H3K27me3 (purple), H3ac (orange), DHS density (green), and RFBR density (light blue) measured continuously across the 1.6-Mb ENm005. All data were generated using HeLa cells. The histone, RNA, DHS and RFBR signals are wavelet-smoothed to an approximately 60-kb scale (see Supplementary Information section 4.7). The HMM segmentation is shown as the blocks labelled 'active' and 'repressed' and the structure of GENCODE genes (not used in the training) is shown at the end. **b**, Enrichment or depletion of annotated sequence features (GENCODE TSSs, CpG islands, LINE1 repeats, Alu repeats, and non-exonic constrained sequences (CSs)) in active versus repressed domains. Note the marked enrichment of TSSs, CpG islands and Alus in active domains, and the enrichment of LINE and LTRs in repressed domains.

majority of constrained bases reside within the non-coding portion of the human genome. Meanwhile, increasingly rich data sets of polymorphisms across the human genome have been used extensively to establish connections between genetic variants and disease, but far fewer analyses have sought to use such data for assessing functional constraint[85].

The ENCODE Project provides an excellent opportunity for more fully exploiting inter- and intra-species sequence comparisons to examine genome function in the context of extensive experimental studies on the same regions of the genome. We consolidated the experimentally derived information about the ENCODE regions and focused our analyses on 11 major classes of genomic elements. These classes are listed in Table 4 and include two non-experimentally derived data sets: ancient repeats (ARs; mobile elements that inserted early in the mammalian lineage, have subsequently become dormant, and are assumed to be neutrally evolving) and constrained sequences (CSs; regions that evolve detectably more slowly than neutral sequences).

Comparative sequence data sets and analysis. We generated 206 Mb of genomic sequence orthologous to the ENCODE regions from 14 mammalian species using a targeted strategy that involved isolating[89] and sequencing[90] individual bacterial artificial chromosome clones. For an additional 14 vertebrate species, we used 340 Mb of orthologous genomic sequence derived from genome-wide sequencing efforts[3–8,91–93]. The orthologous sequences were aligned using three alignment programs: TBA[94], MAVID[95] and MLAGAN[96]. Four independent methods that generated highly concordant results[97] were then used to identify sequences under constraint (PhastCons[88], GERP[87], SCONE[98] and BinCons[86]). From these analyses, we developed a high-confidence set of 'constrained sequences' that correspond to 4.9% of the nucleotides in the ENCODE regions. The threshold for determining constraint was set using a FDR rate of 5% (see ref. 97); this level is similar to previous estimates of the fraction of the human genome under mammalian constraint[4,86–88] but the FDR rate was not chosen to fit this result. The median length of these constrained sequences is 19 bases, with the minimum being 8 bases—roughly the size of a typical transcription factor binding site. These analyses, therefore, provide a resolution of constrained sequences that is substantially better than that currently available using only whole-genome vertebrate sequences[99–102].

Intra-species variation studies mainly used SNP data from Phases I and II, and the 10 re-sequenced regions in ENCODE regions with 48 individuals of the HapMap Project[103]; nucleotide insertion or deletion (indel) data were from the SNP Consortium and HapMap.We also examined the ENCODE regions for the presence of overlaps with known segmental duplications[104] and CNVs.

Experimentally identified functional elements and constrained sequences. We first compared the detected constrained sequences with the positions of experimentally identified functional elements. A total of 40% of the constrained bases reside within protein-coding exons and their associated untranslated regions (Fig. 10) and, in agreement with previous genome-wide estimates, the remaining constrained bases do not overlap the mature transcripts of protein-coding genes[4,5,88,105,106]. When we included the other experimental annotations, we found that an additional 20% of the constrained bases overlap experimentally identified non-coding functional regions, although far fewer of these regions overlap constrained sequences compared to coding exons (see below). Most experimental annotations are significantly different from a random expectation for both base-pair or element-level overlaps (using the GSC statistic, see Supplementary Information section 1.3), with a more striking deviation when considering elements (Fig. 11). The exceptions to this are pseudogenes, Un.TxFrags and RxFrags. The increase in significance moving from base-pair measures to the element level suggests that discrete islands of constrained sequence exist within experimentally identified functional elements, with the surrounding bases apparently not showing evolutionary constraint. This notion is discussed in greater detail in ref. 97.

We also examined measures of human variation (heterozygosity, derived allele-frequency spectra and indel rates) within the sequences of the experimentally identified functional elements (Fig. 12). For these studies, ARs were used as a marker for neutrally evolving sequence. Most experimentally identified functional elements are associated with lower heterozygosity compared to ARs, and a few have lower indel rates compared with ARs. Striking outliers are 3′ UTRs, which have dramatically increased indel rates without an obvious cause. This is discussed in more depth in ref. 107.

These findings indicate that the majority of the evolutionarily constrained, experimentally identified functional elements show evidence of negative selection both across mammalian species and within the human population. Furthermore, we have assigned at least one molecular function to the majority (60%) of all constrained bases in the ENCODE regions.

Conservation of regulatory elements. The relationship between individual classes of regulatory elements and constrained sequences varies considerably, ranging from cases where there is strong evolutionary constraint (for example, pan-vertebrate ultraconserved regions[108,109]) to examples of regulatory elements that are not conserved between orthologous human and mouse genes[110]. Within the ENCODE regions, 55% of RFBRs overlap the high-confidence

Table 4 | Eleven classes of genomic elements subjected to evolutionary and population-genetics analyses

Abbreviation	Description
CDS	Coding exons, as annotated by GENCODE
5′ UTR	5′ untranslated region, as annotated by GENCODE
3′ UTR	3′ untranslated region, as annotated by GENCODE
Un.TxFrag	Unannotated region detected by RNA hybridization to tiling array (that is, unannotated TxFrag)
RxFrag	Region detected by RACE and analysis on tiling array
Pseudogene	Pseudogene identified by consensus pseudogene analysis
RFBR	Regulatory factor binding region identified by ChIP-chip assay
RFBR-SeqSp	Regulatory factor binding region identified only by ChIP-chip assays for factors with known sequence-specificity
DHS	DNaseI hypersensitive sites found in multiple tissues
FAIRE	Region of open chromatin identified by the FAIRE assay
TSS	Transcription start site
AR	Ancient repeat inserted early in the mammalian lineage and presumed to be neutrally evolving
CS	Constrained sequence identified by analysing multi-sequence alignments

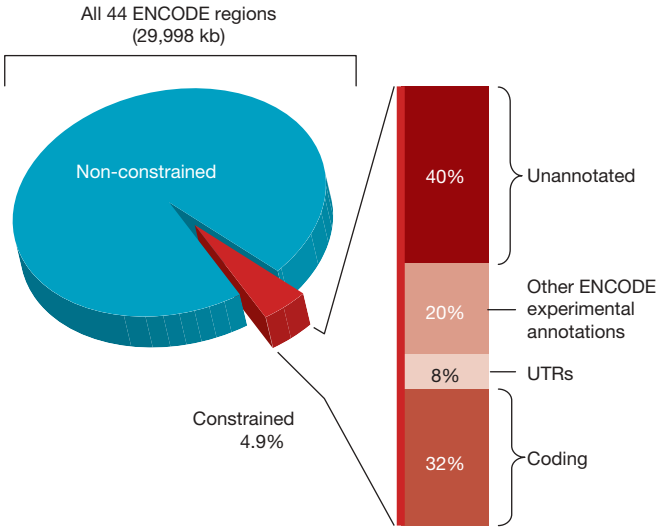


Figure 10 | Relative proportion of different annotations among constrained sequences. The 4.9% of bases in the ENCODE regions identified as constrained is subdivided into the portions that reflect known coding regions, UTRs, other experimentally annotated regions, and unannotated sequence.

constrained sequences. As expected, RFBRs have many unconstrained bases, presumably owing to the small size of the specific binding site. We investigated whether the binding sites in RFBRs could be further delimited using information about evolutionary constraint. For 7 out of 17 factors with either known TRANSFAC or Jaspar motifs, our ChIP-chip data revealed a marked enrichment of the appropriate motif within the constrained versus the unconstrained portions of the RFBRs (see Supplementary Information section 5.1). This enrichment was seen for levels of stringency used for defining ChIP-chip-positive sites (1% and 5% FDR level), indicating that combining sequence constraint and ChIP-chip data may provide a highly sensitive means for detecting factor binding sites in the human genome.

Experimentally identified functional elements and genetic variation. The above studies focus on purifying (negative) selection. We used nucleotide variation to detect potential signals of adaptive (positive) selection. We modified the standard McDonald–Kreitman test (MK-test[111,112]) and the Hudson–Kreitman–Aguade (HKA)[113] test (see Supplementary Information section 5.2.1), to examine whether an entire set of sequence elements shows an excess of polymorphisms or an excess of inter-species divergence. We found that constrained sequences and coding exons have an excess of polymorphisms (consistent with purifying selection), whereas 5′ UTRs show evidence of an excess of divergence (with a portion probably reflecting positive selection). In general, non-coding genomic regions show more variation, with both a large number of segments that undergo purifying selection and regions that are fast evolving.

We also examined structural variation (that is, CNVs, inversions and translocations[114]; see Supplementary Information section 5.2.2). Within these polymorphic regions, we encountered significant over-representation of CDSs, TxFrags, and intra-species constrained sequences ($P < 10^{-3}$, Fig. 13), and also detected a statistically significant under-representation of ARs ($P = 10^{-3}$). A similar overrepresentation of CDSs and intra-species constrained sequences was found within non-polymorphic segmental duplications.

Unexplained constrained sequences. Despite the wealth of complementary data, 40% of the ENCODE-region sequences identified as constrained are not associated with any experimental evidence of function. There is no evidence indicating that mutational cold spots account for this constraint; they have similar measures of constraint to experimentally identified elements and harbour equal proportions of SNPs. To characterize further the unexplained constrained sequences, we examined their clustering and phylogenetic distribution. These sequences are not uniformly distributed across most ENCODE regions, and even in most ENCODE regions the distribution is different from constrained sequences within experimentally identified functional elements (see Supplementary Information section 5.3). The large fraction of constrained sequence that does not match any experimentally identified elements is not surprising considering that only a limited set of transcription factors, cell lines and biological conditions have thus far been examined.

Unconstrained experimentally identified functional elements. In contrast, an unexpectedly large fraction of experimentally identified functional elements show no evidence of evolutionary constraint ranging from 93% for Un.TxFrags to 12% for CDS. For most types of non-coding functional elements, roughly 50% of the individual elements seemed to be unconstrained across all mammals.

There are two methodological reasons that might explain the apparent excess of unconstrained experimentally identified functional elements: the underestimation of sequence constraint or overestimation of experimentally identified functional elements. We do not believe that either of these explanations fully accounts for the large and varied levels of unconstrained experimentally functional sequences. The set of constrained bases analysed here is highly accurate and complete due to the depth of the multiple alignment. Both by bulk fitting procedures and by comparison of SNP frequencies to constraint there is clearly a proportion of constrained bases not captured in the defined 4.9% of constrained sequences, but it is small (see

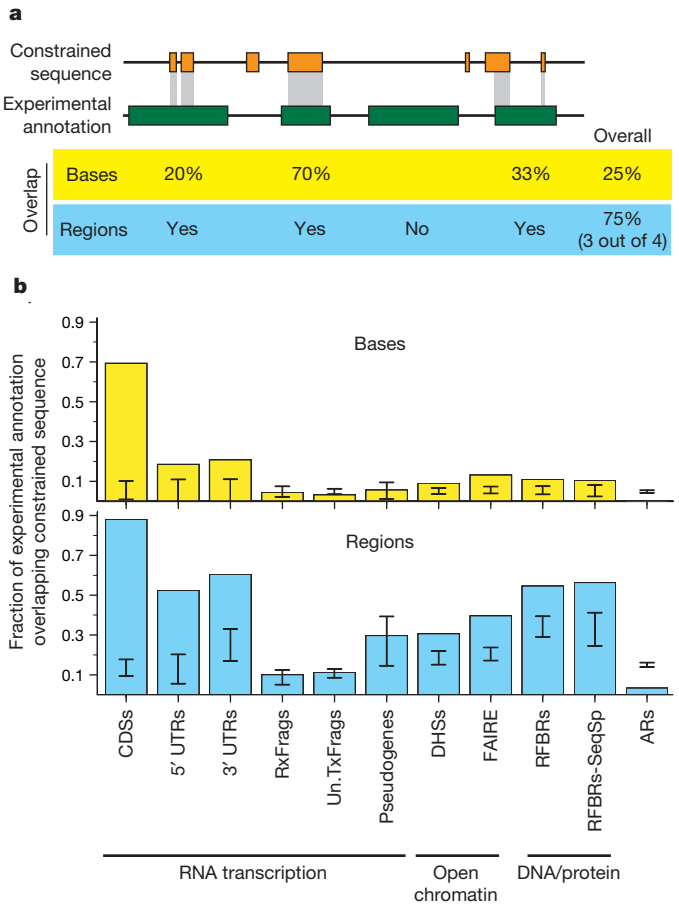


Figure 11 | Overlap of constrained sequences and various experimental annotations. **a**, A schematic depiction shows the different tests used for assessing overlap between experimental annotations and constrained sequences, both for individual bases and for entire regions. **b**, Observed fraction of overlap, depicted separately for bases and regions. The results are shown for selected experimental annotations. The internal bars indicate 95% confidence intervals of randomized placement of experimental elements using the GSC methodology to account for heterogeneity in the data sets. When the bar overlaps the observed value one cannot reject the hypothesis that these overlaps are consistent with random placements.

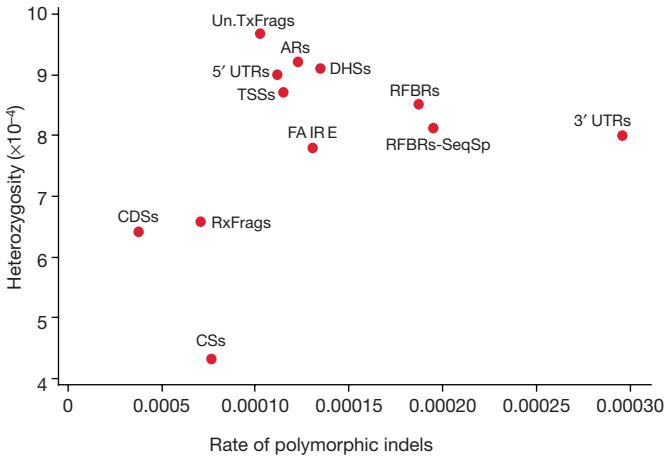


Figure 12 | Relationship between heterozygosity and polymorphic indel rate for a variety of experimental annotations. 3′ UTRs are an expected outlier for the indel measures owing to the presence of low-complexity sequence (leading to a higher indel rate).

Supplementary Information section 5.4 and S5.5). More aggressive schemes to detect constraint only marginally increase the overlap with experimentally identified functional elements, and do so with considerably less specificity. Similarly, all experimental findings have been independently validated and, for the least constrained experimentally identified functional elements (Un.TxFrags and binding sites of sequence-specific factors), there is both internal validation and cross-validation from different experimental techniques. This suggests that there is probably not a significant overestimation of experimentally identified functional elements. Thus, these two explanations may contribute to the general observation about unconstrained functional elements, but cannot fully explain it.

Instead, we hypothesize five biological reasons to account for the presence of large amounts of unconstrained functional elements. The first two are particular to certain biological assays in which the elements being measured are connected to but do not coincide with the analysed region. An example of this is the parent transcript of an miRNA, where the current assays detect the exons (some of which are not under evolutionary selection), whereas the intronic miRNA actually harbours the constrained bases. Nevertheless, the transcript sequence provides the critical coupling between the regulated promoter and the miRNA. The sliding of transcription factors (which might bind a specific sequence but then migrate along the DNA) or the processivity of histone modifications across chromatin are more exotic examples of this. A related, second hypothesis is that delocalized behaviours of the genome, such as general chromatin accessibility, may be maintained by some biochemical processes (such as transcription of intergenic regions or specific factor binding) without the requirement for specific sequence elements. These two explanations of both connected components and diffuse components related to, but not coincident with, constrained sequences are particularly relevant for the considerable amount of unannotated and unconstrained transcripts.

The other three hypotheses may be more general—the presence of neutral (or near neutral) biochemical elements, of lineage-specific functional elements, and of functionally conserved but non-orthologous elements. We believe there is a considerable proportion of neutral biochemically active elements that do not confer a selective advantage or disadvantage to the organism. This neutral pool of sequence elements may turn over during evolutionary time, emerging via certain mutations and disappearing by others. The size of the neutral pool would largely be determined by the rate of emergence and extinction through chance events; low information-content elements, such as transcription factor-binding sites[110] will have larger neutral pools. Second, from this neutral pool, some elements might occasionally acquire a biological role and so come under evolutionary selection. The acquisition of a new biological role would then create a lineage-specific element. Finally, a neutral element from the general pool could also become a peer of an existing selected functional element and either of the two elements could then be removed by chance. If the older element is removed, the newer element has, in essence, been conserved without using orthologous bases, providing a conserved function in the absence of constrained sequences. For example, a common HNF4A binding site in the human and mouse genomes may not reflect orthologous human and mouse bases, though the presence of an HNF4A site in that region was evolutionarily selected for in both lineages. Note that both the neutral turnover of elements and the 'functional peering' of elements has been suggested for *cis*-acting regulatory elements in *Drosophila*[115,116] and mammals[110]. Our data support these hypotheses, and we have generalized this idea over many different functional elements. The presence of conserved function encoded by conserved orthologous bases is a commonplace assumption in comparative genomics; our findings indicate that there could be a sizable set of functionally conserved but non-orthologous elements in the human genome, and that these seem unconstrained across mammals. Functional data akin to the ENCODE Project on other related species, such as mouse, would be critical to understanding the rate of such functionally conserved but non-orthologous elements.

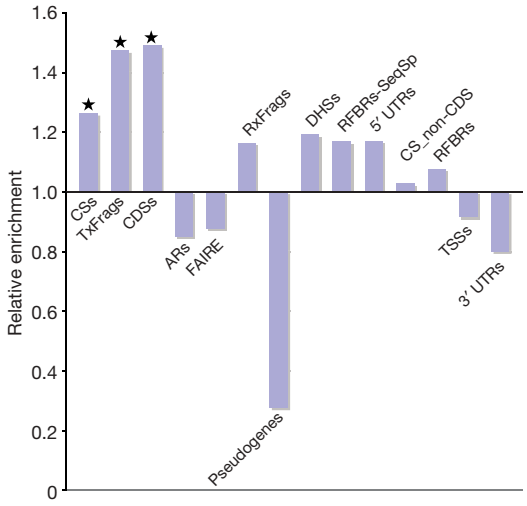


Figure 13 | CNV enrichment. The relative enrichment of different experimental annotations in the ENCODE regions associated with CNVs. CS_non-CDS are constrained sequences outside of coding regions. A value of 1 or less indicates no enrichment, and values greater than 1 show enrichment. Starred columns are cases that are significant on the basis of this enrichment being found in less than 5% of randomizations that matched each element class for length and density of features.

Conclusion

The generation and analyses of over 200 experimental data sets from studies examining the 44 ENCODE regions provide a rich source of functional information for 30 Mb of the human genome. The first conclusion of these efforts is that these data are remarkably informative. Although there will be ongoing work to enhance existing assays, invent new techniques and develop new data-analysis methods, the generation of genome-wide experimental data sets akin to the ENCODE pilot phase would provide an impressive platform for future genome exploration efforts. This now seems feasible in light of throughput improvements of many of the assays and the ever-declining costs of whole-genome tiling arrays and DNA sequencing. Such genome-wide functional data should be acquired and released openly, as has been done with other large-scale genome projects, to ensure its availability as a new foundation for all biologists studying the human genome. It is these biologists who will often provide the critical link from biochemical function to biological role for the identified elements.

The scale of the pilot phase of the ENCODE Project was also sufficiently large and unbiased to reveal important principles about the organization of functional elements in the human genome. In many cases, these principles agree with current mechanistic models. For example, we observe trimethylation of H3K4 enriched near active genes, and have improved the ability to accurately predict gene activity based on this and other histone modifications. However, we also uncovered some surprises that challenge the current dogma on biological mechanisms. The generation of numerous intercalated transcripts spanning the majority of the genome has been repeatedly suggested[13,14], but this phenomenon has been met with mixed opinions about the biological importance of these transcripts. Our analyses of numerous orthogonal data sets firmly establish the presence of these transcripts, and thus the simple view of the genome as having a defined set of isolated loci transcribed independently does not seem to be accurate. Perhaps the genome encodes a network of transcripts, many of which are linked to protein-coding transcripts and to the majority of which we cannot (yet) assign a biological role. Our perspective of transcription and genes may have to evolve and also poses

some interesting mechanistic questions. For example, how are splicing signals coordinated and used when there are so many overlapping primary transcripts? Similarly, to what extent does this reflect neutral turnover of reproducible transcripts with no biological role?

We gained subtler but equally important mechanistic findings relating to transcription, replication and chromatin modification. Transcription factors previously thought to primarily bind promoters bind more generally, and those which do bind to promoters are equally likely to bind downstream of a TSS as upstream. Interestingly, many elements that previously were classified as distal enhancers are, in fact, close to one of the newly identified TSSs; only about 35% of sites showing evidence of binding by multiple transcription factors are actually distal to a TSS. This need not imply that most regulatory information is confined to classic promoters, but rather it does suggest that transcription and regulation are coordinated actions beyond just the traditional promoter sequences. Meanwhile, although distal regulatory elements could be identified in the ENCODE regions, they are currently difficult to classify, in part owing to the lack of a broad set of transcription factors to use in analysing such elements. Finally, we now have a much better appreciation of how DNA replication is coordinated with histone modifications.

At the outset of the ENCODE Project, many believed that the broad collection of experimental data would nicely dovetail with the detailed evolutionary information derived from comparing multiple mammalian sequences to provide a neat 'dictionary' of conserved genomic elements, each with a growing annotation about their biochemical function(s). In one sense, this was achieved; the majority of constrained bases in the ENCODE regions are now associated with at least some experimentally derived information about function. However, we have also encountered a remarkable excess of experimentally identified functional elements lacking evolutionary constraint, and these cannot be dismissed for technical reasons. This is perhaps the biggest surprise of the pilot phase of the ENCODE Project, and suggests that we take a more 'neutral' view of many of the functions conferred by the genome.

METHODS

The methods are described in the Supplementary Information, with more technical details for each experiment often found in the references provided in Table 1. The Supplementary Information sections are arranged in the same order as the manuscript (with similar headings to facilitate cross-referencing). The first page of Supplementary Information also has an index to aid navigation. Raw data are available in ArrayExpress, GEO or EMBL/GenBank archives as appropriate, as detailed in Supplementary Information section 1.1. Processed data are also presented in a user-friendly manner at the UCSC Genome Browser's ENCODE portal (http://genome.ucsc.edu/ENCODE/).

Received 2 March; accepted 23 April 2007.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
3. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
4. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
5. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
6. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).
7. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).
8. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
9. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
10. Zhang, Z. D. *et al.* Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res.* **17,** 787–797 (2007).
11. Euskirchen, G. M. *et al.* Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array and sequencing based technologies. *Genome Res.* **17,** 898–909 (2007).
12. Willingham, A. T. & Gingeras, T. R. TUF love for "junk" DNA. *Cell* **125,** 1215–1220 (2006).
13. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38,** 626–635 (2006).
14. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308,** 1149–1154 (2005).
15. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306,** 2242–2246 (2004).
16. Guigó, R. *et al.* EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.* **7,** (Suppl. 1; S2) 1–31 (2006).
17. Denoeud, F. *et al.* Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* **17,** 746–759 (2007).
18. Tress, M. L. *et al.* The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA* **104,** 5495–5500 (2007).
19. Rozowsky, J. *et al.* The DART classification of unannotated transcription within ENCODE regions: Associating transcription with known and novel loci. *Genome Res.* **17,** 732–745 (2007).
20. Kapranov, P. *et al.* Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15,** 987–997 (2005).
21. Balakirev, E. S. & Ayala, F. J. Pseudogenes: are they "junk" or functional DNA? *Annu. Rev. Genet.* **37,** 123–151 (2003).
22. Mighell, A. J., Smith, N. R., Robinson, P. A. & Markham, A. F. Vertebrate pseudogenes. *FEBS Lett.* **468,** 109–114 (2000).
23. Zheng, D. *et al.* Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription and evolution. *Genome Res.* **17,** 839–851 (2007).
24. Zheng, D. *et al.* Integrated pseudogene annotation for human chromosome 22: evidence for transcription. *J. Mol. Biol.* **349,** 27–45 (2005).
25. Harrison, P. M., Zheng, D., Zhang, Z., Carriero, N. & Gerstein, M. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res.* **33,** 2374–2383 (2005).
26. Washietl, S. *et al.* Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.* **17,** 852–864 (2007).
27. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–1563 (2005).
28. Runte, M. *et al.* The IC-*SNURF–SNRPN* transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for *UBE3A*. *Hum. Mol. Genet.* **10,** 2687–2700 (2001).
29. Seidl, C. I., Stricker, S. H. & Barlow, D. P. The imprinted *Air* ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. *EMBO J.* **25,** 3565–3575 (2006).
30. Parra, G. *et al.* Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.* **16,** 37–44 (2006).
31. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7,** 29–59 (2006).
32. Trinklein, N. D., Aldred, S. J., Saldanha, A. J. & Myers, R. M. Identification and functional analysis of human transcriptional promoters. *Genome Res.* **13,** 308–312 (2003).
33. Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L. & Myers, R. M. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* **16,** 1–10 (2006).
34. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116,** 499–509 (2004).
35. Yelin, R. *et al.* Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnol.* **21,** 379–386 (2003).
36. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309,** 1564–1566 (2005).
37. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290,** 2306–2309 (2000).
38. Iyer, V. R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409,** 533–538 (2001).
39. Horak, C. E. *et al.* GATA-1 binding sites mapped in the β-globin locus by using mammalian CHIp-chip analysis. *Proc. Natl Acad. Sci. USA* **99,** 2924–2929 (2002).
40. Wei, C. L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124,** 207–219 (2006).
41. Kim, J., Bhinge, A. A., Morgan, X. C. & Iyer, V. R. Mapping DNA–protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nature Methods* **2,** 47–53 (2005).
42. Dorschner, M. O. *et al.* High-throughput localization of functional elements by quantitative chromatin profiling. *Nature Methods* **1,** 219–225 (2004).
43. Sabo, P. J. *et al.* Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. *Nature Methods* **3,** 511–518 (2006).
44. Crawford, G. E. *et al.* DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature Methods* **3,** 503–509 (2006).
45. Hogan, G. J., Lee, C. K. & Lieb, J. D. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS Genet.* **2,** e158 (2006).

46. Koch, C. M. *et al.* The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.* **17,** 691–707 (2007).
47. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72,** 449–479 (2003).
48. Mito, Y., Henikoff, J. G. & Henikoff, S. Genome-scale profiling of histone H3.3 replacement patterns. *Nature Genet.* **37,** 1090–1097 (2005).
49. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39,** 311–318 (2007).
50. Yusufzai, T. M., Tagami, H., Nakatani, Y. & Felsenfeld, G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol. Cell* **13,** 291–298 (2004).
51. Kim, T. H. *et al.* Direct isolation and identification of promoters in the human genome. *Genome Res.* **15,** 830–839 (2005).
52. Bieda, M., Xu, X., Singer, M. A., Green, R. & Farnham, P. J. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res.* **16,** 595–605 (2006).
53. Ruppert, S., Wang, E. H. & Tjian, R. Cloning and expression of human $TAF_{II}250$: a TBP-associated factor implicated in cell-cycle regulation. *Nature* **362,** 175–179 (1993).
54. Fernandez, P. C. *et al.* Genomic targets of the human c-Myc protein. *Genes Dev.* **17,** 1115–1129 (2003).
55. Li, Z. *et al.* A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl Acad. Sci. USA* **100,** 8164–8169 (2003).
56. Orian, A. *et al.* Genomic binding by the *Drosophila* Myc, Max, Mad/Mnt transcription factor network. *Genes Dev.* **17,** 1101–1114 (2003).
57. de Laat, W. & Grosveld, F. Spatial organization of gene expression: the active chromatin hub. *Chromosome Res.* **11,** 447–459 (2003).
58. Trinklein, N. D. *et al.* Integrated analysis of experimental datasets reveals many novel promoters in 1% of the human genome. *Genome Res.* **17,** 720–731 (2007).
59. Jeon, Y. *et al.* Temporal profile of replication of human chromosomes. *Proc. Natl Acad. Sci. USA* **102,** 6419–6424 (2005).
60. Woodfine, K. *et al.* Replication timing of the human genome. *Hum. Mol. Genet.* **13,** 191–202 (2004).
61. White, E. J. *et al.* DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. *Proc. Natl Acad. Sci. USA* **101,** 17771–17776 (2004).
62. Schubeler, D. *et al.* Genome-wide DNA replication profile for *Drosophila melanogaster*: a link between transcription and replication timing. *Nature Genet.* **32,** 438–442 (2002).
63. MacAlpine, D. M., Rodriguez, H. K. & Bell, S. P. Coordination of replication and transcription along a *Drosophila* chromosome. *Genes Dev.* **18,** 3094–3105 (2004).
64. Gilbert, D. M. Replication timing and transcriptional control: beyond cause and effect. *Curr. Opin. Cell Biol.* **14,** 377–383 (2002).
65. Schwaiger, M. & Schubeler, D. A question of timing: emerging links between transcription and replication. *Curr. Opin. Genet. Dev.* **16,** 177–183 (2006).
66. Hatton, K. S. *et al.* Replication program of active and inactive multigene families in mammalian cells. *Mol. Cell. Biol.* **8,** 2149–2158 (1988).
67. Gartler, S. M., Goldstein, L., Tyler-Freer, S. E. & Hansen, R. S. The timing of *XIST* replication: dominance of the domain. *Hum. Mol. Genet.* **8,** 1085–1089 (1999).
68. Azuara, V. *et al.* Heritable gene silencing in lymphocytes delays chromatid resolution without affecting the timing of DNA replication. *Nature Cell Biol.* **5,** 668–674 (2003).
69. Cohen, S. M., Furey, T. S., Doggett, N. A. & Kaufman, D. G. Genome-wide sequence and functional analysis of early replicating DNA in normal human fibroblasts. *BMC Genomics* **7,** 301 (2006).
70. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298,** 1039–1043 (2002).
71. Muller, J. *et al.* Histone methyltransferase activity of a *Drosophila* Polycomb group repressor complex. *Cell* **111,** 197–208 (2002).
72. Bracken, A. P., Dietrich, N., Pasini, D., Hansen, K. H. & Helin, K. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev.* **20,** 1123–1136 (2006).
73. Kirmizis, A. *et al.* Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes Dev.* **18,** 1592–1605 (2004).
74. Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125,** 301–313 (2006).
75. Karnani, N., Taylor, C., Malhotra, A. & Dutta, A. Pan-S replication patterns and chromosomal domains defined by genome tiling arrays of human chromosomes. *Genome Res.* **17,** 865–876 (2007).
76. Delaval, K., Wagschal, A. & Feil, R. Epigenetic deregulation of imprinting in congenital diseases of aberrant growth. *Bioessays* **28,** 453–459 (2006).
77. Dillon, N. Gene regulation and large-scale chromatin organization in the nucleus. *Chromosome Res.* **14,** 117–126 (2006).
78. Burkhoff, A. M. & Tullius, T. D. Structural details of an adenine tract that does not cause DNA to bend. *Nature* **331,** 455–457 (1988).
79. Price, M. A. & Tullius, T. D. How the structure of an adenine tract depends on sequence context: a new model for the structure of T_nA_n DNA sequences. *Biochemistry* **32,** 127–136 (1993).
80. Greenbaum, J. A., Parker, S. C. J. & Tullius, T. D. Detection of DNA structural motifs in functional genomic elements. *Genome Res.* **17,** 940–946 (2007).
81. Thurman, R. E., Day, N., Noble, W. S. & Stamatoyannopoulos, J. A. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.* **17,** 917–927 (2007).
82. Gilbert, N. *et al.* Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118,** 555–566 (2004).
83. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302,** 413 (2003).
84. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3,** e7 (2005).
85. Drake, J. A. *et al.* Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genet.* **38,** 223–227 (2006).
86. Margulies, E. H., Blanchette, M., NISC Comparative Sequencing Program, Haussler D. & Green, E. D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13,** 2507–2518 (2003).
87. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
88. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
89. Thomas, J. W. *et al.* Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Res.* **12,** 1277–1285 (2002).
90. Blakesley, R. W. *et al.* An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res.* **14,** 2235–2244 (2004).
91. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297,** 1301–1310 (2002).
92. Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431,** 946–957 (2004).
93. Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* **102,** 4795–4800 (2005).
94. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–715 (2004).
95. Bray, N. & Pachter, L. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.* **14,** 693–699 (2004).
96. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13,** 721–731 (2003).
97. Margulies, E. H. *et al.* Relationship between evolutionary constraint and genome function for 1% of the human genome. *Genome Res.* **17,** 760–774 (2007).
98. Asthana, S., Roytberg, M., Stamatoyannopoulos, J. A. & Sunyaev, S. Analysis of sequence conservation at nucleotide resolution. *PLoS Comp. Biol.* (submitted).
99. Cooper, G. M., Brudno, M., Green, E. D., Batzoglou, S. & Sidow, A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13,** 813–820 (2003).
100. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3,** e10 (2005).
101. Stone, E. A., Cooper, G. M. & Sidow, A. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **6,** 143–164 (2005).
102. McAuliffe, J. D., Jordan, M. I. & Pachter, L. Subtree power analysis and species selection for comparative genomics. *Proc. Natl Acad. Sci. USA* **102,** 7900–7905 (2005).
103. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
104. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437,** 88–93 (2005).
105. Cooper, G. M. *et al.* Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Res.* **14,** 539–548 (2004).
106. Dermitzakis, E. T., Reymond, A. & Antonarakis, S. E. Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nature Rev. Genet.* **6,** 151–157 (2005).
107. Clark, T. G. *et al.* Small insertions/deletions and functional constraint in the ENCODE regions. *Genome Biol.* (submitted) (2007).
108. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304,** 1321–1325 (2004).
109. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3,** e7 (2005).
110. Dermitzakis, E. T. & Clark, A. G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19,** 1114–1121 (2002).
111. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351,** 652–654 (1991).
112. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437,** 1149–1152 (2005).
113. Hudson, R. R., Kreitman, M. & Aguade, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–159 (1987).
114. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Rev. Genet.* **7,** 85–97 (2006).
115. Ludwig, M. Z. *et al.* Functional evolution of a *cis*-regulatory module. *PLoS Biol.* **3,** e93 (2005).
116. Ludwig, M. Z. & Kreitman, M. Evolutionary dynamics of the enhancer region of even-skipped in *Drosophila*. *Mol. Biol. Evol.* **12,** 1002–1011 (1995).
117. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 7, (Suppl. 1; S4) 1–9 (2006).

118. Emanuelsson, O. *et al.* Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome. *Genome Res.* advance online publication, doi: 10.1101/gr.5014606 (21 November 2006).
119. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. Science **296,** 916–919 (2002).
120. Bhinge, A. A., Kim, J., Euskirchen, G., Snyder, M. & Iyer, V. R. Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). *Genome Res.* **17,** 910–916 (2007).
121. Ng, P. *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods* **2,** 105–111 (2005).
122. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17,** 877–885 (2006).
123. Rada-Iglesias, A. *et al.* Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Hum. Mol. Genet.* **14,** 3435–3447 (2005).
124. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436,** 876–880 (2005).
125. Halees, A. S. & Weng, Z. PromoSer: improvements to the algorithm, visualization and accessibility. *Nucleic Acids Res.* **32,** W191–W194 (2004).
126. Bajic, V. B. *et al.* Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol.* 7, (Suppl 1; S3) 1–13 (2006).
127. Zheng, D. & Gerstein, M. B. A computational approach for identifying pseudogenes in the ENCODE regions. *Genome Biol.* **7,** S13.1–S13.10 (2006).
128. Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet* **1,** e78 (2005).
129. Turner, B. M. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nature Struct. Mol. Biol.* **12,** 110–112 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Leja for providing graphical expertise and support. Funding support is acknowledged from the following sources: National Institutes of Health, The European Union BioSapiens NoE, Affymetrix, Swiss National Science Foundation, the Spanish Ministerio de Educación y Ciencia, Spanish Ministry of Education and Science, CIBERESP, Genome Spain and Generalitat de Catalunya, Ministry of Education, Culture, Sports, Science and Technology of Japan, the NCCR Frontiers in Genetics, the Jérôme Lejeune Foundation, the Childcare Foundation, the Novartis Foundations, the Danish Research Council, the Swedish Research Council, the Knut and Alice Wallenberg Foundation, the Wellcome Trust, the Howard Hughes Medical Institute, the Bio-X Institute, the RIKEN Institute, the US Army, National Science Foundation, the Deutsche Forschungsgemeinschaft, the Austrian Gen-AU program, the BBSRC and The European Molecular Biology Laboratory. We thank the Barcelona SuperComputing Center and the NIH Biowulf cluster for computer facilities. The Consortium thanks the ENCODE Scientific Advisory Panel for their advice on the project: G. Weinstock, M. Cherry, G. Churchill, M. Eisen, S. Elgin, J. Lis, J. Rine, M. Vidal and P. Zamore.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. The list of individual authors is divided among the six main analysis groups and five organizational groups. Correspondence and requests for materials should be addressed to the co-chairs of the ENCODE analysis groups (listed in the Analysis Coordination group) E. Birney (birney@ebi.ac.uk); J. A. Stamatoyannopoulos (jstam@u.washington.edu); A. Dutta (ad8q@virginia.edu); R. Guigó (rguigo@imim.es); T. R. Gingeras (Tom_Gingeras@affymetrix.com); E. H. Margulies (elliott@nhgri.nih.gov); Z. Weng (zhiping@bu.edu); M. Snyder (michael.snyder@yale.edu); E. T. Dermitzakis (md4@sanger.ac.uk) or collectively (encode_chairs@ebi.ac.uk).

The ENCODE Project Consortium

Analysis Coordination Ewan Birney[1], John A. Stamatoyannopoulos[2], Anindya Dutta[3], Roderic Guigó[4,5], Thomas R. Gingeras[6], Elliott H. Margulies[7], Zhiping Weng[8,9], Michael Snyder[10,11] & Emmanouil T. Dermitzakis[12]

Chromatin and Replication John A. Stamatoyannopoulos[2], Robert E. Thurman[2,13], Michael S. Kuehn[2,13], Christopher M. Taylor[3], Shane Neph[2], Christoph M. Koch[12], Saurabh Asthana[14], Ankit Malhotra[3], Ivan Adzhubei[14], Jason A. Greenbaum[15], Robert M. Andrews[12], Paul Flicek[1], Patrick J. Boyle[3], Hua Cao[13], Nigel P. Carter[12], Gayle K. Clelland[12], Sean Davis[16], Nathan Day[2], Pawandeep Dhami[12], Shane C. Dillon[12], Michael O. Dorschner[2], Heike Fiegler[12], Paul G. Giresi[17], Jeff Goldy[2], Michael Hawrylycz[18], Andrew Haydock[2], Richard Humbert[2], Keith D. James[12], Brett E. Johnson[13], Ericka M. Johnson[13], Tristan T. Frum[13], Elizabeth R. Rosenzweig[13], Neerja Karnani[3], Kirsten Lee[2], Gregory C. Lefebvre[12], Patrick A. Navas[13], Fidencio Neri[2], Stephen C. J. Parker[15], Peter J. Sabo[2], Richard Sandstrom[2], Anthony Shafer[2], David Vetrie[12], Molly Weaver[2], Sarah Wilcox[12], Man Yu[13], Francis S. Collins[7], Job Dekker[19], Jason D. Lieb[17], Thomas D. Tullius[15], Gregory E. Crawford[20], Shamil Sunyaev[14], William S. Noble[2], Ian Dunham[12] & Anindya Dutta[3]

Genes and Transcripts Roderic Guigó[4, 5], France Denoeud[5], Alexandre Reymond[21,22], Philipp Kapranov[6], Joel Rozowsky[11], Deyou Zheng[11], Robert Castelo[5], Adam Frankish[12], Jennifer Harrow[12], Srinka Ghosh[6], Albin Sandelin[23], Ivo L. Hofacker[24], Robert Baertsch[25,26], Damian Keefe[1], Paul Flicek[1], Sujit Dike[6], Jill Cheng[6], Heather A. Hirsch[27], Edward A. Sekinger[27], Julien Lagarde[5], Josep F. Abril[5,28], Atif Shahab[29], Christoph Flamm[24,30], Claudia Fried[30], Jörg Hackermüller[32], Jana Hertel[30], Manja Lindemeyer[30], Kristin Missal[30,31], Andrea Tanzer[24,30], Stefan Washietl[24], Jan Korbel[11], Olof Emanuelsson[11], Jakob S. Pedersen[26], Nancy Holroyd[12], Ruth Taylor[12], David Swarbreck[12], Nicholas Matthews[12], Mark C. Dickson[33], Daryl J. Thomas[25,26], Matthew T. Weirauch[25], James Gilbert[12], Jorg Drenkow[6], Ian Bell[6], XiaoDong Zhao[34], K.G. Srinivasan[34], Wing-Kin Sung[34], Hong Sain Ooi[34], Kuo Ping Chiu[34], Sylvain Foissac[4], Tyler Alioto[4], Michael Brent[35], Lior Pachter[36], Michael L. Tress[37], Alfonso Valencia[37], Siew Woh Choo[34], Chiou Yu Choo[34], Catherine Ucla[22], Caroline Manzano[22], Carine Wyss[22], Evelyn Cheung[6], Taane G. Clark[38], James B. Brown[39], Madhavan Ganesh[6], Sandeep Patel[6], Hari Tammana[6], Jacqueline Chrast[21], Charlotte N. Henrichsen[21], Chikatoshi Kai[23], Jun Kawai[23,40], Ugrappa Nagalakshmi[10], Jiaqian Wu[10], Zheng Lian[41], Jin Lian[41], Peter Newburger[42], Xueqing Zhang[42], Peter Bickel[43], John S. Mattick[44], Piero Carninci[40],Yoshihide Hayashizaki[23,40], Sherman Weissman[41], Emmanouil T. Dermitzakis[12], Elliott H. Margulies[7], Tim Hubbard[12], Richard M. Myers[33], Jane Rogers[12], Peter F. Stadler[24,30,45], Todd M. Lowe[25], Chia-Lin Wei[34], Yijun Ruan[34], Michael Snyder[10,11], Ewan Birney[1], Kevin Struhl[27], Mark Gerstein[11,46,47], Stylianos E. Antonarakis[22] & Thomas R. Gingeras[6]

Integrated Analysis and Manuscript Preparation James B. Brown[39], Paul Flicek[1], Yutao Fu[8], Damian Keefe[1], Ewan Birney[1], France Denoeud[5], Mark Gerstein[11,46,47], Eric D. Green[7,48], Philipp Kapranov[6], Ulaş Karaöz[8], Richard M. Myers[33], William S. Noble[2], Alexandre Reymond[21,22], Joel Rozowsky[11], Kevin Struhl[27], Adam Siepel[25, 26]†, John A. Stamatoyannopoulos[2], Christopher M. Taylor[3], James Taylor[49,50], Robert E. Thurman[2,13], Thomas D. Tullius[15], Stefan Washietl[24] & Deyou Zheng[11]

Management Group Laura A. Liefer[51], Kris A. Wetterstrand[51], Peter J. Good[51], Elise A. Feingold[51], Mark S. Guyer[51] & Francis S. Collins[52]

Multi-species Sequence Analysis Elliott H. Margulies[7], Gregory M. Cooper[33]†, George Asimenos[53], Daryl J. Thomas[25,26], Colin N. Dewey[54], Adam Siepel[25,26]†, Ewan Birney[1], Damian Keefe[1], Minmei Hou[49,50], James Taylor[49,50], Sergey Nikolaev[22], Juan I. Montoya-Burgos[55], Ari Löytynoja[1], Simon Whelan[1]†, Fabio Pardi[1], Tim Massingham[1], James B. Brown[39], Haiyan Huang[43], Nancy R. Zhang[43,56], Peter Bickel[43], Ian Holmes[57], James C. Mullikin[7,48], Abel Ureta-Vidal[1], Benedict Paten[1], Michael Seringhaus[11], Deanna Church[58], Kate Rosenbloom[26], W. James Kent[25,26], Eric A. Stone[33], NISC Comparative Sequencing Program*, Baylor College of Medicine Human Genome Sequencing Center*, Washington University Genome Sequencing Center*, Broad Institute*, Children's Hospital Oakland Research Institute*, Mark Gerstein[11,46,47], Stylianos E. Antonarakis[22], Serafim Batzoglou[53], Nick Goldman[1], Ross C. Hardison[50,59], David Haussler[25,26,60], Webb Miller[49,50,61], Lior Pachter[36], Eric D. Green[7,48] & Arend Sidow[33,62]

Transcriptional Regulatory Elements Zhiping Weng[8,9], Nathan D. Trinklein[33]†, Yutao Fu[8], Zhengdong D. Zhang[11], Ulaş Karaöz[8], Leah Barrera[68], Rhona Stuart[68], Deyou Zheng[11], Srinka Ghosh[6], Paul Flicek[1], David C. King[50, 59], James Taylor[49, 50], Adam Ameur[69], Stefan Enroth[69], Mark C. Bieda[70], Christoph M. Koch[12], Heather A. Hirsch[27], Chia-Lin Wei[34], Jill Cheng[6], Jonghwan Kim[71], Akshay A. Bhinge[71], Paul G. Giresi[17], Nan Jiang[72], Jun Liu[34], Fei Yao[34], Wing-Kin Sung[34], Kuo Ping Chiu[34], Vinsensius B. Vega[34], Charlie W.H. Lee[34], Patrick Ng[34], Atif Shahab[29], Edward A. Sekinger[27], Annie Yang[27], Zarmik Moqtaderi[27], Zhou Zhu[27], Xiaoqin Xu[70], Sharon Squazzo[70], Matthew J. Oberley[73], David Inman[73], Michael A. Singer[72], Todd A. Richmond[72], Kyle J. Munn[72,74], Alvaro Rada-Iglesias[74], Ola Wallerman[74], Jan Komorowski[69], Gayle K. Clelland[12], Sarah Wilcox[12], Shane C. Dillon[12], Robert M. Andrews[12], Joanna C. Fowler[12], Phillippe Couttet[12], Keith D. James[12], Gregory C. Lefebvre[12], Alexander W. Bruce[12], Oliver M. Dovey[12], Peter D. Ellis[12], Pawandeep Dhami[12], Cordelia F. Langford[12], Nigel P. Carter[12], David Vetrie[12], Philipp Kapranov[6], David A. Nix[6], Ian Bell[6], Sandeep Patel[6], Joel Rozowsky[11], Ghia Euskirchen[10], Stephen Hartman[10], Jin Lian[41], Jiaqian Wu[10], Alexander E. Urban[10], Peter Kraus[10], Sara Van Calcar[68], Nate Heintzman[68], Tae Hoon Kim[68], Kun Wang[68], Chunxu Qu[68], Gary Hon[68], Rosa Luna[75], Christopher K. Glass[75], M. Geoff Rosenfeld[75], Shelley Force Aldred[33], Sara J. Cooper[33], Anason Halees[8], Jane M. Lin[9], Hennady P. Shulha[9], Xiaoling Zhang[8], Mousheng Xu[8], Jaafar N. S. Haidar[9], Yong Yu[9], Ewan Birney*,[1], Sherman Weissman[41], Yijun Ruan[34], Jason D. Lieb[17], Vishwanath R. Iyer[71], Roland D. Green[72], Thomas R. Gingeras[6], Claes Wadelius[74], Ian Dunham[12], Kevin Struhl[27], Ross C. Hardison[50,59], Mark Gerstein[11,46,47], Peggy J. Farnham[70], Richard M. Myers[33] Bing Ren[68] & Michael Snyder[10,11]

UCSC Genome Browser Daryl J. Thomas[25,26], Kate Rosenbloom[26], Rachel A. Harte[26], Angie S. Hinrichs[26], Heather Trumbower[26], Hiram Clawson[26], Jennifer Hillman-Jackson[26], Ann S. Zweig[26], Kayla Smith[26], Archana Thakkapallayil[26], Galt Barber[26], Robert M. Kuhn[26], Donna Karolchik[26], David Haussler[25,26,60] & W. James Kent[25,26]

Variation Emmanouil T. Dermitzakis[12], Lluis Armengol[76], Christine P. Bird[12], Taane G. Clark[38], Gregory M. Cooper[33]†, Paul I. W. de Bakker[77], Andrew D. Kern[26], Nuria Lopez-Bigas[5], Joel D. Martin[50,59], Barbara E. Stranger[12], Daryl J. Thomas[25,26], Abigail Woodroffe[78], Serafim Batzoglou[53], Eugene Davydov[53], Antigone Dimas[12], Eduardo Eyras[5], Ingileif B. Hallgrímsdóttir[79], Ross C. Hardison[50,59], Julian Huppert[12], Arend Sidow[33,62], James Taylor[49,50], Heather Trumbower[26], Michael C. Zody[77], Roderic Guigó[4,5], James C. Mullikin[7], Gonçalo R. Abecasis[78], Xavier Estivill[76,80] & Ewan Birney[1].

*__NISC Comparative Sequencing Program__ Gerard G. Bouffard[7,48], Xiaobin Guan[48], Nancy F. Hansen[48], Jacquelyn R. Idol[7], Valerie V.B. Maduro[7], Baishali Maskeri[48], Jennifer C. McDowell[48], Morgan Park[48], Pamela J. Thomas[48], Alice C. Young[48] & Robert W. Blakesley[7,48] **Baylor College of Medicine, Human Genome Sequencing Center** Donna M. Muzny[63], Erica Sodergren[63], David A. Wheeler[63], Kim C. Worley[63], Huaiyang Jiang[63], George M. Weinstock[63] & Richard A. Gibbs[63]; **Washington University Genome Sequencing Center** Tina Graves[64], Robert Fulton[64], Elaine R. Mardis[64] & Richard K. Wilson[64] **Broad Institute** Michele Clamp[65], James Cuff[65], Sante Gnerre[65], David B. Jaffe[65], Jean L. Chang[65], Kerstin Lindblad-Toh[65] & Eric S. Lander[65,66] **Children's Hospital Oakland Research Institute** Maxim Koriabine[67], Mikhail Nefedov[67], Kazutoyo Osoegawa[67], Yuko Yoshinaga[67], Baoli Zhu[67] & Pieter J. de Jong[67]

Affiliations for participants: [1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [2]Department of Genome Sciences, 1705 NE Pacific Street, Box 357730, University of Washington, Seattle, Washington 98195, USA. [3]Department of Biochemistry and Molecular Genetics, Jordan 1240, Box 800733, 1300 Jefferson Park Ave, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA. [4]Genomic Bioinformatics Program, Center for Genomic Regulation, [5]Research Group in Biomedical Informatics, Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra, c/o Dr. Aiguader 88, Barcelona Biomedical Research Park Building, 08003 Barcelona, Catalonia, Spain. [6]Affymetrix, Inc., Santa Clara, California 95051, USA. [7]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [8]Bioinformatics Program, Boston University, 24 Cummington St., Boston, Massachusetts 02215, USA. [9]Biomedical Engineering Department, Boston University, 44 Cummington St., Boston, Massachusetts 02215, USA. [10]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA. [11]Department of Molecular Biophysics and Biochemistry, Yale University, PO Box 208114, New Haven, Connecticut 06520, USA. [12]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. [13]Division of Medical Genetics, 1705 NE Pacific Street, Box 357720, University of Washington, Seattle, Washington 98195, USA. [14]Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [15]Department of Chemistry and Program in Bioinformatics, Boston University, 590 Commonwealth Avenue, Boston, Massachusetts 02215, USA. [16]Genetics Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [17]Department of Biology and Carolina Center for Genome Sciences, CB# 3280, 202 Fordham Hall, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [18]Allen Institute for Brain Sciences, 551 North 34th Street, Seattle, Washington 98103, USA. [19]Program in Gene Function and Expression and Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA. [20]Institute for Genome Sciences & Policy and Department of Pediatrics, 101 Science Drive, Duke University, Durham, North Carolina 27708, USA. [21]Center for Integrative Genomics, University of Lausanne, Genopode building, 1015 Lausanne, Switzerland. [22]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. [23]Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan. [24]Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria. [25]Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. [26]Center for Biomolecular Science and Engineering, Engineering 2, Suite 501, Mail Stop CBSE/ITI, University of California, Santa Cruz, California 95064, USA. [27]Department of Biological Chemistry & Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA. [28]Department of Genetics, Facultat de Biologia, Universitat de Barcelona, Av Diagonal, 645, 08028, Barcelona, Catalonia, Spain. [29]Bioinformatics Institute, 30 Biopolis Street, #07-01 Matrix, Singapore, 138671, Singapore. [30]Bioinformatics Group, Department of Computer Science, [31]Interdisciplinary Center of Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany. [32]Fraunhofer Institut für Zelltherapie und Immunologie - IZI, Deutscher Platz 5e, D-04103 Leipzig, Germany. [33]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA. [34]Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore. [35]Laboratory for Computational Genomics, Washington University, Campus Box 1045, Saint Louis, Missouri 63130, USA. [36]Department of Mathematics and Computer Science, University of California, Berkeley, California 94720, USA. [37]Spanish National Cancer Research Centre, CNIO, Madrid, E-28029, Spain. [38]Department of Epidemiology and Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK. [39]Department of Applied Science & Technology, University of California, Berkeley, California 94720, USA. [40]Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan. [41]Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA. [42]Department of Pediatrics, University of Massachusetts Medical School, 55 Lake Avenue, North Worcester, Massachusetts 01605, USA. [43]Department of Statistics, University of California, Berkeley, California 94720, USA. [44]Institute for Molecular Bioscience, University of Queensland, St. Lucia, QLD 4072, Australia. [45]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA. [46]Department of Computer Science, Yale University, PO Box 208114, New Haven, Connecticut 06520-8114, USA. [47]Program in Computational Biology & Bioinformatics, Yale University, PO Box 208114, New Haven, Connecticut 06520-8114, USA. [48]NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [49]Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [50]Center for Comparative Genomics and Bioinformatics, Huck Institutes for Life Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [51]Division of Extramural Research, National Human Genome Research Institute, National Institute of Health, 5635 Fishers Lane, Suite 4076, Bethesda, Maryland 20892-9305, USA. [52]Office of the Director, National Human Genome Research Institute, National Institute of Health, 31 Center Drive, Suite 4B09, Bethesda, Maryland 20892-2152, USA. [53]Department of Computer Science, Stanford University, Stanford, California 94305, USA. [54]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, 6720 MSC, 1300 University Ave, Madison, Wisconsin 53706, USA. [55]Department of Zoology and Animal Biology, Faculty of Sciences, University of Geneva, 1205 Geneva, Switzerland. [56]Department of Statistics, Stanford University, Stanford, California 94305, USA. [57]Department of Bioengineering, University of California, Berkeley, California 94720-1762, USA. [58]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA. [59]Department of Biochemistry and Molecular Biology, Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [60]Howard Hughes Medical Institute, University of California, Santa Cruz, California 95064, USA. [61]Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [62]Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA. [63]Human Genome Sequencing Center and Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. [64]Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, Saint Louis, Missouri 63108, USA. [65]Broad Institute of Harvard University and Massachusetts Institute of Technology, 320 Charles Street, Cambridge, Massachusetts 02141, USA. [66]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [67]Children's Hospital Oakland Research Institute, BACPAC Resources, 747 52nd Street, Oakland, California 94609, USA. [68]Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, California 92093-0653, USA. [69]The Linnaeus Centre for Bioinformatics, Uppsala University, BMC, Box 598, SE-75124 Uppsala, Sweden. [70]Department of Pharmacology and the Genome Center, University of California, Davis, California 95616, USA. [71]Institute for Cellular & Molecular Biology, The University of Texas at Austin, 1 University Station A4800, Austin, Texas 78712, USA. [72]NimbleGen Systems, Inc., 1 Science Court, Madison, Wisconsin 53711, USA. [73]University of Wisconsin Medical School, Madison, Wisconsin 53706, USA. [74]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, SE-75185 Uppsala, Sweden. [75]University of California, San Diego School of Medicine, 9500 Gilman Drive, La Jolla, California 92093, USA. [76]Genes and Disease Program, Center for Genomic Regulation, c/o Dr. Aiguader 88, Barcelona Biomedical Research Park Building, 08003 Barcelona, Catalonia, Spain. [77]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [78]Center for Statistical Genetics, Department of Biostatistics, SPH II, 1420 Washington Heights, Ann Arbor, Michigan 48109-2029, USA. [79]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK. [80]Universitat Pompeu Fabra, c/o Dr. Aiguader 88, Barcelona Biomedical Research Park Building, 08003 Barcelona, Catalonia, Spain. †Present addresses: Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA (G.M.C.); Department of Biological Statistics & Computational Biology, Cornell University, Ithaca, New York 14853, USA (A.S.); Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK (S.W.); SwitchGear Genomics, 1455 Adams Drive, Suite 2015, Menlo Park, California 94025, USA (N.D.T.; S.F.A.).

Article

Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations

David L. Mobley, Christopher I. Bayly, Matthew D. Cooper, Michael R. Shirts, and Ken A. Dill

J. Chem. Theory Comput., **2009**, 5 (2), 350-358• DOI: 10.1021/ct800409d • Publication Date (Web): 02 January 2009

JCTC Journal of Chemical Theory and Computation

Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations

David L. Mobley,*,† Christopher I. Bayly,§ Matthew D. Cooper,§ Michael R. Shirts,|| and Ken A. Dill‡

Department of Chemistry, University of New Orleans, New Orleans, Louisiana 70148 Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California 94158 Merck-Frosst Canada Ltd., 16711 TransCanada Highway, Kirkland, Quebec, Canada H9H 3L1, and Department of Chemical Engineering, University of Virginia, P.O. Box 400741, Charlottesville, Virginia 22904-4741

Received September 30, 2008

Abstract: Using molecular dynamics free energy simulations with TIP3P explicit solvent, we compute the hydration free energies of 504 neutral small organic molecules and compare them to experiments. We find, first, good general agreement between the simulations and the experiments, with an rms error of 1.24 kcal/mol over the whole set (i.e., about 2 kT) and a correlation coefficient of 0.89. Second, we use an automated procedure to identify systematic errors for some classes of compounds and suggest some improvements to the force field. We find that alkyne hydration free energies are particularly poorly predicted due to problems with a Lennard-Jones well depth and find that an alternate choice for this well depth largely rectifies the situation. Third, we study the nonpolar component of hydration free energies—that is, the part that is not due to electrostatics. While we find that repulsive and attractive components of the nonpolar part both scale roughly with surface area (or volume) of the solute, the total nonpolar free energy does not scale with the solute surface area or volume, because it is a small difference between large components and is dominated by the deviations from the trend. While the methods used here are not new, this is a more extensive test than previous explicit solvent studies, and the size of the test set allows identification of systematic problems with force field parameters for particular classes of compounds. We believe that the computed free energies and components will be valuable to others in the future development of force fields and solvation models.

I. Introduction

Aqueous solvation (hydration) of molecules is important for much of chemistry and biochemistry. Many experimental hydration free energies are available, providing a wonderful opportunity for testing force fields and computational treatments of solvation.

There have been a number of extensive tests of hydration free energies computed using continuum representations of water and static solute conformations.[1−4] One recent study extended this by sampling ensembles of solute conformations using classical molecular dynamics and using these to compute hydration free energies.[5] Continuum representations of solvent, however, have known limitations,[6,7] and explicit treatment of solvent provides a "gold standard" for molecular simulations. Early explicit solvent hydration free energy studies were limited by computational cost to a few compounds and, more recently, by the availability of

* Corresponding author e-mail: dmobley@gmail.com.
† University of New Orleans.
§ Merck-Frosst Canada Ltd.
|| University of Virginia.
‡ University of California at San Francisco.

10.1021/ct800409d CCC: $40.75
Published on Web 01/02/2009

parameters for small molecules. Thus a variety of studies have looked at hydration free energies of amino acid side chain analogs in explicit solvent (for example, refs 8−11), but few have studied a more diverse set.

With recent computational and methodological developments, both of these hurdles—computational cost and parameters—are now at least partially surmountable. Hydration free energy calculations can now be conducted more efficiently,[8,12] and computers are faster. Recent developments also make possible semiautomatic parametrization of small molecules, using general small molecule force fields like the general Amber force field (GAFF)[13] and parameter assignment tools like Antechamber.[14] Thus, two recent studies have examined hydration free energies of a total of roughly 60 small molecules in explicit solvent.[4,12]

Here, we perform a much more extensive test of explicit solvent modeling, on a test set of 504 molecules previously used for implicit solvent hydration free energy calculations[5]—more than 10 times larger than the largest previous explicit solvent tests.[12] Because this test is so extensive, we believe it provides a good benchmark for the best results that can currently be expected from molecular dynamics models of hydration. We also hope that others will find this compilation of computational and experimental results useful for analysis and force field parametrization efforts.

II. Simulation Methods

A. General Simulation Parameters. In this work, we use alchemical free energy calculations to compute hydration free energies in explicit solvent for 504 small molecules, using the compound set from a previous implicit solvent study.[5] Simulation protocols were similar to those used in previous explicit solvent studies.[4,12] Hydration free energies were computed using the Bennett acceptance ratio (BAR).[15] A brief summary of the methods follows, and we note the deviations from the previous studies.[4,12]

Here, starting molecular conformations were the same as those for the previous implicit solvent study,[5] except that here a single starting conformation was used for each molecule (rather than 5) due to computational limitations relating to the size of the set. Simulations were performed in GROMACS 3.3.1[16,17] using the GAFF small molecule parameters[13] as assigned by Antechamber[14] (as in the implicit solvent study).[5] Here, AM1-BCC[18,19] partial charges were assigned using the Merck-Frosst implementation of AM1-BCC.

This data set contains several nitro-containing compounds which did not have improper torsions for the nitro-ring system in the GAFF parameter set, specifically improper torsions for GAFF types ca-o-no-o and c3-o-no-o. We added these using generic GAFF values (that is, the values used for the majority of the improper torsions in GAFF)—a barrier height of 2.2 kcal/mol, a phase shift of 180°, and a periodicity of 2.

After setup in Antechamber and Leap, small molecule parameters were converted to GROMACS topology and coordinate files using a Perl conversion script developed previously.[20] Small molecules were then solvated using GROMACS utilities in a dodecahedral water box with at least 1.2 nm from the solute to the nearest box edge using the TIP3P model of water.[21] The number of water molecules varied depending on the solute size. Simulations were performed separately at a variety of different alchemical intermediate λ values, with the number of λ values and the amount of equilibration as described previously.[12] Production simulations were 5 ns in length at each λ value, and free energies and uncertainties were computed as described previously.[4,12] Uncertainties were computed using the block bootstrap procedure described previously. Cutoffs and simulation parameters were as described previously except that the real-space electrostatic cutoff was 10 Å rather than 9 Å.

We computed the electrostatic and nonpolar components of solvation. The electrostatic component was computed as the free energy of turning on the solute partial charges in water, less the free energy of the same transformation in vacuum. The nonpolar component was the free energy of turning on the Lennard-Jones interactions between the uncharged solute and water, as in previous studies.[4,12] Alternative definitions of the nonpolar component are possible.[44]

B. Analysis of the Nonpolar Component. In implicit solvent models, the nonpolar component of solvation is often assumed to correlate with the surface area and/or the volume based on theoretical arguments relating to cavity creation cost.[22−26] To explore this we computed the solvent accessible surface area and volume for all of the solutes considered here using GROMACS tool g_sas with a probe radius of 1.4 nm.

We also further dissected the nonpolar part (due the Lennard-Jones interactions) into repulsive and attractive components using the Weeks-Chandler-Andersen (WCA) separation.[27] To do this, we implemented the WCA separation in a modified version of GROMACS 3.3.1.[45]

In our main study, we simply computed the total nonpolar component and retained the trajectories. The attractive component for each solute was then obtained by applying the WCA separation to stored trajectories of the fully interacting solute and reprocessing these simulations with the attractive interactions turned off to re-evaluate the energies. We computed the free energy for turning off the attractive interactions using exponential averaging (the Zwanzig relation)[28] and standard error analysis. This assumes that phase-space overlap is good between the ensemble where the solute has attractive interactions with water and that where it does not. Error analysis should tell us if this is not the case. We further tested this by recomputing the attractive contribution using simulations at a series of separate λ values (where λ modifies only the attractive interactions) for selected solutes (phenol, *p*-xylene, pyridine, and toluene) and found that computed free energies were within uncertainty of the values computed using exponential averaging, indicating overlap was sufficient.

With these attractive components, we then obtained repulsive components by subtracting the attractive component from the total nonpolar component. This probably results in slightly larger uncertainties in computed repulsive compo-

nents than would have resulted from computing the repulsive component separately, but it also saves a large amount of computer time since we had already computed the total nonpolar component, and the repulsive portion of the calculation is the most difficult to converge.

C. Identification of Systematic Errors. Some functional groups may lead to systematic errors, resulting in errors which are larger for some types of molecules than for other types. Alternatively, there might be no systematic errors. We seek an approach to easily identify systematic errors and prioritize functional groups which have the largest errors.

We make a list of compounds and sort it by rms error, from largest error to smallest error. Following a method that is often used to determine *enrichment factors* for drug discovery, we look at the cumulative distribution function (CDF) for each functional group—the probability of compounds with that functional group having a ranked rms error up to rank x. Those functional groups that are systematically wrong will tend to cluster at high rms error and will result in a rapid rise in the CDF versus x. This can be assessed easily by computing the area under the CDF, biased by a weighting function to give the most weight to high rms errors. Here, we do this using the recently developed BEDROC metric,[29] which evaluates the integral of the CDF multiplied by an exponentially decaying weighting factor and then rescales this to run from 0 to 1. Chemical groups which occur most often in compounds with high rms errors will have larger BEDROC values, while chemical groups which have more random errors will have smaller BEDROC values (the expected BEDROC value for a uniform distribution can be computed analytically).[29] Chemical groups that only occur in compounds with low rms errors have the smallest BEDROC values. In Section III we report BEDROC values for a variety of chemical groups and atom types. Uncertainties were computed using the standard deviation of the mean for 40 iterations of a bootstrap procedure where BEDROC values for each chemical group are recomputed using a new list of compounds made up of a random selection of compounds from the original list.

Here, BEDROC values were computed using a weighting factor of $\alpha = 1.0$. This value was obtained empirically by experimenting with different α values to see what gave the best ability to recognize functional groups which differ substantially from random. If α is too large, the weighting is too strong, and only compounds at the very highest rms errors matter. If α is too small, making BEDROC equivalent to the ROC metric, the weighting of the early part of the curve is too weak, also apparently reducing the ability to recognize systematic errors. $\alpha = 1.0$ was a good compromise.

To avoid having to assign functional groups to all of the compounds in the test set by hand, we used the program Checkmol,[30] which automatically assigns chemical groups to molecules. We used MDL molfiles generated by OpenEye's OEChem toolkit as input. This resulted in an extremely large set of chemical groups, so we retained only those chemical groups which occurred in at least 5 molecules. We also combined some small groups. For example, we made a single group of amines, containing all types of amines. We also did the same for amides, ethers, esters, thiols, acids, and alcohols. We also manually created a "hypervalent S" group and included the appropriate compounds in this group. The resulting list of molecules assigned to chemical groups was used to generate BEDROC values for these chemical groups.

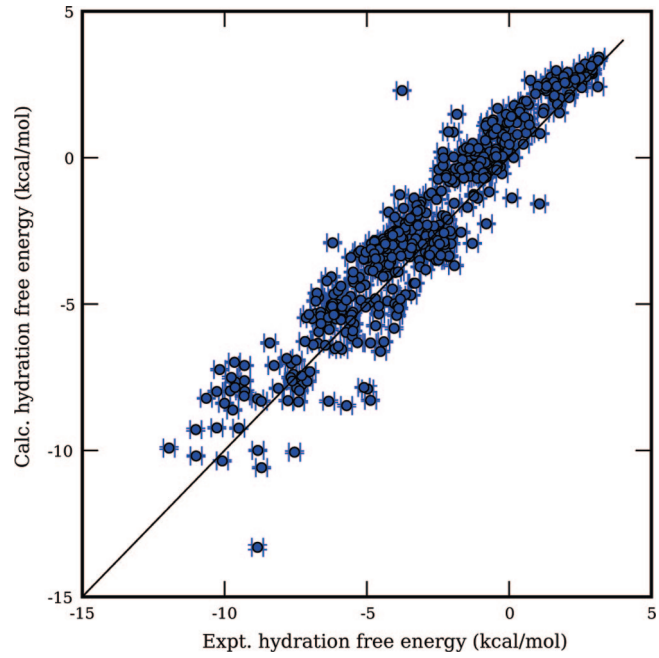


Figure 1. Calculated hydration free energies versus experiment. Shown are the calculated hydration free energies versus experiment for the full test set. The diagonal line is $x = y$. Vertical error bars denote computed uncertainties, and horizontal error bars are a conservative estimate.

We also tried using Student's *t*-test to look for systematic errors to supplement the BEDROC approach. We used our own implementations of the *t*-test and SciPy's implementation of the incomplete beta function for computing the significance. Results from this are discussed below.

III. Results and Discussion

A. The Mean Error Relative to Experiment Is Less than 1 kcal/mol. Here we evaluate the agreement between computed hydration free energies and the experimental values for the full test set. A previous study on the same 504 small-molecule test set compared the accuracy of several different implicit solvent models[5] using molecular dynamics free energy calculations. rms errors ranged from 2.014 ± 0.008 kcal/mol to 2.433 ± 0.002 kcal/mol depending on the implicit solvent model, with correlation coefficients (r^2) from 0.685 ± 0.001 to 0.774 ± 0.001. In all four solvent models tested, the computed hydration free energies were systematically too negative relative to experiments (the solutes preferred the water phase too much in the simulations), so the mean error was negative (-0.65 ± 0.09 to -1.1 ± 0.1).

Here, using explicit TIP3P water, we find an rms error of 1.26 ± 0.01 kcal/mol, with a correlation coefficient of 0.889 ± 0.006 and a mean error of 0.676 ± 0.002 (Figure 1). Hence, on average, explicit solvent simulations give significantly better agreement with experiments than our earlier implicit solvent study, consistent with an earlier comparison.[4] Interestingly, the systematic errors of explicit and implicit

solvent studies are in opposite directions. In explicit solvent, the hydration free energies here are systematically too positive. These differences are likely due to the solvent models rather than the force field parameters, since the solute parameters are very similar in the two cases. Systematic errors in other explicit solvent models tended to be in the same direction as the explicit solvent deviation here,[8] so perhaps limitations of the water model are playing a role. Another potential source of such differences is the neglect, in implicit solvent models, of asymmetries in the response of water to solutes of different polarities.[7] Another origin of differences could be the nonpolar term in the implicit models. That is, the term $\gamma \times A$ (where A is the surface area) in implicit solvent models involves an adjustable parameter which can change the errors. A table of the full results from this study is available in the Supporting Information.

B. Improving the Alkyne Lennard-Jones Parameters and Identifying Other Systematic Errors. Are there systematic errors in the force field parameters for molecules in our test set? We found that the computed hydration free energies for alkynes were systematically much too positive (Figure 1 and the Supporting Information). There were six alkynes in the set, and the mean error was 1.92 ± 0.21 kcal/mol. All of the computed hydration free energies were actually around 2 kcal/mol, while experimental values are close to zero. For all of the alkynes, the electrostatic component of hydration is quite small (-0.8 to -0.9 kcal/mol), since these molecules are largely nonpolar. We reasoned that errors in alkyne parameters are thus not likely to be in the electrostatic terms, nor are the errors expected to come from the bonded parameters (bond stretching, angle bending, etc.), which should affect hydration free energies only weakly. Hence, we focused on the alkyne Lennard-Jones parameters. In GAFF, the alkyne carbon Lennard-Jones parameters are identical to those for all carbons except selected sp2 carbons (the 'c2' atom type) and are taken directly from comparable carbons in older AMBER force fields.[13] We were particularly concerned about the parameters for the GAFF "c1" atom type, for the triple bonded carbons in alkynes. These apparently originated with the work of Howard et al., where they "were obtained by analogy to the Weiner et al. and Cornell et al. force fields".[31] In that work, those Lennard-Jones parameters were taken to be the same as for the other carbons.

Many AMBER Lennard-Jones parameters were originally taken from the OPLS force field, so we examined the OPLS choices for triple bonded carbons. It turned out that OPLS uses several different atom types for alkyne carbons, originating from simulations of linear and substituted alkynes,[32−34] and some of these have much stronger dispersion interactions than those for the GAFF c1 type, which is intuitively reasonable. It seemed likely that missing dispersion interactions could account for at least part of the error we were seeing for alkynes, thus we examined modifying the Lennard-Jones well-depth for alkynes in GAFF.

We sought to avoid adding additional atom types to GAFF, but OPLS has several different carbon well-depths for alkynes, depending on whether the carbon is terminal ($\varepsilon = 0.086$ kcal/mol), nonterminal with an attached atom having two or three hydrogens ($\varepsilon = 0.210$ kcal/mol), nonterminal with an attached atom having one hydrogen ($\varepsilon = 0.135$ kcal/mol), or nonterminal with an attached phenyl or other atom having no hydrogens ($\varepsilon = 0.100$ kcal/mol).[32−34] To avoid adding additional atom types to GAFF, we needed to pick just one of these, so we chose the one which gave the most accurate hydration free energies when used for all alkyne triple bonded carbons. This was $\varepsilon = 0.210$ kcal/mol. The original GAFF well depth was $\varepsilon = 0.086$ kcal/mol.

Using this new ε value for triple-bonded carbons, the computed hydration free energies for alkynes are much closer to zero (although still slightly positive); now the mean error is 0.49 ± 0.07 kcal/mol, down from 1.92 ± 0.21 kcal/mol initially. Increasing the well depth further could reduce this somewhat more, but this might cause other inconsistencies within the force field. Nevertheless, the systematic error here on alkynes is compelling, and we recommend that future GAFF studies use a well depth of $\varepsilon = 0.210$ kcal/mol for triple bonded carbons (GAFF types c1, cg, and ch).[46]

The alkynes also provide an example of how the BEDROC metric works for identifying systematic errors. Before the adjustment of the well depth for alkynes, the BEDROC value (with $\alpha = 1$) for alkynes was 0.90 ± 0.02 (compared to 0.49 for a random distribution with this α),[47] indicating that alkynes were systematically wrong. After the fix, the BEDROC value was 0.26 ± 0.05, indicating that alkynes now actually are considerably better than other typical compounds (Figure 2). Although our correction of ε was done without regard for the carbonitriles, the change results in a decrease in BEDROC for the carbonitriles from 0.86 ± 0.05 to 0.73 ± 0.06 (compared to 0.49 for uniform). So carbonitriles are now improved too but still have substantial systematic errors. With this change, the overall rms error decreases slightly to 1.24 ± 0.01 kcal/mol, and the correlation coefficient remains essentially the same (0.891 ± 0.006). In all that follows we report values computed with the new well depth.

We believe that the approach utilized here (looking for compounds that are over-represented at the highest rms errors) is a general and useful strategy for identifying systematic flaws in the energy functions used for molecular modeling simulations and prioritize reparameterization efforts. Functional groups which tend to cause significant errors should occur frequently at the high-rms error end of the set, while functional groups which are not necessarily linked to the errors should be roughly randomly distributed over the test set. For example, one would intuitively expect that whether a compound is aromatic or not will have little to do with whether it is systematically mispredicted. Indeed, aromatic compounds have a BEDROC value of 0.48 ± 0.03, roughly randomly distributed (Figure 2). BEDROC values by functional group for the set are shown in Table 1. These BEDROC values show that cyclic hydrocarbons, alkynes (with the fix), alkanes, aldehydes, and ketones are now particularly well predicted. On the other hand, there appear to be systematic errors for alcohols, alkyl bromides, and carbonitriles.

We also tried another approach for identifying systematic errors involving Student's *t*-test, which compares the means

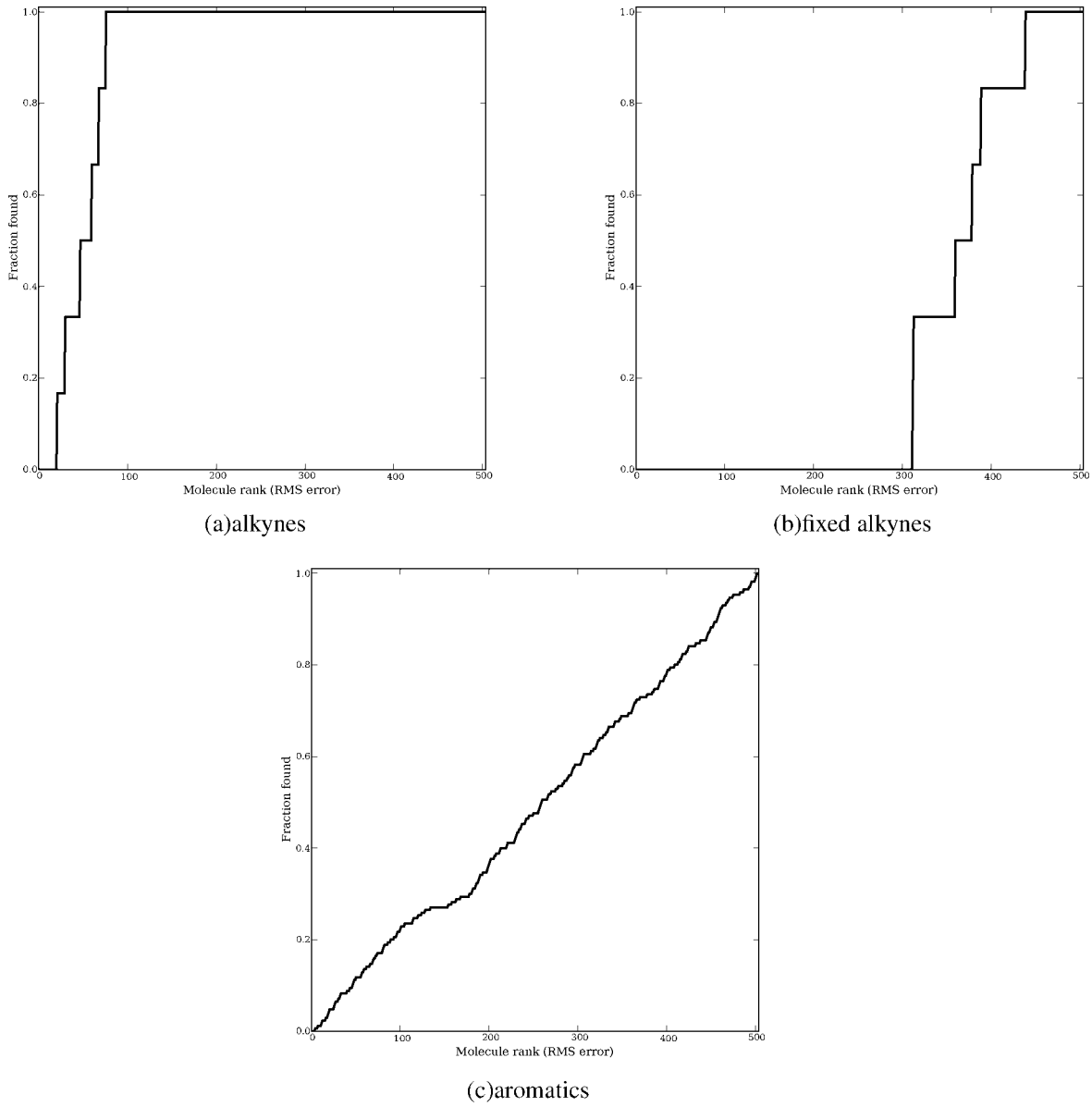


Figure 2. CDFs for selected functional groups versus error. Shown are cumulative distribution functions for finding compounds with particular functional groups at a given ranked error. Compounds found far to the left have very large errors; compounds far to the right have very small errors. An ideal random distribution of errors would give rise to a linear rise in the CDF. CDFs are shown for (a) alkynes before fixing the Lennard-Jones well-depth, (b) alkynes after fixing the Lennard-Jones well-depth, and (c) aromatics.

of two distributions and provides a measure of the significance of any difference in the means. We applied this approach in two different ways:

(1) We compared the mean experimental value for each functional group with the mean calculated value for each functional group (Supporting Information, Table 5). This proved not to be particularly useful, as these means are significantly different for almost every functional group. This is not surprising given the fact that the mean error across the entire test set is 0.676 ± 0.002, so most computed values (in all functional groups) are too positive. This does show that results could be improved across the entire set by addressing this systematic offset, but it does not provide any insight into which functional groups are particularly problematic.

(2) We compared the error for the compounds in each functional group with the error for the entire set (Table 2). This shows which functional groups have a significantly *different* performance than the overall set, though this performance could be better or worse. We also show the mean error for each functional group in Table 2; functional groups with mean errors around 0.676 kcal/mol are typical, while those with larger mean errors are worse than average, and those with smaller mean errors are better than average. The *t*-test tells us which of these differences are significant, and many are. This appears to be a useful analysis that complements the BEDROC analysis. The advantage of the BEDROC analysis is that it tells us which functional groups have the worst errors, while this analysis can tell us which functional groups have the most significant errors.

Table 1. BEDROC Values by Functional Group for the Different Functional Groups Represented in the Test Set, Compared to What Would Be Expected for the Same Number of Compounds Distributed Randomly Across the Test Set[a]

functional group	number	BEDROC
acid	73	0.48 ± 0.03
alcohol	38	0.76 ± 0.03
aldehyde	20	0.22 ± 0.04
alkanes	28	0.16 ± 0.03
alkene	35	0.55 ± 0.04
alkyl bromide	17	0.72 ± 0.08
alkyl chloride	31	0.61 ± 0.05
alkyl iodide	9	0.44 ± 0.06
alkyne	6	0.26 ± 0.04
amine	44	0.47 ± 0.04
aromatic compound	170	0.48 ± 0.03
aryl chloride	20	0.54 ± 0.05
carbonitrile	12	0.73 ± 0.07
cyclic hydrocarbon	8	0.14 ± 0.03
ester	8	0.46 ± 0.11
ether	42	0.60 ± 0.04
halogen derivative	22	0.58 ± 0.07
heterocyclic compound	48	0.60 ± 0.04
hypervalent S	5	0.62 ± 0.20
ketone	25	0.26 ± 0.06
nitro compound	17	0.63 ± 0.08
other	29	0.62 ± 0.06
phenol or hydroxyhetarene	33	0.60 ± 0.05
thiol	5	0.46 ± 0.04

[a] Functional groups with high BEDROC values (relative to the value for random, roughly 0.5 here) are overrepresented in compounds with high RMS errors.

Table 2. Statistics from Applying Student's *t*-Test to the Difference between the Calculated and Experimental Means by Functional Group[a]

functional group	number	*t*-value	significance	mean error
acid	73	−7.43	4e-13	−0.34
alcohol	38	3.62	0.0003	1.29
aldehyde	20	−3.04	0.003	−0.07
alkanes	28	−1.69	0.09	0.31
alkene	35	2.34	0.02	1.07
alkyl bromide	17	3.31	0.001	1.50
alkyl chloride	31	2.31	0.02	1.09
alkyl iodide	9	0.59	0.6	0.86
alkyne	6	−0.38	0.7	0.49
amine	44	−0.65	0.5	0.55
aromatic compound	170	−1.05	0.3	0.55
aryl chloride	20	1.65	0.1	1.04
carbonitrile	12	3.22	0.001	1.63
cyclic hydrocarbon	8	−1.18	0.2	0.21
ester	8	−1.69	0.09	0.02
ether	42	2.18	0.03	1.01
halogen derivative	22	0.32	0.8	0.73
heterocyclic compound	48	2.38	0.02	1.02
hypervalent S	5	−4.55	7e-06	−1.50
ketone	25	−2.77	0.006	0.05
nitro compound	17	1.86	0.06	1.13
other	29	−0.48	0.6	0.55
phenol or hydroxyhetarene	33	2.72	0.007	1.16
thiol	5	0.51	0.6	0.89

[a] Shown are the number of compounds in each functional group, the calculated *t* value, the computed significance (probability that *t* could be this large or larger by chance), and the mean error for this group (in kcal/mol). The overall mean error is 0.676 ± 0.002 kcal/mol, so groups with mean errors smaller than this may be significantly better than average (until the mean error becomes negative), while those with mean errors larger than this may be significantly worse.

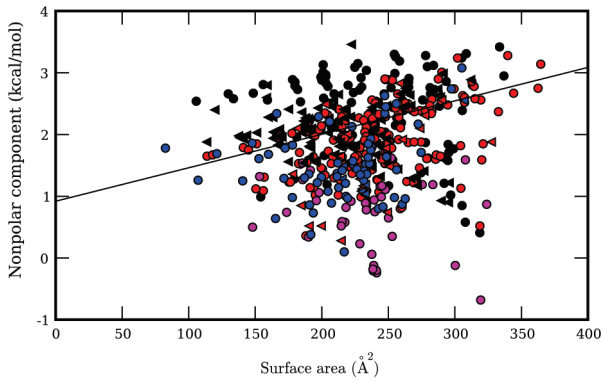
The study done here uses one particular charge model. Charge model may affect which compounds are particularly poorly predicted, though in two recent tests, the compounds which were poorly predicted tended to be poorly predicted by most charge models.[4,35] Still, our analysis here does not in general point to a specific source of error. Errors may be due to the charge model, Lennard-Jones parameters, or bonded parameters, some combination, or even due to the water model. In the case of the alkynes, we can be fairly confident that the source of error is the Lennard-Jones parameters for the reasons noted above. But for the other cases noted here, further work will be required to determine the source of error.

C. The Total Nonpolar Component Does Not Correlate with Surface Area or Volume. We examined the nonpolar components (the nonelectrostatic component of the hydration free energy) for our data set. The total nonpolar contribution to the solvation free energy is typically assumed to correlate with surface area or volume in implicit solvent models. Yet we find that there is essentially no correlation. Plots of nonpolar components versus surface area and volume are shown in Figure 3. The correlation of the nonpolar component with surface area is $r^2 = 0.019 \pm 0.001$, and that with volume is $r^2 = 0.011 \pm 0.001$. The molecules in this test set are small enough that surface area and volume are highly correlated ($r^2 = 0.991 \pm 0.001$).

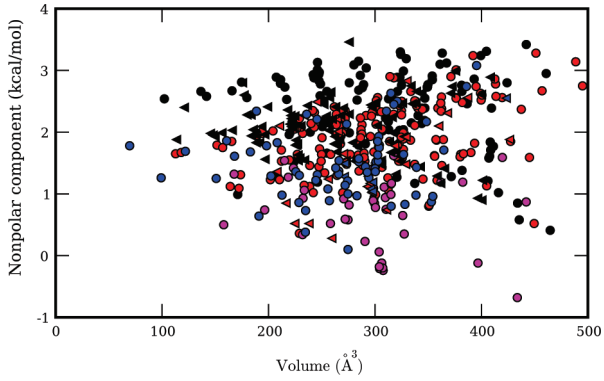
We further dissect the nonpolar component using the WCA separation of the Lennard-Jones potential energy (and thus the nonpolar component) into attractive and repulsive parts. The potential is split based on the sign of the force, as discussed in the Methods section. We find that both the attractive and repulsive components individually correlate strongly with surface area and volume (repulsive: $r^2 = 0.964 \pm 0.002$ with surface area, $r^2 = 0.952 \pm 0.002$ with volume; attractive: $r^2 = 0.945 \pm 0.002$ with surface area, $r^2 = 0.946 \pm 0.002$ with volume; Figure 4), and it is only the total (the small difference of the two large individual components) that does not correlate well with surface area or volume. This is in accord with previous work on a smaller set of compounds.[36] Essentially, the total nonpolar component is the sum of two anticorrelated quantities, and so the total ends up being dominated by the scatter in these quantities. It is interesting to note that the minimum in the Lennard-Jones potential is precisely where these two forces, the attractive and repulsive components, are very well balanced, so it is perhaps not surprising that the attractive and repulsive components correlate so well.

The observed poor correlation, and the importance of attractive interactions, is consistent with several previous studies which have found that the nonpolar component of solvation does not correlate well with surface area.[36−39]

Why is the correlation with surface area so poor? In Figure 3, it is apparent that compounds containing only carbon and hydrogen have a nonpolar component that is less favorable to solvation than molecules of an equal size which additionally contain nitrogen and/or oxygen. The likely reason for this is that nitrogen and oxygen atoms tend to have stronger attractive dispersion interactions with their environment than

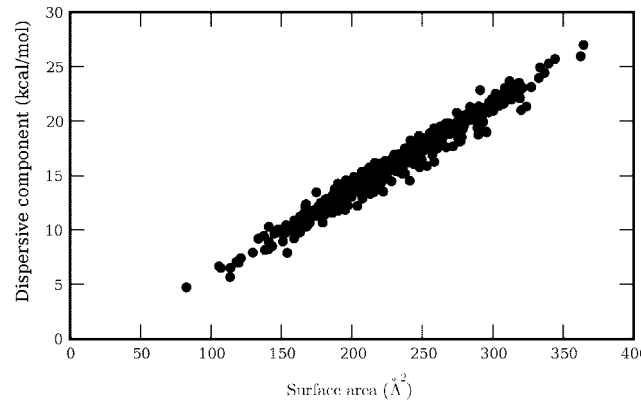


(a)Nonpolar component versus surface area

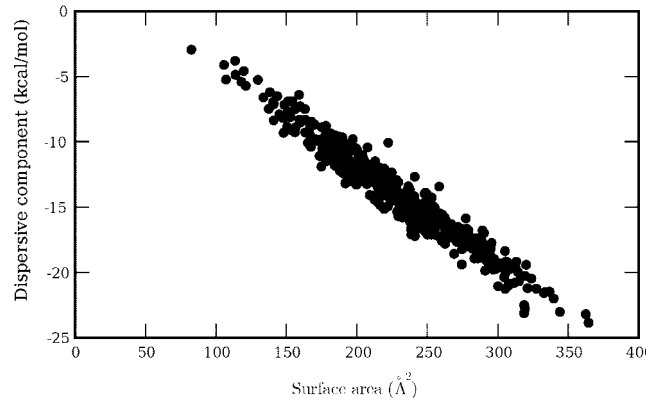


(b)Nonpolar component versus volume

Figure 3. Nonpolar components versus solvent accessible surface area and volume. Shown are the calculated nonpolar component of the hydration free energies versus solvent accessible surface area and volume for the compounds in the set. Carbon and hydrogen containing compounds are black, those with oxygen additionally are red, those with nitrogen additionally are blue, and those with nitrogen and oxygen both are magenta. Compounds with diamond symbols contain other elements in addition to C, H, N, and O. In the surface area plot, the line is a typical implicit solvent nonpolar component estimate of $G_{np} = (0.00542 \cdot SA + 0.92)$ kcal/mol[1].



(a)Repulsive part of nonpolar component versus surface area



(b)Attractive part of nonpolar component versus surface area

Figure 4. Repulsive and attractive parts of the nonpolar component versus surface area. Shown are the repulsive (a) and attractive (b) parts of the nonpolar component, as calculated using the WCA separation, plotted versus the solvent accessible surface area for solutes in the test set. Similar plots comparing the repulsive and attractive components to volume are given in the Supporting Information.

do carbon and hydrogen. Several other studies have noted that dispersion interactions play an important role in nonpolar solvation.[36−39] Even interior solute atoms contribute to these attractive interactions in an important way.[40] Other factors may also contribute to the poor correlation with surface area. For example, geometric effects may play an important role as well.

Overall, our results strongly support the growing consensus that implicit solvent models should move beyond the simple surface area model for treatment of the nonpolar component, perhaps at least to include a treatment of dispersion interactions. A number of alternate models have already been proposed.[36,38,39,41,42]

IV. Conclusions

We used molecular dynamics simulations in explicit TIP3P water to compute the hydration free energies for a set of 504 neutral compounds. We compared the results with experimental data in the most extensive such test in explicit solvent to date. We find a good correlation (r^2 of 0.891 ± 0.06) and an rms error of 1.24 ± 0.01 kcal/mol or roughly 2 kT. We believe this is representative of the accuracy that can be expected from the best current physical models for hydration free energies. It may be possible to develop new models which can do somewhat better, though we expect that it may be very hard to increase accuracies past 1 kT. A key finding is that these explicit solvent free energies are considerably more accurate than the corresponding implicit solvent values for the same data set.

At the same time, many of the molecules in this test set are relatively small and simple compared to typical druglike molecules, which may be highly polyfunctional. Recent work suggests that overall performance of the approach applied here may be significantly worse in tests where the compounds involved are more polar and polyfunctional.[4,35] This may suggest we need much more hydration free energy data on more polyfunctional, druglike molecules in order to refine our force fields.

Here, we also propose a way to identify systematic errors in force field parameters for particular functional groups. We do this using the BEDROC method.[29] Using this approach,

we were able to fix a systematic problem with alkyne Lennard-Jones parameters. We also identified several other classes of compounds which appear to have systematic errors, and for which further force field development should be done. Having a method to systematically identify problematic compound classes provides good opportunities for force field improvements.

In addition, we studied the nonpolar component of the hydration free energy for the compounds in the test set. We find that while the large repulsion and attraction terms both correlate well with the size (area or volume) of the solute, the total nonpolar component, which is a small difference between these two quantities, does not. This strongly suggests that implicit solvent models need to move away from treating the nonpolar component as simply dependent on the surface area. The data additionally suggest that new models must address the nonlinear behavior arising from the delicate balance of repulsive and attractive nonpolar terms. Furthermore, implicit solvent models that have been parametrized to match experimental hydration free energies using a simple surface area-based nonpolar term may need to be reparameterized.

Here, the real value of this study is not the methods presented—the methods were used in previous work. Rather, it is the extensive nature of the test, which provides the opportunity to actually identify systematic errors in the force field descriptions of particular functional groups. It also provides guidance into what compounds are likely to be particularly challenging to study computationally with current force fields.

Because we believe the real value of this study is these results, we have deposited the full set of computed free energies, components, starting molecular structures, and parameters for this work in the Supporting Information. We hope that others find this experimental data set and the computational results to be useful in future studies of solvation and for force field development.

Acknowledgment. We thank John D. Chodera (Stanford University) for helpful discussions. We appreciate the support of NIH grant GM 63592 to K.A.D.

Supporting Information Available: Coordinate files (mol2) with AM1-BCC partial charges for the small molecules in the test set used here; computed hydration free energies, electrostatic and nonpolar components, and the experimental values; AMBER parameter and coordinate files for all of the molecules in the test set; plots of attractive and repulsive components versus solute volume; a table mapping the names used for the files to IUPAC names; a table of computed solvent accessible surface area and volume for each solute; and results from Student's *t*-test comparing the mean experimental and calculated values for each functional group.This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. *J. Chem. Theory Comput.* **2006**, *2*, 128–139.

(2) Bordner, A. J.; Cavasotto, C. N.; Abagyan, R. A *J. Phys. Chem. B* **2002**, *106*, 11009–11015.

(3) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6532–6542.

(4) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. *J. Med. Chem.* **2008**, *51*, 769–778.

(5) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2008**, *111*, 938–946.

(6) Chorny, I.; Dill, K. A.; Jacobson, M. P. *J. Phys. Chem. B* **2005**, *109*, 24056–24060.

(7) Mobley, D. L.; Barber, A. E., II; Fennell, C. J.; Dill, K. A. *J. Phys. Chem. B* **2008**, *112*, 2405–2414.

(8) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.

(9) Hess, B.; van der Vegt, N. F. A. *J. Phys. Chem. B* **2006**, *110*, 17616–17626.

(10) Deng, Y.; Roux, B. *J. Chem. Phys.* **2004**, *108*, 16567–16576.

(11) Villa, A.; Mark, A. E. *J. Comput. Chem.* **2002**, *23*, 548–553.

(12) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.

(13) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(14) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *26*, 247260.

(15) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.

(16) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(17) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.

(18) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.

(19) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(20) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Phys.* **2006**, *125*, 084902.

(21) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(22) Stillinger, F. H. *J. Solution Chem.* **1973**, *2*, 141–158.

(23) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717–726.

(24) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754.

(25) Chothia, C. *Nature* **1974**, *248*, 338.

(26) Reynolds, J. A.; Gilbert, D. B.; Tanford, C. *Proc. Natl. Acad. Sci. U.S.A.* **1974**, *71*, 2925.

(27) Weeks, J. D.; Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1971**, *54*, 5237–5247.

(28) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

(29) Truchon, J.-F.; Bayly, C. I. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(30) Haider, N. Checkmol. http://merian.pch.univie.ac.at/ nhaider/cheminf/cmmm.html (accessed July 20, 2007).

(31) Howard, A. E.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 243–261.

(32) Jorgensen, W. L., personal communication, 2007.

(33) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665–6670.

(34) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689–1700.

(35) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. J. Phys. Chem. BAccepted for publication.

(36) Tan, C.; Tan, Y.-H.; Luo, R. *J. Phys. Chem. B* **2007**, *111*, 12263–12274.

(37) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. *J. Am. Chem. Soc.* **1999**, *121*, 9243–9244.

(38) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.

(39) Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8331–8336.

(40) Pitera, J. W.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2001**, *123*, 3163–3164.

(41) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.

(42) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.

(43) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740–5761.

(44) Here, the alchemical pathway used turns off all solute partial charges, meaning that the nonpolar component is calculated without solute intramolecular electrostatic interactions. An alternative pathway[8,43] involves turning off only intermolecular electrostatic interactions between the solute and its environment, while maintaining intramolecular electrostatic interactions. While the two pathways must give equivalent hydration free energies, the breakdown into electrostatic and nonpolar components will be slightly different, since the conformational ensemble sampled during the nonpolar component of the calculation will be altered by the presence (or lack thereof) of intramolecular electrostatic interactions.

(45) Other separation schemes and probe radii are possible, as are other surface definitions, but our main conclusions here should not depend significantly on these factors, as suggested by the work of Tan et al.[36]

(46) We do not believe it is necessary to perform additional tests to examine other alkyne properties (such as pure liquid properties) before making this recommendation for two reasons. First, the AMBER force field does not typically use these properties to inform the parameterization process, so including them would be a deviation from AMBER parameterization strategies. Second, the AMBER force field and GAFF claim (in the force field files) to use the OPLS Lennard-Jones parameters for alkynes. The suggested modification simply makes this claim true and brings AMBER/GAFF back into conformity with OPLS.

(47) BEDROC values for a random distribution actually depend on the number of compounds being considered relative to the total. But here the BEDROC value for the random distribution is 0.49 for all of the sizes of our chemical groups except for aromatics, where it is 0.50. To simplify our tables, then, we simply compare all BEDROC values to 0.49.

CT800409D

THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations

David L. Mobley, Christopher I. Bayly, Matthew D. Cooper, and Ken A. Dill

J. Phys. Chem. B, **Article ASAP** • DOI: 10.1021/jp806838b • Publication Date (Web): 09 March 2009

Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations[†]

David L. Mobley,*[,‡] Christopher I. Bayly,[§] Matthew D. Cooper,[§] and Ken A. Dill[||]

Department of Chemistry, University of New Orleans, New Orleans, Louisiana 70148, Merck-Frosst Canada Ltd., 16711 TransCanada Highway, Kirkland, Quebec, Canada H9H 3L1, and Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California 94158

Received: July 31, 2008; Revised Manuscript Received: October 3, 2008

Here, we computed the aqueous solvation (hydration) free energies of 52 small drug-like molecules using an all-atom force field in explicit water. This differs from previous studies in that (1) this was a blind test (in an event called SAMPL sponsored by OpenEye Software) and (2) the test compounds were considerably more challenging than have been used in the past in typical solvation tests of all-atom models. Overall, we found good correlations with experimental values which were subsequently made available, but the variances are large compared to those in previous tests. We tested several different charge models and found that several standard charge models performed relatively well. We found that hypervalent sulfur and phosphorus compounds are not well handled using current force field parameters and suggest several other possible systematic errors. Overall, blind tests like these appear to provide significant opportunities for improving force fields and solvent models.

1. Introduction

Hydration free energies provide an important metric of the accuracy of physics-based methods used in molecular simulations. Since these can now be calculated very precisely, they can be compared with experiment to test force fields and identify systematic errors.[1−4] They also can provide insight into the underlying solvation effects such as hydrophobicity,[5] surface effects,[6] and solvent asymmetries.[7] For these and other reasons, there have been a wide range of recent computational studies of small molecule hydration free energies from explicit solvent simulations.[1,2,4,8−14]

A major advantage of physical methods is their potential ability to predict properties of compounds that have not been previously studied. Ideally this ability could be used in drug discovery and other applications. With this in mind, it is important to test methods not only in retrospective tests but also prospectively, as they would be used in real applications. Here we report the results of a blind test for computing hydration free energies with explicit solvent molecular dynamics simulations. This test was done as part of OpenEye's Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenge. Hydration free energies were computed with no knowledge of experimental values, then submitted to the moderators of the SAMPL project, who then provided the experimental values.

2. Methods

Starting mol2 files were provided by the organizers of OpenEye's SAMPL event; names and 2D structures are provided in the work of Guthrie.[15] We then prepared five partial charge sets for use with AMBER small molecule parameters: a negative control, two positive controls, and two sets for testing. The negative control was Merck molecular mechanics force field (MMFF) charges, which we expected to perform poorly.[16] Positive controls were RESP HF/6-31G* and AM1-BCC partial charges. We tested PM3-BCC v0.2 and PM3-BCC v0.3 partial charges, which are under development by C. I. Bayly and collaborators as potential successors to AM1-BCC. MMFF charges were computed using routine Merck & Co. internal software. RESP charges were computed as described previously,[16] except that a B3LYP (cc-pVTZ) minimization was done on an extended conformation holding all non-H-containing dihedrals constant; the restraint weight was 0.001 in both stages; and for all topologically equivalent atoms, charges were averaged as the last step. For time reasons, geometry optimization was not entirely completed for molecule 23, though the forces were in the last significant figure before the convergence threshold. AM1-BCC charges were computed as described previously.[16,17]

The approach for the free energy calculations here was very similar to that in several previous studies of hydration free energies.[2,4,11] We used explicit solvent molecular dynamics simulations with the TIP3P water model[18] and Amber GAFF[19,20] small molecule parameters. Simulations were conducted using the April 2, 2007 CVS version of the GROMACS 3.3.1 software package[21] (which incorporated several bugfixes past the 3.3.1 release itself). The hydration free energy calculations involved several components as described previously,[4] with each simulation conducted independently from the same starting structure. First, solute electrostatics are turned off in water linearly with the variable λ (where $\lambda = 0$, 0.25, 0.5, 0.75, and 1.0 in turn). Second, solute−water Lennard-Jones interactions are turned off in water using soft core potentials[22] with the parameters suggested by Shirts[8] ($\alpha = 0.5$, with a soft core exponent of 1), as previously.[4] For this step, λ values were 0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, and 1.0. Finally, solute electrostatics were turned back on in vacuum (with $\lambda = 0$, 0.25, 0.5, 0.75, and 1.0). The free energy of each of these component steps was computed using the Bennett

[†] Part of the special section "Calculation of Aqueous Solvation Energies of Drug-Like Molecules: A Blind Challenge".

* To whom correspondence should be addressed. E-mail: dmobley@gmail.com.

[‡] University of New Orleans.

[§] Merck-Frosst Canada Ltd.

[||] University of California at San Francisco.

10.1021/jp806838b CCC: $40.75

acceptance ratio (BAR)[23] and then the total hydration free energy was computed as $\Delta G_{hyd} = \Delta G_{chg,vac} - \Delta G_{chg} - \Delta G_{LJ}$, where ΔG_{chg} denotes the free energy of turning off the electrostatics in water, $\Delta G_{chg,vac}$ denotes the same quantity for vacuum, and ΔG_{LJ} denotes the free energy of turning off the solute-water Lennard-Jones interactions in water.

Protocols were generally as described previously.[4] Briefly, at each lambda value, the (same) starting structure was minimized using steepest descents minimization. The resulting structures were then run through an equilibration procedure consisting of 10 ps of constant volume equilibration, followed by 100 ps of constant pressure equilibration. Production simulations were 5 ns at each λ. We did make some minor modifications to our previous protocols. We replaced the L-BFGS minimizations with up to 5000 steps of steepest descents minimization for each molecule at each λ value (because the GROMACS L-BFGS minimizer would often terminate too early, resulting in forces that remained too large; we achieved better minimization with the steepest descents minimizer). For simulations in water, we used a neighbor list cutoff of 1.0 Å and an electrostatic cutoff of 1.0 Å (this change was because the 3.3.1 version of GROMACS requires these two cutoffs be equal when using lattice-sum electrostatics). Small molecules were solvated using GROMACS utilities in a dodecahedral simulation box with at least 1.2 nm from the solute to the nearest simulation box edge; in some cases, previous simulations used slightly smaller box sizes. For each charge set, a separate set of electrostatic annihilation calculations was performed, rather than computing the free energy of changing the charges from a reference set as in the previous work (this change was to avoid introducing the potential for additional error by adding an extra step to the calculation). Additionally, following constant pressure equilibration, at each lambda value, we performed an affine transformation on the atomic coordinates to scale the volume to the average box volume from the constant pressure equilibration. The box was then fixed at this size during the subsequent constant volume simulations, and an additional 100 ps of data was discarded to equilibration before collecting data for analysis. This change was made because occasionally box volumes at the end of constant pressure equilibration could be far from the mean, and fixing the box volume to this value for constant volume production could lead to artifactual densities. Adding the affine transformation ensures the box volume, and hence density, is correct for the constant volume production. Data and error analysis was as described previously;[2,4,11] computed uncertainties reported in the Supporting Information represent the estimated standard error in the mean. Nonpolar components, which do not depend on the charge model, were only calculated once.

WIth the data we generate, we want to be able to identify whether there are particular functional groups that tend to cause systematic errors, or whether errors are not particularly linked to functional groups. We begin with the realization that, if a functional group is not associated with systematic errors, it should be roughly as likely to occur in compounds that have large errors relative to experiment as in compounds with small errors relative to experiment. For example, a previous study found that whether a compound is aromatic or not has no bearing on whether it is well- or poorly predicted.[11] So, to identify systematic errors, what we seek to find is chemical groups that are statistically over-represented in the compounds with the largest errors. There are many potential ways to perform such a search, and here we choose just one such way that appears to work well for us. We first sort the molecules by the absolute value of the error relative to experiment, from largest to smallest errors. We then use the package Checkmol[24] to group compounds by functional group. We then want a statistical metric to assess which are over-represented at the largest errors. We choose the BEDROC metric[25] for this task, as in one previous study.[11] Basically, BEDROC computes a Boltzmann-weighted area under the cumulative probability distribution function for finding compounds (with a particular chemical group) at a particular error, then rescales the resulting numbers to fall between 0 and 1. The weighting simply makes the early (high-error) part of the curve dominate. Here, we compute BEDROC values (with $\alpha = 1$) for different functional groups. Those functional groups which have particularly high BEDROC values (relative to random) are typically associated with large errors (and thus may have parameter or other problems), as noted previously.[11] Thus, the BEDROC values we report here are simply a numerical metric that tells us whether or not a particular functional group is especially likely to be associated with large errors relative to experiment.

Experimental results are taken from the tables of Guthrie,[15] and experimental error bars shown in the plot are taken from the uncertainty estimates described in that work. Some potential sources of error are discussed there as well.

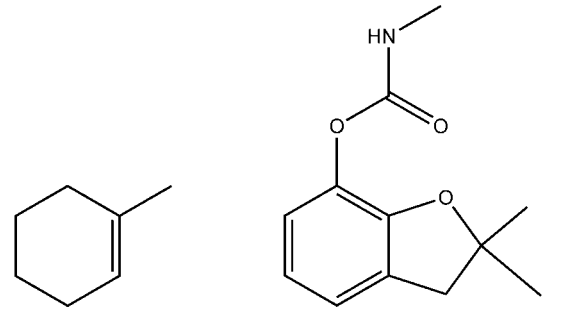


Figure 1. Representative molecules from the test sets. Shown are reference molecules with the typical number of heavy atoms in the previous (a) and this (b) test sets. 1-methylcyclohexene is shown in (a) and has 7 heavy atoms; 2,2-dimethyl-2,3-dihydrobenzofuran-7-yl methylcarbamate (also known as carbofuran) is shown in (b) and has 16 heavy atoms.

TABLE 1: Statistics for the Charge Models Tested in This Study[a]

charge model	rms error (kcal/mol)	R^2 (kcal/mol)	mean error (kcal/mol)
RESP	3.51 ± 0.20	0.76 ± 0.08	−1.68±0.42
AM1-BCC	3.82 ± 0.21	0.83 ± 0.09	−1.88 ± 0.45
MMFF	5.75 ± 0.20	0.60 ± 0.08	−3.92 ± 0.57
PM3BCC v0.2	4.13 ± 0.22	0.76 ± 0.09	−2.57 ± 0.44
PM3BCC v0.3	4.05 ± 0.21	0.80 ± 0.09	−2.47±0.43

[a] Shown are RMS error, correlation coefficient (R^2), and mean error.

3. Results and Discussion

A variety of previous explicit solvent hydration free energy studies had rms errors relative to experiment in the 0.8−1.6 kcal/mol range,[2,4,11] which might have implied similar accuracies here. However, the composition of this test set is very different. Prior test sets contained mainly monofunctional molecules with relatively standard or common functional groups. They were,

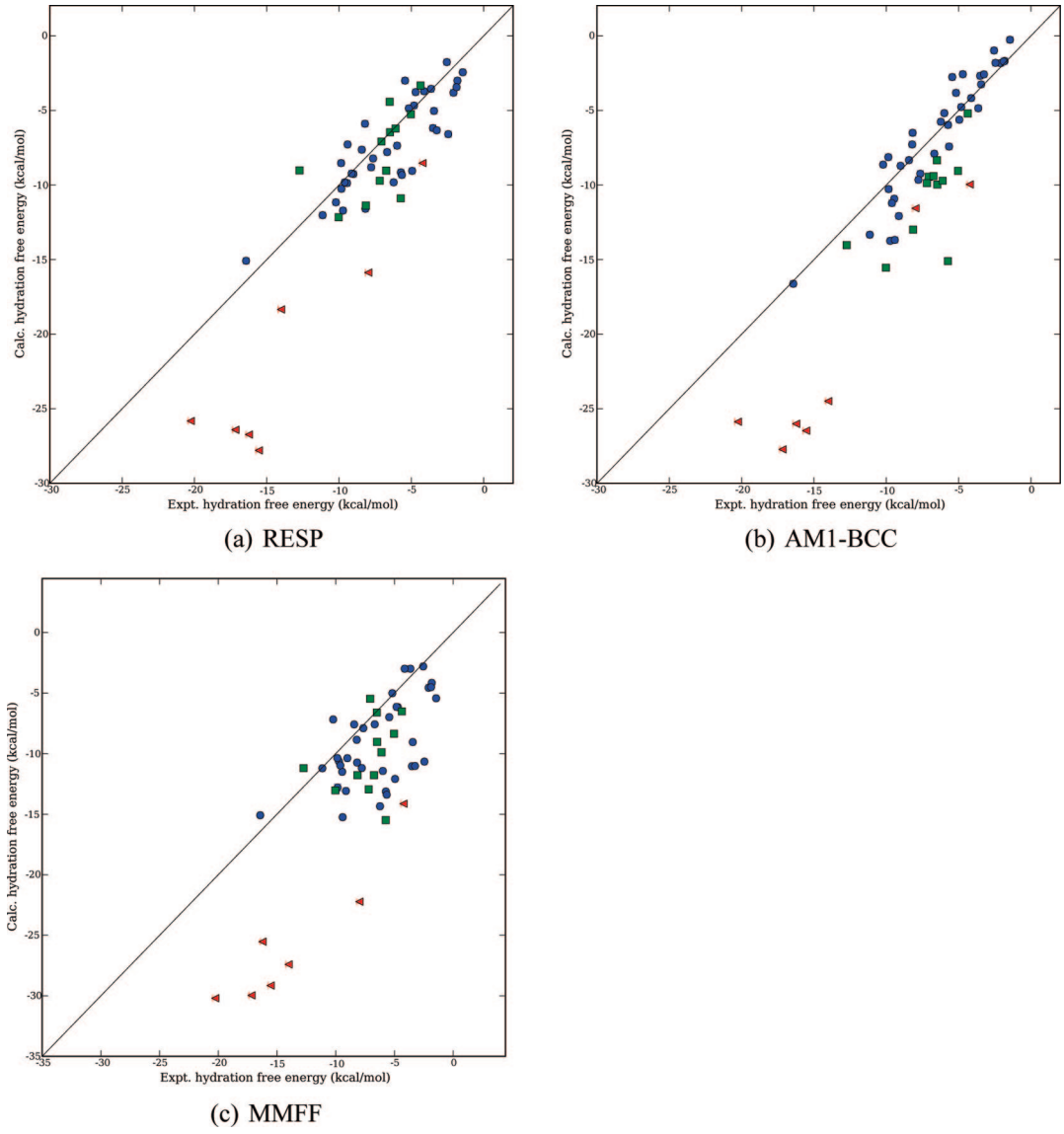


Figure 2. Computed versus experimental hydration free energies for the charge models studied. Shown are computed hydration free energies versus experiment, for partial charges from (a) RESP HF/6-31G*; (b) AM1-BCC; and (c) MMFF. Red triangles denote compounds containing hypervalent sulfur; green squares denote those containing hypervalent phosphorus, and blue circles denote the remainder of the compounds.

in some ways, unlike typical drugs because they lacked the polyfunctionality common in many drugs. On the other hand, the SAMPL set is much different, and in some respects more drug-like (many of the compounds are pesticides). Most of the SAMPL molecules are larger (16.3 heavy atoms on average, compared to 7.1 in a previous extensive test),[11] highly polyfunctional, and very polar (see the discussion and structures in ref 15). Also, a number of these functional groups have been rarely, if ever, studied with fixed-charge force fields. Finally, this test set takes us fairly far afield from the usual functional groups (such as amino acid side chain analogs, nucleic acids, and common cosolvents) which have been studied in developing the GAFF parameter set.[19,20] These factors make this SAMPL test much more challenging. To illustrate the difference, representative molecules are shown in Figure 1.

Here, we tried several different charge models with the same GAFF bonded and nonbonded parameters. We expected the RESP HF/6-31G*[26] and AM1-BCC[16,17] charge models would do fairly well, and as a negative control we used MMFF charges, which we expected to perform relatively poorly, as they were developed for a different force field. We also tested two other charge models to see how they compared against these others. Statistics are shown in Table 1. We use rms error and R^2, the correlation coefficient, as our metrics for quality, and also report mean error to show whether there is a systematic offset in computed values. We find that, as expected, MMFF performs worst. Both PM3-BCC charge models are intermediate in terms of rms error, and comparable to RESP and AM1-BCC in terms of R^2, and RESP and AM1-BCC have the lowest rms errors. Here, RESP has the lowest rms error, 3.5 $\pm$ 0.2 kcal/mol, and an R^2 value of 0.76 $\pm$ 0.08. Except for MMFF, rms errors fall between 3.5 and 4.1 kcal/mol, and R^2 values are decent, running from 0.76 $\pm$ 0.08 to 0.83 $\pm$ 0.09. Computed hydration free energies versus experiment are shown in Figures 2 and 3, and a full table of computed values and components is provided in the Supporting Information.

Overall, rms errors here are markedly higher than in previous studies,[2,4,11] probably reflecting the difficulty of this highly polar and polyfunctional test set, as well as its deviations from the regions of chemical space the force field has been tested in. Previous work on more typical functional groups showed that computed results for more polar compounds with more negative

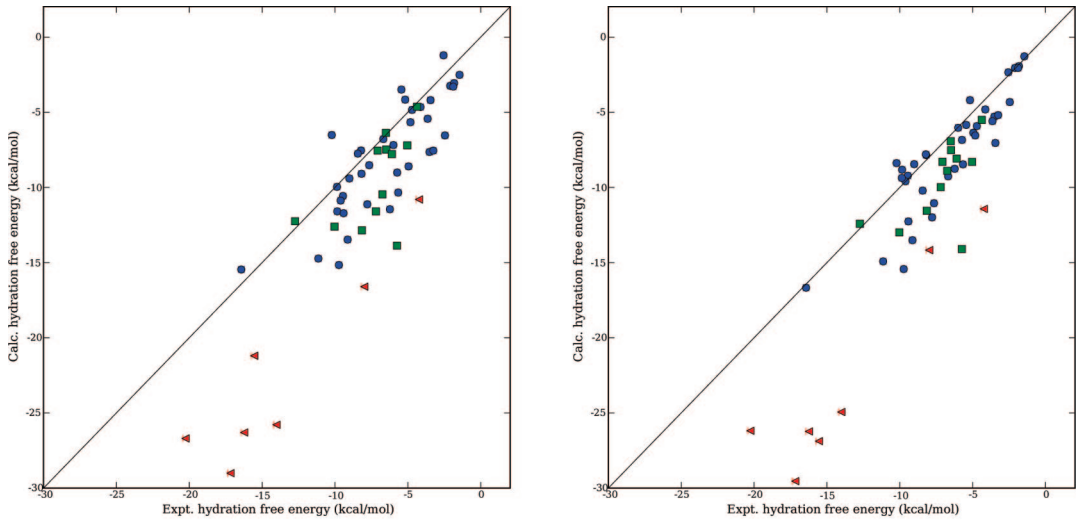


Figure 3. Computed versus experimental hydration free energies for the PM3-BCC charge models. Shown are computed hydration free energies versus experiment, for partial charges from (a) PM3-BCC v0.2 and (b) PM3-BCC v0.3. Red triangles denote compounds containing hypervalent sulfur; green squares denote those containing hypervalent phosphorus, and blue circles denote the remainder of the compounds.

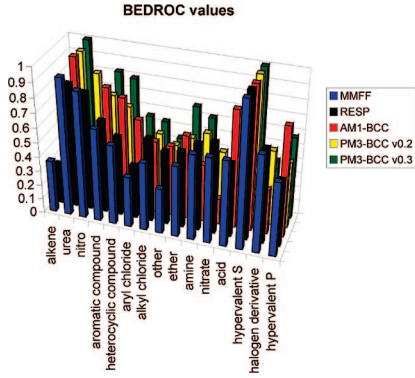


Figure 4. BEDROC values by functional group and charge model. Shown are computed BEDROC values ($\alpha = 1.0$) for different functional groups represented in the test set, for all of the charge models examined. A table of values, including uncertainties, is shown in the Supporting Information.

hydration free energies had larger errors;[11] this set has a higher proportion of highly polar compounds, which may have played a significant role in the lower accuracy here. Here, we group the compounds by functional group, sort the list by the magnitude of the error relative to experiment, and use the BEDROC metric to look for functional groups that are disproportionately associated with large errors. High BEDROC values mean a particular functional group occurs mostly in compounds with large errors, while low BEDROC values mean it occurs mostly in compounds with small errors, and intermediate values mean the functional group is distributed roughly randomly. Thus high BEDROC values may be an indication of force field errors for a particular functional group.

BEDROC values for the functional groups we examined are shown in Figure 4 by charge model. A random distribution (for these numbers of compounds) gives a BEDROC value of 0.49−0.50. Some functional groups show particularly significant deviations from random. BEDROC values for ureas, compounds with hypervalent sulfur, sulfonamides, and compounds with hypervalent phosphorus are all significantly worse than random with AM1-BCC. In contrast, nitrates are particularly well predicted. These trends are consistent across all the charge models, except that nitro-containing compounds also perform poorly in MMFF and PM3-BCC v0.2, and hypervalent phosphorus compounds are reasonably well predicted with PM3-BCC and RESP.

It seems clear that something is significantly wrong with the calculations or experiments for the compounds with hypervalent sulfur. Figure 2a shows the compounds with hypervalent sulfur for the AM1-BCC charge set, with those containing hypervalent sulfur and phosphorus highlighted with a different color and symbol. All of the computed values for these compounds are off from the experimental values in the same direction by several kcal/mol (mean error −8.10±0.40 kcal/mol with AM1-BCC). Hypervalent phosphorus compounds are also particularly poorly predicted with AM1-BCC and several other charge models (Figure 2a). Together, these two groups account for the worst outliers; if hypervalent sulfur and phosphorus compounds are excluded, the AM1-BCC rms error is 1.6 ± 0.2 kcal/mol (down from 3.8 ± −0.2) and the R^2 increases to 0.9 ± 0.1 (from 0.8 ± 0.1), more in line with the accuracies seen in previous studies.[2,11] Of course, excluding outliers always makes results better, and is only possible retrospectively.

Should we have known that hypervalent sulfur and phosphorus compounds might be a problem? Our previous retrospective study[11] only had five hypervalent sulfur compounds, and we find a BEDROC value of 0.6 ± 0.2, within uncertainty of random, so the data is inconclusive. The situation was even worse for hypervalent phosphorus compounds, for which there were only two representatives.

We believe this analysis suggests a systematic problem with force field parameters for hypervalent sulfur and possibly hypervalent phosphorus compounds that was statistically insignificant in earlier work, essentially due to the small number of such compounds in the earlier test set.

What might be the problem with these parameters? While AMBER[27] and GAFF[28] use a variety of atom types for sulfur and phosphorus, the Lennard-Jones parameters for all sulfur atom types are identical. Similarly, the Lennard-Jones parameters for all phosphorus atom types are identical. This seems surprising, as the chemical environment seems likely to affect the strength of dispersion interactions between these atoms and their surroundings. Another study recently found that the Lennard-Jones parameters for triple bonded carbons had been

taken from those for aromatic carbons in AMBER and GAFF and that this underestimated the attractive interactions between, for example, alkynes and water, leading to systematic errors.[4] Something similar could be going on here. Apparently the original AMBER sulfur parameters were taken from OPLS[27] (though the AMBER force field files indicate that free energy perturbation calculations also played a role), but now OPLS has moved to using different Lennard-Jones parameters for different sulfur atom types, while AMBER has maintained the single set of parameters. The (single set of) phosphorus parameters for AMBER were originally developed by Weiner et al.[29] and GAFF simply took this set and applied it to all the phosphorus atom types.[28] This practice of taking Lennard-Jones parameters derived for an element in one particular environment and applying it to the same element in a substantially different chemical environment, differing even in terms of the number of bonds, could be the cause of some of these systematic errors.

rms errors in the range of those reported here (3.5 kcal/mol and up) are large for practical applications. For example, a 3.5 kcal/mol error in a binding free energy calculation would be larger than the range in binding affinities in many lead series! In some sense, the compounds tested here are "drug-like", so this is at first a discouraging result. But as noted, rms errors are much better (and more in line with previous studies) if the hypervalent sulfur and phosphorus compounds are excluded, so the poor accuracy seen here may be simply pointing the way toward the need for refinements of the force field for these particular compounds.

We are not aware of any force fields or methods that would be expected to give better results for this set. Other participants in the SAMPL challenge seemed to achieve at best comparable results.[15] The same held true in another recent prospective test,[2] where the best Poisson−Boltzmann based approach gave accuracies no better than molecular dynamics free energy calculations. Even a later retrospective study using a quantum mechanical continuum solvation model gave accuracies that were roughly comparable (rms errors between 1.08 and 1.88 kcal/mol,[30] versus rms errors between 1.33 and 2.0 kcal/mol with explicit solvent in the previous study)[2]. So, from a methods point of view, there is no reason to expect that other methods should perform any better on this set. However, if the dominant source of error here is indeed a parameter problem for a small subset of the compounds, it might suggest that a method more grounded in quantum mechanics might do substantially better here.

4. Conclusions

Opportunities for prospective or blind tests of computational free energy methods have been relatively rare. This represents the second such test for calculations of hydration free energies.[2] These tests are helpful, as they provide a way to avoid any possibility of being influenced by knowledge of the "right answer" and thus to genuinely test the method with no adjustments to parameters.

Overall, this prospective test has provided an opportunity to test explicit solvent simulations in a region of chemical space in which they have been rarely applied. Stepping into the "wilderness" in this way appears to present some risks, as errors were larger here than in previous studies considering simpler, often monofunctional, compounds that were more similar to typical protein or nucleic acid components. It was encouraging that correlations with experimental values remained fairly strong (R^2 of 0.75 and higher, except for our negative control charge model), though errors were relatively high. Some functional groups were particularly poorly predicted, suggesting that further force field development for these functional groups may improve accuracies. We believe that regular studies of this nature will provide substantial benefits for the development of solvation models and force fields, and will aid in identifying systematic errors with force fields and making improvements.

Acknowledgment. We are grateful to J. Peter Guthrie (University of Western Ontario) for curating and providing the experimental data, and we appreciate OpenEye Software for organizing SAMPL; particular thanks go to Geoff Skillman and Anthony Nicholls (OpenEye). We appreciate the support of NIH Grant GM 63592 to K.A.D.

Supporting Information Available: A full table of all of the computed hydration free energies and components, for all charge models, along with the experimental values; computed BEDROC values and uncertainties by functional group and charge model; AMBER parameter and coordinate files for the molecules in the set. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740–5761.
(2) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Pande, V. S. *J. Med. Chem.* **2008**, *51*, 769–778.
(3) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2008**, *112*, 938–946.
(4) Mobley, D. L.; Dumont, È.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.
(5) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. *J. Am. Chem. Soc.* **1999**, *121*, 9243–9244.
(6) Chorny, I.; Dill, K. A.; Jacobson, M. P. *J. Phys. Chem. B* **2005**, *109*, 24056–24060.
(7) Mobley, D. L.; Barber, A. E.; Fennell, C. J.; Dill, K. A *J. Phys. Chem. B.* **2008**, *112*, 2405–2414.
(8) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.
(9) Hess, B.; van der Vegt, N. F. A. *J. Phys. Chem. B* **2006**, *110*, 17616–17626.
(10) Deng, Y.; Roux, B. *J. Chem. Phys.* **2004**, *108*, 16567–16576.
(11) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. **2008**, submitted.
(12) Villa, A.; Mark, A. E. *J. Comput. Chem.* **2002**, *23*, 548–553.
(13) Xu, Z.; Luo, H. H.; Tieleman, D. P. *J. Comput. Chem.* **2007**, *28*, 689–697.
(14) Maccallum, J. L.; Tieleman, D. P. *J. Comput. Chem.* **2003**, *24*, 1930–5.
(15) Guthrie, J. P. *J. Phys. Chem. B* **2009**, accepted.
(16) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.
(17) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.
(18) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
(19) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
(20) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Model.* **2006**, *26*, 247260.
(21) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
(22) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529–539.
(23) Bennett, C. H. *J. Comp. Phys.* **1976**, *22*, 245–268.
(24) Haider, N."Checkmol", http://merian.pch.univie.ac.at/ nhaider/cheminf/cmmm.html.
(25) Truchon, J.-F.; Bayly, C. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
(26) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
(27) Cornell, W.; Cieplak, P.; Bayly, C. I.; Gould, I. R., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
(28) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
(29) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230–252.
(30) Chamberlin, A. C.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2008**, *112*, 8651–8655.

JP806838B

Structure
Review

Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get"

David L. Mobley[1,*] and Ken A. Dill[2]
[1]Department of Chemistry, University of New Orleans, New Orleans, LA 70148, USA
[2]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94158-2517, USA
*Correspondence: dmobley@gmail.com
DOI 10.1016/j.str.2009.02.010

We review insights from computational studies of affinities of ligands binding to proteins. The power of structural biology is in translating knowledge of protein structures into insights about their forces, binding, and mechanisms. However, the complementary power of computer modeling is in showing "the rest of the story" (i.e., how motions and ensembles and alternative conformers and the entropies and forces that cannot be seen in single molecular structures also contribute to binding affinities). Upon binding to a protein, a ligand can bind in multiple orientations; the protein or ligand can be deformed by the binding event; waters, ions, or cofactors can have unexpected involvement; and conformational or solvation entropies can sometimes play large and otherwise unpredictable roles. Computer modeling is helping to elucidate these factors.

Introduction: Computer Modeling Is an Important Tool for Understanding Ligand Binding to Proteins

Structure-based computer modeling of ligand-protein interactions is now a core component of modern drug discovery (Charifson and Kuntz, 1997). It is now difficult to imagine the drug discovery process without computation (Jorgensen, 2004). Computational methods have played a key role in the drug discovery process for a growing number of marketed drugs, including HIV protease inhibitors (Charifson and Kuntz, 1997; Greer et al., 1994; Jorgensen, 2004) and zanamivir (an antiviral neuraminidase inhibitor) (von Itzstein et al., 1993), and in the development of new drug candidates, such as HIV integrase inhibitors (Hazuda et al., 2004; Schames et al., 2004), hepatitis C protease inhibitors (Liverton et al., 2008; Thomson and Perni, 2006), and beta-secretase inhibitors (BACE-1) (Stauffer et al., 2007).

An early step in this field was the invention of the DOCK method in 1982 (Kuntz et al., 1982). There are now at least four classes of physical computer methods (listed from fastest to slowest, and least physical to most physical): (1) very fast molecular docking methods, including DOCK, Glide, AutoDock, FlexX, ICN, PMF, and GOLD, (2) approximate free energy methods, in which the solvent and protein motions are taken into account with fewer approximations, (3) relative binding free energy (RBFE) methods, which include full solvent and protein motions, but which require prior knowledge of the structure of a complex of the protein with a ligand that is similar to the one of interest, and (4) absolute binding free energy (ABFE) methods, which are the most expensive computationally, but which include the physics in the most rigorous way that is currently practical (see Figure 1). ABFE methods start from an unbound ligand and potentially the unbound structure of the protein to attempt to predict the structures, affinities, and thermal properties of the complexes of interest. Mining minima is another method that is very nearly in this last category and has provided insight into binding (Chang and Gilson, 2004; Gilson and Zhou, 2007; Head et al., 1997).

Different Computer Methods Trade Off Speed versus Physical Accuracy

First, we define some terms. A lead compound is a molecule, typically in early-stage drug discovery, that can be further chemically modified to improve its properties as a possible drug candidate. A complex is a receptor and ligand bound together. A pose is one conformation of a ligand in a complex and specifies both the ligand conformation and its position relative to the receptor. A pose can refer either to a conformation that is known from an experimental structure of a complex, or to a hypothetical conformation generated in a computer model. The apo form refers to the structure of the protein that has no ligand bound to it. The holo form refers to the structure of the protein when it is complexed with ligand. The binding free energy, ΔG^o, is the free energy of the complex minus the free energies of the ligand and apo protein separately in aqueous solution. The binding free energy is related to the equilibrium association constant, K_a, (in units of M^{-1}) by $\Delta G^o = -RT \ln (C^o K_a)$, where R is the gas constant, T is the absolute temperature, and C^o is the standard concentration (1 M). The binding affinity, or dissociation constant, equals $1/K_a$. The binding free has two components, $\Delta G = \Delta H - T\Delta S$, where H is the enthalpy and S is the entropy. Here are some of the key approaches used for studying binding.

Docking

Docking methods start with a known protein structure and a known ligand structure and aim to rapidly generate an optimal protein-ligand bound conformation. Docking was designed to be very rapid (seconds or less per compound), which is desirable for screening large libraries in the short times required for modern pharmaceutical lead discovery. Docking explores many ligand conformations and orientations, and in some cases even different potential binding sites. The different poses are rank ordered by a score, a quantity that ideally would correlate with the free energy of binding, and is obtained either from a physical or knowledge-based potential. Often, docking approaches treat the protein as completely rigid, having a single fixed receptor conformation. Other docking methods treat protein motions by

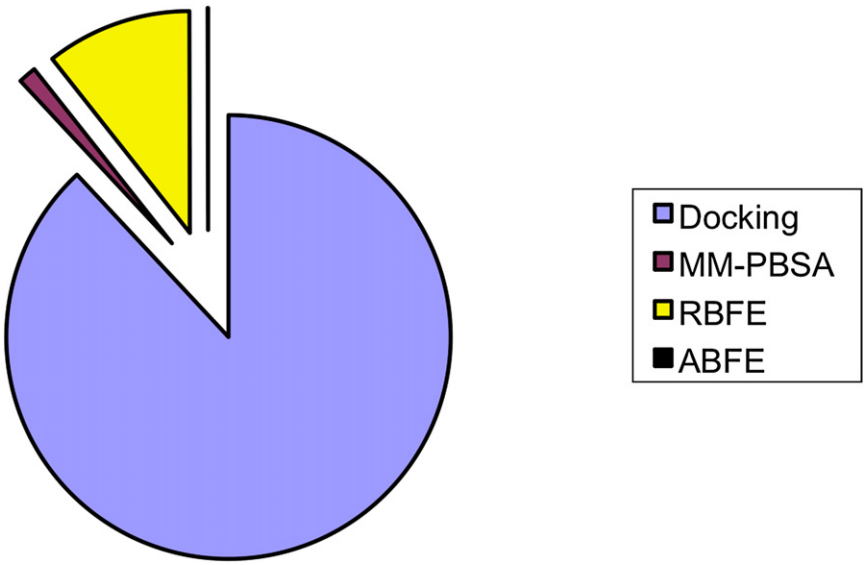


Figure 1. Relative Publication Numbers for Different Computational Methods
From Google Scholar (July 2008). MM-PBS, molecular-mechanics with Poisson-Boltzmann surface area; RBFE, relative binding free energy; ABFE, absolute binding free energy. Percentages, in the order listed in the legend, are 88%, 1%, 11%, and 0.04%.

moving only certain atoms out of the way. Though some modern docking approaches can allow for some motions of side chains or backbone (Corbeil et al., 2007; Cozzini et al., 2008; Leach, 1994; Meiler and Baker, 2006; Sherman et al., 2006; Wei et al., 2004), treating these degrees of freedom slows down the computations considerably. Docking is an appealing way to generate leads (Shoichet et al., 2002) because of its speed and ability to screen large libraries of potential leads (Huang and Jacobson, 2007; Babaoglu et al., 2008). But because docking trades off physical accuracy for speed, it is seldom accurate enough to predict binding affinities or rank-order compounds. Its power to discriminate binders from nonbinders varies widely depending on the target protein (Graves et al., 2008; Warren et al., 2006). But because of its speed, docking approaches are the method of choice for filtering out compounds that are likely nonbinders and for identifying native-like poses.

MM-PBSA/GBSA

MM-PBSA/GBSA is more physically rigorous than docking. The acronym stands for molecular mechanics with Poisson-Boltzmann + surface area or MM-GBSA (GB stands for Generalized Born), and the method was originated by the Kollman and Case labs in the late 1990s (Cheatham et al., 1998; Kollman et al., 2000; Srinivasan et al., 1998), with parallel work by others (Vorobjev and Hermans, 1999). It involves greater computational cost than docking (at least several hours per compound), but also is more physical in its more extensive conformational sampling. MM-PBSA aims to estimate the binding free energies, or relative binding free energies, of related compounds. Here, a computer generates representative bound and unbound structures by molecular mechanics simulations or by energy minimization of a protein-ligand complex, usually in explicit solvent. The goal is to estimate the change in enthalpy on binding by comparing the average enthalpy of bound and unbound states, but this would be a small difference of two large, noisy energies. So after the all-atom simulations, the water is removed and the enthalpies and binding free energies are estimated using an implicit (Poisson-Boltzmann or Generalized Born) representation of water. The binding free energy estimate includes the enthalpy change plus the change in salvation free energy from the implicit solvent model.

In many cases, an approximate value of the entropy is also estimated from these simulations. Because MM-PBSA/GBSA invests more effort in sampling and entropies, it is closer to a true free energy calculation. However, often, because of limitations in the approximations for estimating entropy (Gilson and Zhou, 2007), entropic contributions are omitted when estimating relative binding strengths, in the hope that these contributions will cancel when comparing similar ligands (Gilson and Zhou, 2007; Shirts et al., 2009).

Early results with the MM-PBSA method were quite promising (typically with mean-squared errors under 3 kcal/mol for the first several years) (Huo et al., 2002; Kuhn and Kollman, 2000; Mardis et al., 2001; Rizzo et al., 2004; Schwarzl et al., 2002; Shirts et al., 2009), but more recent studies have seen larger errors in some cases (Shirts et al., 2009). Applications have typically been limited to single targets, so it is difficult to evaluate how well the method does generally.

The drawbacks of MM-PBSA/GBSA are that it, too, is sometimes not predictive (Pearlman, 2005; Shirts et al., 2009; Steinbrecher et al., 2006) and it requires prior knowledge of a likely bound complex as a starting point, although such starting conformations can be taken from prior docking (Steinbrecher et al., 2006).

Relative Binding Free Energies

A still more rigorous approach uses the energetics of a physical force field and extensive conformational sampling from molecular dynamics simulations to actually compute differences in binding free energies between similar ligands. This can be done using computational alchemy to obtain the difference in binding free energies, $\Delta\Delta G_{A\rightarrow B}$. This is the free energy of changing ligand A into ligand B in the receptor, minus the free energy of changing A into B in solution. To compute this free energy difference for just one pair of ligands binding to the same protein can cost several hundred CPU days. These relative free energies can be computed precisely—given sufficiently long molecular dynamics simulations—using one of several different analysis techniques (Shirts et al., 2007). Though the accuracy of the binding free energies obtained from this method depends on the accuracy of the underlying molecular mechanics force field, it does treat fully, at least in principle, free energies associated with conformational change as well as entropies.

The first alchemical calculations were performed in the 1980s in the McCammon lab (Tembe and McCammon, 1984; Wong

and McCammon, 1986), and then by others (Hermans and Subramaniam, 1986; Warshel et al., 1986; Bash et al., 1987; Shirts et al., 2007). Limitations of these methods are the high computational costs and the need to know at least one bound structure of a similar ligand in the protein as a starting point. Accuracies are generally better than for MM-PBSA (Pearlman, 2005; Steinbrecher et al., 2006) and docking (Mobley et al., 2007b; Pearlman and Charifson, 2001), but few systematic comparisons have been done. These methods are only useful for comparing related ligands or receptors.

Absolute Binding Free Energies

The most powerful approach, in principle, is the method of absolute binding free energies (ABFE) (Boresch et al., 2003; Hermans and Wang, 1997; Roux et al., 1996). Like RBFE methods, ABFE methods also use full molecular dynamics simulations with fully detailed atomic force fields, and also involve separate sets of simulations for the solvated ligand, the solvated protein, and the complex. But ABFE methods do not require prior knowledge of the binding affinity of a related ligand, hence the term *absolute*. There have been two groups of ABFE approaches. The first begins with the structure of the ligand of interest bound to the protein. However, the ultimate goal is to begin with no prior knowledge of either the structure or affinity of the ligand complex. A second, more recent group replaces starting knowledge of the structure with one or more docking poses (Mobley et al., 2006, 2007b; Jayachandran et al., 2006). Various studies suggest that ABFE methods are fairly accurate, with good correlations to experimental binding affinities and with RMS errors often less than 3 kcal/mol (Deng and Roux, 2006; Fujitani et al., 2005; Jayachandran et al., 2006; Mobley et al., 2007b; Shirts et al., 2007; Wang et al., 2006), and sometimes much better.

Ligand Binding Is Described by Energy Landscapes, Not Just Single Structures

The enterprise of structural biology has given us powerful "eyes" to see single structures—specific native structures and specific bound complexes—and some of the driving forces that hold them together: hydrogen bonds, hydrophobic interactions, ion pairing, and van der Waals packing. However, "what you see" is not always "what you get." Other equally important forces, namely the entropies, are not visible in native structures.

To capture both the observable and nonobservable contributions to the energetics, it is important to note that binding takes place on an energy landscape. Exploring energy landscapes often requires modeling and computer simulations. For binding, the energy landscape is the free energy of the system as a function of its degrees of freedom, which are many, and include translational, rotational, conformational, and solvation degrees of freedom.

Upon Binding, a Ligand Loses Translational and Rotational Entropy

Relative to a receptor, a ligand has three translational degrees of freedom (x, y, and z directions) and three orientational degrees of freedom. When bound, motion in these degrees of freedom becomes restricted. This loss of freedom results in an entropic and free energy cost, opposing binding and favoring the dissociated state (Chang et al., 2007; Chang and Gilson, 2004; Chen et al., 2004; Deng and Roux, 2006; Lee and Olson, 2006; Wang et al., 2006). The loss of freedom depends on the mobility

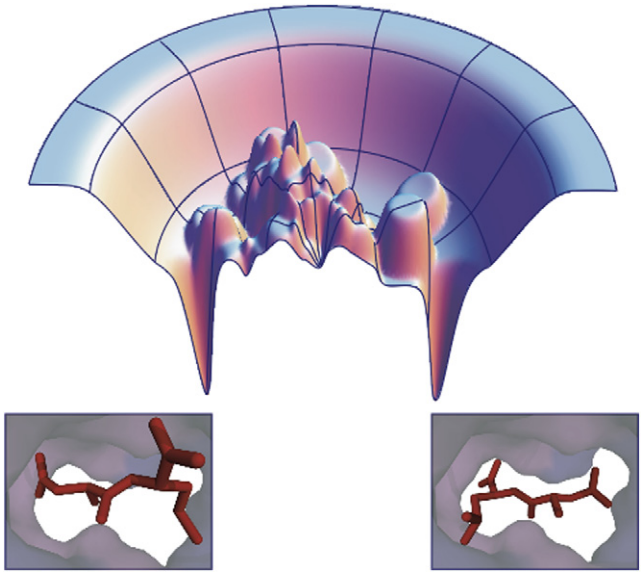


Figure 2. Hypothetical Ligand Binding Energy Landscape
Ligand binding energy landscapes (top) can be rough, with multiple minima. These multiple minima can correspond to multiple distinct ligand binding conformations in the receptor (bottom).

remaining in the binding site, so that in a series of increasingly tightly bound structures there will be increasing losses in translational and rotational entropies, resulting in a contribution opposing binding.

Upon Binding, a Ligand Can Lose Internal Freedom and Entropy

Some simple models assume that the ligand entropy lost on binding correlates with the number of rotatable bonds in the ligand (Böhm, 1993; Gilson and Zhou, 2007; Huey et al., 2007; Laederach and Reilly, 2003; Taylor et al., 2002; Gohlke and Klebe, 2002). The ligand is envisioned to start in its unbound state having access to all possible conformers and to end in its bound state having a single conformer. However, interestingly, more rigorous recent computational studies in host-guest systems indicate that losses in ligand conformational entropy on binding are not strongly correlated with the number of rotatable bonds (Chang and Gilson, 2004; Chen et al., 2004; Guimaraes and Cardozo, 2008). Recent work on salvation free energies of small molecules has led to similar conclusions (Mobley et al., 2008). Not all small-molecule conformers are populated equally in solution. Thus computing accurate ligand affinities (and entropy losses) requires more accurate treatments of the different ligand populations in solution. Entropic contributions can also vary between different conformations of the same ligand in a particular receptor (Chen et al., 2004; Gilson and Zhou, 2007), which may be important even for docking (Ruvinsky and Kozintsev, 2005).

A Ligand Can Bind to a Receptor in Different Poses

A ligand can sometimes adopt multiple different conformations or orientations upon binding (see Figure 2). These different poses can be separated by energetic barriers. In some cases the different poses are due to ligand or protein symmetries. HIV-1 protease is a dimer with a nearly symmetric active site; as a result, many HIV protease inhibitors have two nearly identically

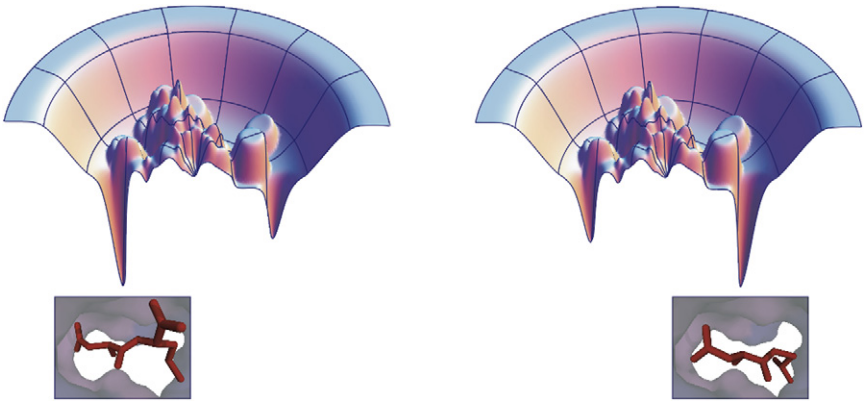


Figure 3. Small Changes in a Ligand May Modify the Binding Landscape
It is not uncommon to find that small modifications in a ligand (bottom) may lead to drastic changes in the observed binding mode (bottom) (Stout et al., 1999; Badger et al., 1988; Böhm and Klebe, 1996; Kim, 2007a, 2007b; Pei et al., 2006; Reich et al., 1995; Stoll et al., 2003). This can be explained by an energy landscape with multiple minima (top), which is altered slightly by minor modifications to the ligand (top left versus top right), leading to a substantial change in the binding mode.

binding modes (e.g., see Protein Data Bank [PDB] codes 1AXA, 1IZH, 1MUI, and 1U8G). Ligand symmetries can lead to trivial cases of multiple binding modes, which have significant entropic implications. Multiple binding modes are observed also when symmetries do not play a role. Computational studies show multiple distinct ligand binding modes in binding sites in T4 lysozyme (Mobley et al., 2006, 2007b), neutrophil elastase (Steinbrecher et al., 2006), estrogen receptor inhibitors (Oostenbrink and van Gunsteren, 2004), FKBP inhibitors (Jayachandran et al., 2006), biotin and streptavidin (Lazaridis et al., 2002), and cytochrome P450cam (Paulsen and Ornstein, 1992).

Do experimental studies support these predictions of multiple ligand conformations? The challenge is that multiple conformers are difficult to determine experimentally. But there is at least some direct crystallographic evidence suggesting multiple relevant orientations: in T4 lysozyme (Graves et al., 2005; Mobley et al., 2006, 2007b), influenza neuraminidase (Stoll et al., 2003) and possibly in trypsin (Stubbs et al., 2002), where the binding mode is affected by pH. Multiple binding modes of fragment-like kinase inhibitors have also been observed (Constantine et al., 2008). Multiple orientations or binding modes have also been seen in thymidylate synthase (Montfort et al., 1990), in the binding of an HIV-1 cell entry inhibitor to the core of HIV-1 gp41 (Zhou et al., 2000), the binding of a transition state analog to an AmpC beta lactamase mutant (Chen et al., 2006), the binding of thiocamphor to cytochrome P450cam (Raag and Poulos, 1991), the binding of flavin to *para* hydroxybenzoate hydrolase (Gatti et al., 1994), and in the binding of some HIV-1 protease inhibitors (Murthy et al., 1992). There is additional evidence for multiple orientations in several other cases (Birdsall et al., 1989; Böhm and Klebe, 1996; Lazaridis et al., 2002; Mewshaw et al., 2005; Orville et al., 1997; Uytterhoeven et al., 2002). Spectroscopic data (Deng et al., 2001) and studies of drastically different binding modes of related inhibitors (Figure 3) (Stout et al., 1999; Badger et al., 1988; Böhm and Klebe, 1996; Kim, 2007a, 2007b; Pei et al., 2006; Reich et al., 1995; Stoll et al., 2003), some of which may have multiple binding modes (Montfort et al., 1990), suggest that multiple binding modes may be relatively common.

Proteins Wiggle, and May Have Multiple Conformers in Both the Bound and Unbound States

It is not only ligands that can have multiple binding modes. Proteins can, too (Mobley et al., 2007a, 2007b). We refer here not to induced fit, where the ligand binding event causes a change in protein conformation. Rather, we focus on internal motions or freedom of the protein that occur either in the apo structure of the protein itself or in the complex itself. Comparisons of different apo structures of the same proteins show that there are some rotamer changes near binding sites even in the absence of the ligand (Najmanovich et al., 2000), suggesting that multiple rotameric states may be relevant; this is also supported by NMR data (Chou and Bax, 2001). Structural data in the apolar T4 lysozyme binding cavity suggests that helix F, which borders on the binding cavity, can undergo substantial motions of 1.5–2.5 Å with little free energy cost (Morton and Matthews, 1995); various other motions occur in T4 lysozyme as well (Zhang et al., 1995). DHFR appears to have multiple relevant conformations, both in isolation and when binding ligands, and the populations are modulated by pH (Birdsall et al., 1989); each state in the catalytic cycle appears to have at least partially occupied conformations that resemble those before or after it in the cycle (Boehr et al., 2006). Crystallographic evidence suggests multiple protein conformations due to domain motion in some cases (Ma et al., 2002). Multiple conformers are also seen in host-guest binding (Chang and Gilson, 2004; Chen et al., 2004) and can be critical for protein mechanisms such as enzyme catalytic motions (Eisenmesser et al., 2005; Arora and Brooks, 2007; Henzler-Wildman et al., 2007; Henzler-Wildman and Kern, 2007). Several other studies also have provided evidence for multiple protein conformations (Eisenmesser et al., 2002; Gerstein et al., 1994; Min et al., 2005a, 2005b; Fragai et al., 2006).

Strain: A Measure of the Free Energy of Deformation

Ligand binding to a protein may induce strain. Strain refers to an energy cost, usually associated with a deformation of some sort. To achieve the lowest free energy of binding in the complex, the protein and/or ligand may become deformed relative to its unbound state in solvent, which costs energy (strain). Computational studies in the apolar lysozyme model binding site have found that protein strain energies for a valine side-chain rotamer change can be 3–4 kcal/mol (Deng and Roux, 2006; Mobley et al., 2007a, 2007b). When such strain energies are not taken into account, it leads to errors in predicted binding free energies (Mobley et al., 2007a, 2007b).

Computational modeling suggests that ligand strain free energies can be significant. In a survey of 150 crystallographic protein-ligand complexes, Perola and Charifson (2004) used molecular mechanics scoring functions to assess strain energies

and found that roughly 50% of ligands with 4–6 rotatable bonds had strain energies more than 3 kcal/mol, and, overall, 40% of ligands had strain energies more than 5 kcal/mol. Another study computed quantum mechanical torsional potentials for a variety of PDB ligands and found that typical strain energies could be on the order of 0.6 kcal/mol per torsion motif, amounting to roughly 3 kcal/mol for a ligand with five torsion motifs (Hao et al., 2007). Another study found, for several ligands, that free energy costs of restricting the ligand to the bound conformation could be a few kcal/mol (Tirado-Rives and Jorgensen, 2006). More recent work suggests that these values could be overestimates of the true strain, as crystal structures (from which strain is estimated) may be refined with a different force field than is used in estimating the strain, introducing artifacts. Nevertheless, strain energies often appear to be greater than several *kT* (Huang et al., 2006). See also Warren and Perola's (Warren and Perola, 2008) presentation on the topic from OpenEye's CUP meeting (http://www.eyesopen.com/about/events/cups-2008/pdfs-CUP/CUP9-Field-of-Extremes.pdf). Apparently, binding interactions can be strong enough to pay a substantial strain price for deforming one or both partners. Hence the true bound structure of a complex will not be the one that maximizes the interaction energy between the receptor and ligand, but rather the one that best balances the tradeoff between gaining additional favorable interactions while also inducing strain (Sharp, 2005).

Some experiments support this contention that strain free energies can be substantial. In an NMR study on maltose binding protein, Tang et al. (2007) found that the unbound protein was predominantly (roughly 95%) in the open apo conformation, and had a smaller (roughly 5%) population in a minor apo conformation that was more like the holo conformation, but with no evidence that it populated the holo conformation at all in the absence of the ligand. Thus the minor apo conformation is roughly 1.7 kcal/mol less favorable than the major apo conformation, and the holo conformation is probably still more unfavorable. In another instance, in $NtrC^r$, a conformational switch in bacteria that undergoes a conformational change upon phosphorylation, the active conformation has been shown to be partially populated even when the protein is unphosphorylated (Volkman et al., 2001), but with a smaller population. Based on the populations, this active-like conformation is about 2 kcal/mol less favorable than the norm in active conformation. So, functional protein conformational changes can make significant contributions to the thermodynamics.

Ligand Binding Can Cause Conformational Change in Protein Structures

When a ligand binds to a protein, it causes conformational changes in the protein. This may or may not be accompanied by strain in the protein, as strain is an (invisible) energy cost.

Ligand-induced protein conformational changes are not rare events. Comparisons of apo and holo structures from the PDB show that backbone conformational motions on ligand binding are relatively common; 20% of binding residues (Gutteridge and Thornton, 2005) and 25% of binding sites (Najmanovich et al., 2000) across a variety of proteins have backbone C_α motions more than 1 Å. And 15% of binding site residues have side-chain motions of more than 2 Å (Gutteridge and Thornton, 2005), whereas only 30%–40% of binding sites have been shown to undergo no side-chain rotamer changes (Najmanovich et al., 2000). More anecdotal reports of conformational changes on ligand binding are available for a wide range of systems; kinases, for example, are notoriously flexible (Noble et al., 2004; Vajpai et al., 2008; Weisberg et al., 2007), as are many other proteins (Böhm and Klebe, 1996; Kim, 2007a; Meiler and Baker, 2006; Teague, 2003). An extreme example may be natively disordered proteins in which large parts of the protein may become ordered upon interacting with binding partners (Hilser and Thompson, 2007; Radivojac et al., 2007; Wright and Dyson, 1999).

In addition, a given protein can adopt different conformations for different ligands. A PDB study of 206 binding site pairs (each pair consisting of two structures of the same protein with different, similar ligands in the binding site) showed that in 83% of the cases there were significant conformational changes in the binding sites between pair members (Bostrom et al., 2006); changes were judged significant if the RMSD for all side-chain atoms if at least one amino acid residue within 5 Å of the ligand is greater than 1.0 Å. The most frequent differences were changes in water architecture and side-chain conformation (both occurring in over 50% of the pairs). Significant backbone conformational changes occurred in only 7% of the set; changes were judged significant if the RMSD of at least one backbone heavy atom in three or more consecutive amino acids is more than 0.5 Å. A smaller study found examples of substantial conformational changes on binding similar ligands for a variety of systems as well (Kim, 2007a). It is even possible for a single ligand to bind to different protein conformations under different solution conditions (Miller and Dill, 1997). Thus, changes in binding site architecture, at least at the side-chain and water level, should be regarded as the rule, rather than the exception.

Small Changes in Conformation Can Cause Big Changes in Binding Affinities

Some computational studies predict that even when a binding site structure is not perturbed very much, its energetics can change substantially. For example, simulations show that neglecting even small protein motions can lead to large errors (RMS errors relative to experiment of nearly 20 kcal/mol when protein motions are not allowed), relative to much smaller errors (1.7 kcal/mol RMS) when protein motion is allowed. Even small relaxations of the protein reduced the RMS errors to 4–5 kcal/mol) (Mobley et al., 2007b). This is important for both conceptual and practical reasons. Conceptually, it means the strength or quality of binding interactions is sensitive to minute details of the bound structure and is not easily assessed by simple metrics like hydrogen bond counts or apparent fit. Practically, it means that free energy methods that include these protein conformational changes can potentially have much higher accuracy than docking methods that neglect them.

The conclusion that these small changes can make big differences in the energetics is supported by a variety of docking and (re)scoring studies that have looked at the effects of introducing small amounts of protein flexibility. Though scores do not necessarily improve for all ligands, they do typically change substantially, showing some improvement (Graves et al., 2008; Huang et al., 2006; Meiler and Baker, 2006; Sousa et al., 2006; Wei et al., 2004). But introducing protein flexibility without accounting for protein strain energies can potentially increase false positive rates by making binding sites too permissive (Graves et al., 2008;

Sousa et al., 2006; Wei et al., 2004). This likely highlights the role of conformational change and strain.

Differences in Solvation Can Contribute to Binding Affinities

Several detailed binding free energy studies have suggested that differences in solvation may play an important role in differences in binding free energy between relatively similar compounds (Jiao et al., 2008; Reddy and Erion, 2001). Two molecules might have similar interactions with a protein, similar strain energies, etc., but have different solvation properties in water, leading to solvation-driven differences in binding free energies. These differences may not always be intuitive. For example, the N-methylacetamide/amine "problem" (Rizzo and Jorgensen, 1999) suggests that adding a hydrophobic methyl group to acetamide or ammonia increases the affinity for water, whereas subsequent methylations decrease the affinity.

The importance of solvation and desolvation is supported by an emerging trend toward including approximate estimates of solvation/desolvation energies in approximate docking methods for scoring protein-ligand binding. Including such estimates appears to result in improved scoring (Ferrara et al., 2004; Gilson and Zhou, 2007;Shoichet et al., 1999). Without these contributions, charged ligands can wrongly appear to bind better than polar ligands in a polar binding site. A charged ligand may make favorable electrostatic interactions in a polar binding site, but it also costs a huge amount of energy to remove it from water (Brenk et al., 2006; Gilson and Zhou, 2007; Shoichet et al., 1999). In other cases, a small modification to a ligand can potentially lead to affinity gains due to a change in the desolvation cost (Kangas and Tidor, 2001).

Bound Waters Usually Contribute Favorably to Ligand Binding, But Not Always, and Their Contributions Are Highly Variable

Computer simulations have been used to study the role of crystallographic waters in binding thermodynamics (Barillari et al., 2007; Hamelberg and McCammon, 2004; Lu et al., 2006; Olano and Rick, 2004; Zhang and Hermans, 1996; see also Helms and Wade [1995, 1998a, 1998b] for desolvation of a buried binding cavity). In many cases, binding or ordering of waters occurs concurrently with ligand binding, so it can be extremely difficult to experimentally assess the contribution of water binding to overall binding thermodynamics. Computational methods can directly compute the free energy of inserting or removing a water molecule from a binding site, providing key insight that is hard to obtain experimentally.

These computational studies indicate that bound waters contribute substantially to binding free energies, contributing as much as −10 kcal/mol for some waters (Barillari et al., 2007), but smaller values between −3 and −6 kcal/mol are more typical (Barillari et al., 2007; Hamelberg and McCammon, 2004; Lu et al., 2006; Olano and Rick, 2004; Zhang and Hermans, 1996). In some cases, crystallographic waters appear substantially unfavorable relative to bulk, raising the possibility of problems with refinement or force fields (Barillari et al., 2007; Olano and Rick, 2004). Perhaps ligands can be designed with improved affinities by recognizing nearby sites where waters can be easily displaced (Abel et al., 2008; Pan et al., 2007).

Sometimes ligand binding can involve concerted reordering of many water molecules. In some hydrophobic sites in proteins that bind fatty acids or lipids, whole networks of more than a half-dozen water molecules shift their structures to form a "hydrophobic" interface with the ligand (LaLonde et al., 1994; Sulsky et al., 2007).

Protonation States Can Change on Binding, Influencing Affinity

Binding free energies can also be affected by other unseen and unexpected factors. For example, protonation states can change on binding (Czodrowski et al., 2007; Dullweber et al., 2001; Gohlke and Klebe, 2002; Steuber et al., 2007), as can tautomeric states (Pospisil et al., 2003) and other factors. In some cases, multiple protonation or tautomeric states can be relevant, as observed crystallographically for one CDK2 inhibitor (Furet et al., 2002) and hypothesized in another instance (Lee et al., 1996). "Similar" ligands may also adopt different protonation states on binding (Dullweber et al., 2001).

Perspective

We have reviewed some recent computational studies of ligand binding to proteins. Ultimately, to predict accurate binding affinities, it will be necessary to go beyond predicting a single "dominant" conformation of the ligand complexed with the protein. Binding free energy is not driven by a single conformation, but rather by the free energy landscape. It is the shape of the energy landscape that is crucial, the shape and width of the minima influences entropies. Entropies are key contributors to binding thermodynamics and are not observable in single bound structures. Other factors about the full landscape also play key roles, such as multiple ligand poses, protein conformations, strain energies, changes in water structure, and solvation and protonation all play roles. And none of these are observable in single structures. Computational tools can help provide insight into the unseen landscape, so those doing crystallographic studies may want to complement their work by using computational tools to explore this landscape. And those relying on crystallographic data (e.g., in a drug design context) should be aware that there are various binding possibilities that might not be captured in a single crystal structure.

ACKNOWLEDGMENTS

We thank Scott P. Brown (Abbott Laboratories), Sarah Boyce (University of California, San Francisco), John van Drie, and John D. Chodera (Stanford University), for help with references; Sarah Boyce, Eric Manas (GlaxoSmithKline), and Gabe Rocklin (University of California, San Francisco) for comments on the manuscript; and Sarina Bromberg for preparing figures. We acknowledge financial support of the National Institutes of Health (Grant GM34993 to K.A.D.).

REFERENCES

Abel, R., Young, T., Farid, R., Berne, B.J., and Friesner, R.A. (2008). Role of the active-site solvent thermodynamics of factor Xa ligand binding. J. Am. Chem. Soc. *130*, 2817–2831.

Arora, K., and Brooks, C.L., III. (2007). Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. Proc. Natl. Acad. Sci. U.S.A. *104*, 18496–18501.

Babaoglu, K., Simeonov, A., Irwin, J.J., Nelson, M.E., Feng, B., Thomas, C.J., Cancian, L., Costi, M.P., Maltby, D.A., Jadhav, A., et al. (2008). Comprehensive mechanistic analysis of hits from high-throughput and docking screens against beta-lactamase. J. Med. Chem. *51*, 2502–2511.

Badger, J., Minor, I., Kremer, M.J., Oliveira, M.A., Smith, T.J., Griffith, J.P., Guerin, D.M.A., Krishnaswamy, S., Luo, M., Rossmann, M.G., et al. (1988). Structural analysis of a series of antiviral agents complexed with human rhinovirus 14. Proc. Natl. Acad. Sci. U.S.A. *85*, 3304–3308.

Barillari, C., Taylor, J., Viner, R., and Essex, J.W. (2007). Classification of water molecules in protein binding sites. J. Am. Chem. Soc. *129*, 2577–2587.

Bash, P.A., Singh, U.C., Brown, F.K., Langridge, R., and Kollman, P.A. (1987). Calculation of the relative change in binding free energy of a protein-inhibitor complex. Science *235*, 574–576.

Birdsall, B., Feeney, J., Tendler, S.J.B., Hammond, S.J., and Roberts, G.C.K. (1989). Dihydrofolate reductase: multiple conformations and alternative modes of substrate binding. Biochemistry *28*, 2297–2305.

Boehr, D.D., McElheny, D., Dyson, H.J., and Wright, P.E. (2006). The dynamic energy landscape of dihydrofolate reductase catalysis. Science *313*, 1638–1642.

Böhm, H.J. (1993). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J. Comput. Aided Mol. Des. *8*, 243–256.

Böhm, H.J., and Klebe, G. (1996). What can we learn from molecular recognition in protein-ligand complexes for the design of new drugs? Angew. Chem. Int. Ed. Engl. *35*, 2588–2614.

Boresch, S., Tettinger, F., Leitgeb, M., and Karplus, M. (2003). Absolute binding free energies: a quantitative approach for their calculation. J. Phys. Chem. A *107*, 9535–9551.

Bostrom, J., Hogner, A., and Schmitt, S. (2006). Do structurally similar ligands bind in a similar fashion? J. Med. Chem. *49*, 6716–6725.

Brenk, R., Vetter, S.W., Boyce, S.E., Goodin, D.B., and Shoichet, B.K. (2006). Probing molecular docking at a charged model binding site. J. Mol. Biol. *357*, 1449–1470.

Chang, C.E., and Gilson, M.K. (2004). Free energy, entropy, and induced fit in host-guest recognition: calculations with the second-generation mining minima algorithm. J. Am. Chem. Soc. *126*, 13156–13164.

Chang, C.E., Chen, W., and Gilson, M.K. (2007). Ligand configurational entropy and protein binding. Proc. Natl. Acad. Sci. U.S.A. *104*, 1534–1539.

Charifson, P.S., and Kuntz, I.D. (1997). Recent successes and continuing limitations in computer aided drug design. In Practical Application of Computer Aided Drug Design, P.S. Charifson, ed. (New York: Dekker).

Cheatham, T.E., Srinivasan, J., Case, D.A., and Kollman, P.A. (1998). Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution. J. Biomol. Struct. Dyn. *16*, 265–280.

Chen, W., Chang, C.E., and Gilson, M.K. (2004). Calculation of cyclodextrin binding affinities: energy, entropy, and implications for drug design. Biophys. J. *87*, 3035–3049.

Chen, Y., Minasov, G., Roth, T.A., Prati, F., and Shoichet, B.K. (2006). The deacylation mechanism of AmpC beta-lactamase at ultrahigh resolution. J. Am. Chem. Soc. *128*, 2970–2976.

Chou, J.J., and Bax, A. (2001). Protein side-chain rotamers from dipolar couplings in a liquid crystalline phase. J. Am. Chem. Soc. *123*, 3844–3845.

Constantine, K.L., Mueller, L., Metzler, W.J., McDonnell, P.A., Todderud, G., Goldfarb, V., Fan, Y., Newitt, J.A., Keifer, S.E., Gao, M., et al. (2008). Multiple and single binding modes of fragment-like kinase inhibitors revealed by molecular modeling, residue type-selective protonation, and nuclear overhauser effects. J. Med. Chem. *51*, 6225–6229.

Corbeil, C.R., Englebienne, P., and Moitessier, N. (2007). Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. J. Chem. Inf. Model. *47*, 435–449.

Cozzini, P., Kellogg, G.E., Spyrakis, F., Abraham, D.J., Constntino, G., Emerson, A., Fanelli, F., Gohlke, H., Kuhn, L.A., Morris, G.M., et al. (2008). Target flexibility: an emerging consideration in drug discovery and design. J. Med. Chem. *51*, 6237–6255.

Czodrowski, P., Sotriffer, C.A., and Klebe, G. (2007). Atypical protonation states in the active site of HIV-1 protease: a computational study. J. Chem. Inf. Model. *47*, 1590–1598.

Deng, Y., and Roux, B. (2006). Calculation of standard binding free energies: aromatic molecules in the T4 lysozyme L99A mutant. J. Chem. Theory Comput. *2*, 1255–1273.

Deng, H., Zhadin, N., and Callender, R. (2001). Dynamics of protein ligand binding on multiple time scales: NADH binding to lactate dehydrogenase. Biochemistry *40*, 3767–3773.

Dullweber, F., Stubbs, M.T., Musil, D., Sturzebecher, J., and Klebe, G. (2001). Factorising ligand affinity: a combined thermodynamic and crystallographic study of trypsin and thrombin inhibition. J. Mol. Biol. *313*, 593–614.

Eisenmesser, E.Z., Bosco, D.A., Akke, M., and Kern, D. (2002). Enzyme dynamics during catalysis. Science *295*, 1520–1523.

Eisenmesser, E.Z., Millet, O., Labeikovsky, W., Korzhnev, D.M., Wolf-Watz, M., Bosco, D.A., Skalicky, J.J., Kay, L.E., and Kern, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. Nature *438*, 117–121.

Ferrara, P., Gohlke, H., Price, D.J., Klebe, G., Charles, L., and Brooks, I.I.I. (2004). Assessing scoring functions for protein-ligand interactions. J. Med. Chem. *47*, 3032–3047.

Fragai, M., Luchinat, C., and Parigi, G. (2006). "Four-dimensional" protein structures: examples from metalloproteins. Acc. Chem. Res. *39*, 909–917.

Fujitani, H., Tanida, Y., Ito, M., Shirts, M.R., Jayachandran, G., Snow, C.D., Sorin, E.J., and Pande, V.S. (2005). Direct calculation of the binding free energies of FKBP ligands. J. Chem. Phys. *123*, 084108.

Furet, P., Meyer, T., Strauss, A., Raccuglia, S., and Rondeau, J.M. (2002). Structure-based design and protein X-ray analysis of a protein kinase inhibitor. Bioorg. Med. Chem. Lett. *12*, 221–224.

Gatti, D.L., Palfrey, B.A., Lah, M.S., Entsch, B., Massey, V., Ballou, D.P., and Ludwig, M.L. (1994). The mobile flavin of 4-OH benzoate hydroxylase. Science *266*, 110–114.

Gerstein, M., Lesk, A.M., and Chothia, C. (1994). Structural mechanisms for domain movements in proteins. Biochemistry *33*, 6739–6749.

Gilson, M.K., and Zhou, H.-X. (2007). Calculation of protein-ligand binding affinities. Annu. Rev. Biophys. Biomol. Struct. *36*, 21–42.

Gohlke, H., and Klebe, G. (2002). Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew. Chem. Int. Ed. *41*, 2645–2676.

Graves, A.P., Brenk, R., and Shoichet, B.K. (2005). Decoys for docking. J. Med. Chem. *48*, 3714–3728.

Graves, A.P., Shivakumar, D.M., Boyce, S.E., Jacobson, M.P., Case, D.A., and Shoichet, B.K. (2008). Rescoring docking hit lists for model cavity sites: predictions and experimental testing. J. Mol. Biol. *377*, 914–934.

Greer, J., Erickson, J.W., Baldwin, J.J., and Varney, M.D. (1994). Application of the three-dimensional structures of protein target molecules in structure-based drug design. J. Med. Chem. *37*, 1035–1047.

Guimaraes, C.R.W., and Cardozo, M. (2008). MM-GB/SA rescoring of docking poses in structure-based lead optimization. J. Chem. Inf. Model. *48*, 958–970.

Gutteridge, A., and Thornton, J. (2005). Conformational changes observed in enzyme crystal structures upon substrate binding. J. Mol. Biol. *346*, 21–28.

Hamelberg, D., and McCammon, J.A. (2004). Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method. J. Am. Chem. Soc. *126*, 7683–7689.

Hao, M.H., Haq, O., and Muegge, I. (2007). Torsion angle preference and energetics of small-molecule ligands bound to proteins. J. Chem. Inf. Model. *47*, 2242–2252.

Hazuda, D.J., Anthony, N.J., Gomez, R.P., Jolly, S.M., Wai, J.S., Zhuang, L., Fisher, T.E., Embrey, M., Guare, J.P. Jr., Egbertson, M.S., et al. (2004). A naphthyridine carboxamide provides evidence for discordant resistance between mechanistically identical inhibitors of HIV-1 integrase. Proc. Natl. Acad. Sci. U.S.A. *101*, 11233–11238.

Head, M.S., Given, J.A., and Gilson, M.K. (1997). "Mining minima": direct computation of conformational free energy. J. Phys. Chem. A *101*, 1609–1618.

Helms, V., and Wade, R.C. (1995). Thermodynamics of water mediating protein-ligand interactions in cytochrome P450cam: a molecular dynamics study. Biophys. J. *69*, 810–824.

Helms, V., and Wade, R.C. (1998a). Computational alchemy to calculate absolute protein-ligand binding free energy. J. Am. Chem. Soc. *120*, 2710–2713.

Helms, V., and Wade, R.C. (1998b). Hydration energy landscape of the active site cavity in cytochrome P450cam. Proteins Struct. Funct. Genet. *32*, 381–396.

Henzler-Wildman, K., and Kern, D. (2007). Dynamic personalities of proteins. Nature *450*, 964–972.

Henzler-Wildman, K.A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M.A., Petsko, G.A., Karplus, M., et al. (2007). Intrinsic motions along an enzymatic reaction trajectory. Nature *450*, 838–844.

Hermans, J., and Subramaniam, S. (1986). The free energy of xenon binding to myoglobin from molecular dynamics simulation. Isr. J. Chem. *27*, 225–227.

Hermans, J., and Wang, L. (1997). Inclusion of the loss of translational and rotational freedom in theoretical estimates of free energies of binding. Application to a complex of benzene and mutant T4 lysozyme. J. Am. Chem. Soc. *119*, 2707–2714.

Hilser, V.J., and Thompson, E.B. (2007). Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. Proc. Natl. Acad. Sci. U.S.A. *104*, 8311–8315.

Huang, N., and Jacobson, M.P. (2007). Physics-based methods for studying protein-ligand interactions. Curr. Opin. Drug Discov. Dev. *10*, 325–331.

Huang, N., Kalyanaraman, C., Bernacki, K., and Jacobson, M.P. (2006). Molecular mechanics methods for predicting protein-ligand binding. Phys. Chem. Chem. Phys. *8*, 5166–5177.

Huey, R., Morris, G.M., Olson, A.J., and Goodsell, D.S. (2007). A semiempirical free energy force field with charge-based desolvation. J. Comput. Chem. *28*, 1145–1152.

Huo, S., Wang, J., Cieplak, P., Kollman, P.A., and Kuntz, I.D. (2002). Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design. J. Med. Chem. *45*, 1412–1419.

Jayachandran, G., Shirts, M.R., Park, S., and Pande, V.S. (2006). Parallelized-over-parts computation of absolute binding free energy with docking and molecular dynamics. J. Chem. Phys. *125*, 084901.

Jiao, D., Golubkov, P.A., Darden, T.A., and Ren, P. (2008). Calculation of protein-ligand binding free energy using a polarizable potential. Proc. Nat. Acad. Sci U.S.A. *105*, 6290–6295.

Jorgensen, W.L. (2004). The many roles of computation in drug discovery. Science *303*, 1813–1818.

Kangas, E., and Tidor, B. (2001). Electrostatic complementarity at ligand binding sites: application to chorismate mutase. J. Phys. Chem. B *105*, 880–888.

Kim, K.H. (2007a). Outliers in SAR and QSAR: 2. Is a flexible binding site a possible source of outliers? J. Comput. Aided Mol. Des. *21*, 421–435.

Kim, K.H. (2007b). Outliers in SAR and QSAR: Is unusual binding mode a possible source of outliers? J. Comput. Aided Mol. Des. *21*, 63–86.

Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., et al. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc. Chem. Res. *33*, 889–897.

Kuhn, B., and Kollman, P.A. (2000). Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. J. Med. Chem. *43*, 3786–3791.

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., and Ferrin, T.E. (1982). A geometric approach to macromolecule-ligand interactions. J. Mol. Biol. *161*, 269–288.

Laederach, A., and Reilly, P.J. (2003). Specific empirical free energy function for automated docking of carbohydrates to proteins. J. Comput. Chem. *24*, 1748–1757.

LaLonde, J.M., Bernlohr, D.A., and Banaszak, L.J. (1994). X-ray crystallographic structures of adipocyte lipid-binding protein complexed with palmitate and hexadecanesulfonic acid. Properties of cavity binding sites. Biochemistry *33*, 4885–4895.

Lazaridis, T., Matsunov, A., and Gandolfo, F. (2002). Contributions to the binding free energy of ligands to avidin and streptavidin. Proteins *47*, 194–208.

Leach, A.R. (1994). Ligand docking to proteins with discrete side-chain flexibility. J. Mol. Biol. *235*, 345–356.

Lee, M.S., and Olson, M.A. (2006). Calculation of absolute protein-ligand binding affinity using path and endpoint approaches. Biophys. J. *90*, 864–877.

Lee, H., Reyes, V.M., and Kraut, J. (1996). Crystal structures of Escherichia coli dihydrofolate reductase complexed with 5-formyltetrahydrofolate (folinic acid) in two space groups: evidence for enolization of pteridine O4. Biochemistry *35*, 7012–7020.

Liverton, N.J., Holloway, M.K., McCauley, J.A., Rudd, M.T., Butcher, J.W., Carroll, S.S., DiMuzio, J., Fandozzi, C., Gilbert, K.F., Mao, S.-S., et al. (2008). Molecular modeling based approach to potent P2-P4 macrocyclic inhibitors of hepatitis C NS3/4A protease. J. Am. Chem. Soc. *130*, 4607–4609.

Lu, Y., Wang, C.-Y., and Wang, S. (2006). Binding free energy contributions of interfacial waters in HIV-1 protease/inhibitor complexes. J. Am. Chem. Soc. *128*, 11830–11839.

Ma, B., Shatsky, M., Wolfson, H.J., and Nussinov, R. (2002). Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. Protein Sci. *11*, 184–197.

Mardis, K.L., Luo, R., and Gilson, M.K. (2001). Interpreting trends in the binding of cyclic ureas to HIV-1 protease. J. Mol. Biol. *309*, 507–517.

Meiler, J., and Baker, D.A. (2006). ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. Proteins *65*, 538–548.

Mewshaw, R.E., Edsall, R.J., Jr., Yang, C., Manas, E.S., Xu, Z.B., Henderson, R.A., Keith, J.C., Jr., and Harris, H.A. (2005). ERbeta ligands. 3. Exploiting two binding orientations of 2-phenylnapthalene scaffold to achieve ERbeta selectivity. J. Med. Chem. *48*, 3953–3979.

Miller, D.W., and Dill, K.A. (1997). Ligand binding to proteins: the binding landscape model. Protein Sci. *6*, 2166–2179.

Min, W., English, B.P., Luo, G., Cherayil, B.J., Kou, S.C., and Xie, X.S. (2005a). Fluctuating enzymes: lessons from single-molecule studies. Acc. Chem. Res. *38*, 923–931.

Min, W., Luo, G., Cherayil, B.J., Kou, S.C., and Xie, X.S. (2005b). Observation of a power-law memory kernel for fluctuations within a single protein molecule. Phys. Rev. Lett. *94*, 198302.

Mobley, D.L., Chodera, J.D., and Dill, K.A. (2006). On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. J. Chem. Phys. *125*, 084902.

Mobley, D.L., Chodera, J.D., and Dill, K.A. (2007a). Confine-and-release method: obtaining correct binding free energies in the presence of protein conformational change. J. Chem. Theory Comput. *3*, 1231–1235.

Mobley, D.L., Graves, A.P., Chodera, J.D., McReynolds, A.C., Shoichet, B.K., and Dill, K.A. (2007b). Predicting absolute ligand binding free energies to a simple model site. J. Mol. Biol. *371*, 1118–1134.

Mobley, D.L., Dill, K.A., and Chodera, J.D. (2008). Treating entropy and conformational changes in implicit solvent simulations of small molecules. J. Phys. Chem. B *112*, 938–946.

Montfort, W.R., Perry, K.M., Fauman, E.B., Finer-Moore, J.S., Maley, G.F., Hardy, L., Maley, F., and Stroud, R.M. (1990). Structure, multiple site binding, and segmental accomodation in thymidylate synthase on binding dUMP and an anti-folate. Biochemistry *29*, 6964–6977.

Morton, A., and Matthews, B.W. (1995). Specificity of ligand binding in a buried nonpolar cavity of {T4} lysozyme: linkage of dynamics and structural plasticity. Biochemistry *34*, 8576–8588.

Murthy, K.H.M., Winborne, E.L., Minnich, M.D., Culp, J.S., and Debouck, C. (1992). The crystal structures at 2.2-A resolution of hydroxyethylene-based inhibitors bound to human immunodeficiency virus type 1 protease show that the inhibitors are present in two distinct orientations. J. Biol. Chem. *267*, 22770–22778.

Najmanovich, R., Kuttner, J., Sobolev, V., and Edelman, M. (2000). Side-chain flexibility in proteins upon ligand binding. Proteins *39*, 261–268.

Noble, M.E., Endicott, J.A., and Johnson, L.N. (2004). Protein kinase inhibitors: insights into drug design from structure. Science *303*, 1800–1805.

Olano, L.R., and Rick, S.W. (2004). Hydration free energies and entropies for water in protein interiors. J. Am. Chem. Soc. *126*, 7991–8000.

Oostenbrink, C., and van Gunsteren, W.F. (2004). Free energies of binding of polychlorinated biphenyls to the estrogen receptor from a single simulation. Proteins *54*, 237–246.

Orville, A.M., Elango, N., Lipscomb, J.D., and Ohlendorf, D.H. (1997). Structures of competitive inhibitor complexes of protocatechuate 3,4-dioxygenase: multiple exogenous ligand binding orientations within the active site. Biochemistry *36*, 10039–10051.

Pan, C., Mezei, M., Mujtaba, S., Muller, M., Zeng, L., Li, J., Wang, Z., and Zhou, M.M. (2007). Structure-guided optimization of small molecules inhibiting human immunodeficiency virus 1 Tat association with the human coactivator p300/CREB binding protein-associated factor. J. Med. Chem. *50*, 2285–2288.

Paulsen, M.D., and Ornstein, R.L. (1992). Predicting the product specificity and coupling of cytochrome P450cam. J. Comput. Aided Mol. Des. *6*, 449–460.

Pearlman, D.A. (2005). Evaluating the molecular mechanics poisson-boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase. J. Med. Chem. *48*, 7796–7807.

Pearlman, D.A., and Charifson, P.S. (2001). Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system. J. Med. Chem. *44*, 3417–3423.

Pei, Z., Li, X., Longenecker, K., Geldern, T.W., Wiedeman, P.E., Lubben, T.H., Zinker, B.A., Stewart, K., Ballaron, S.J., Stashko, M.A., Mika, A.K., Beno, D.W., et al. (2006). Discovery, structure-activity relationship, and pharmacological evaluation of (5-substituted-pyrrolidinyl-2-carbonyl)-2-cyanopyrrolidines as potent dipeptidyl peptidase IV inhibitors. J. Med. Chem. *49*, 3520–3535.

Perola, E., and Charifson, P.S. (2004). Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. J. Med. Chem. *47*, 2499–2510.

Pospisil, P., Ballmer, P., Scapozza, L., and Folkers, G. (2003). Tautomerism in computer-aided drug design. Journal of Receptors and Signal Transduction *23*, 361–371.

Raag, R., and Poulos, T.L. (1991). Crystal structures of cytochrome P-450CAM complexed with camphane, thiocamphor, and adamantane: factors controlling P-450 substrate hydroxylation. Biochemistry *30*, 2674–2684.

Radivojac, P., Iakoucheva, L.M., Oldfield, C.J., Obradovic, Z., Uversky, V.N., and Dunker, A.K. (2007). Intrinsic disorder and functional proteomics. Biophys. J. *92*, 1439–1456.

Reddy, M.R., and Erion, M.D. (2001). Calculation of relative binding free energy differences for fructose 1,6-bisphosphatase inhibitors using the thermodynamic cycle perturbation approach. J. Am. Chem. Soc. *123*, 6246–6252.

Reich, S.H., Melnick, M., II, Davies, J.F., Appelt, K., Lewis, K.K., Fuhry, M.A., Pino, M., Trippe, A.J., Nguyen, D., Dawson, H., et al. (1995). Protein structure-based design of potent orally bioavailable, nonpeptide inhibitors of human immunodeficiency virus protease. Proc. Natl. Acad. Sci. U.S.A. *92*, 3298–3303.

Rizzo, R.C., and Jorgensen, W.L. (1999). OPLS all-atom model for amines: resolution of the amine hydration problem. J. Am. Chem. Soc. *121*, 4827–4836.

Rizzo, R.C., Toba, S., and Kuntz, I.D. (2004). A molecular basis for the selectivity of thiadiazole urea inhibitors with stromelysin-1 and gelatinase-A from generalized born molecular dynamics simulations. J. Med. Chem. *47*, 3065–3074.

Roux, B., Nina, M., Poms, R., and Smith, J.C. (1996). Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy perturbation study. Biophys. J. *71*, 670–681.

Ruvinsky, A.M., and Kozintsev, A.V. (2005). New and fast statistical-thermodynamic method for computation of protein-ligand binding entropy substantially improves docking accuracy. J. Comput. Chem. *26*, 1089–1095.

Schames, J.R., Henchman, R.H., Siegel, J.S., Sotriffer, C.A., Ni, H., and McCammon, J.A. (2004). Discovery of a novel binding trench in HIV integrase. J. Med. Chem. *47*, 1879–1881.

Schwarzl, S.M., Tschopp, T.B., Smith, J.C., and Fischer, S. (2002). Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas entropy correction? J. Comput. Chem. *23*, 1143–1149.

Sharp, K.A. (2005). Important considerations impacting molecular docking. In Virtual Screening in Drug Discovery, B.K. Shoichet and J. Alvarez, eds. (Boca Raton, FL: CRC Press).

Sherman, W., Day, T., Jacobson, M.P., Friesner, R.A., and Farid, R. (2006). Novel procedure for modeling ligand/receptor induced fit effects. J. Med. Chem. *49*, 534–553.

Shirts, M.R., Mobley, D.L., and Brown, S.P. (2009). Free energy calculations in structure-based drug design. In Structure Based Drug Discovery, K.M. Merz, D. Ringe, and C.H. Reynolds, eds. (Cambridge: Cambridge University Press).

Shirts, M.R., Mobley, D.L., and Chodera, J.D. (2007). Alchemical free energy calculations: ready for prime time? Ann. Rep. Comput. Chem. *3*, 41–59.

Shoichet, B.K., Leach, A.R., and Kuntz, I.D. (1999). Ligand solvation in molecular docking. Proteins *34*, 4–16.

Shoichet, B.K., McGovern, S.L., Wei, B., and Irwin, J.J. (2002). Lead discovery using molecular docking. Curr. Opin. Chem. Biol. *6*, 439–446.

Sousa, S.F., Fernandes, P.A., and Ramos, M.J. (2006). Protein-ligand docking: current status and future challenges. Proteins *65*, 15–26.

Srinivasan, J., Cheatham, T.E. III, Cieplak, P., Kollman, P.A., and Case, D.A. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. J. Am. Chem. Soc. *120*, 9401–9409.

Stauffer, S.R., Stanton, M.G., Gregro, A.R., Steinbeiser, M.A., Shaffer, J.R., Nantermet, P.G., Barrow, J.C., Rittle, K.E., Collusi, D., Espeseth, A.S., et al. (2007). Discovery and SAR of isonicotinamide BACE-1 inhibitors that bind [beta]-secretase in a N-terminal 10s-loop down conformation. Bioorg. Med. Chem. Lett. *17*, 1788–1792.

Steinbrecher, T., Case, D.A., and Labahn, A. (2006). A multistep approach to structure-based drug design: studying ligand binding at the human neutrophil elastase. J. Med. Chem. *49*, 1837–1844.

Steuber, H., Czodrowski, P., Sotriffer, C.A., and Klebe, G. (2007). Tracing changes in protonation: a prerequisite to binding to aldose reductase. J. Mol. Biol. *373*, 1305–1320.

Stoll, V., Stewart, K.D., Maring, C.J., Muchmore, S., Giranda, V., Gu, Y.G., Wang, G., Chen, Y., Sun, M., Zhao, C., et al. (2003). Influenza neuraminidase inhibitors: structure-based design of a novel inhibitor series. Biochemistry *42*, 718–727.

Stout, T.J., Tondi, D., Rinaldi, M., Barlocco, D., Pecorari, P., Santi, D.V., Kuntz, I.D., Stroud, R.M., Shoichet, B.K., and Costi, M.P. (1999). Structure-based design of inhibitors specific for bacterial thymidylate synthase. Biochemistry *38*, 1607–1617.

Stubbs, M.T., Reyda, S., Dullweber, F., Moller, M., Klebe, G., Dorsch, D., Mederski, W.W.K.R., and Wurziger, H. (2002). pH-dependent binding modes observed in trypsin crystals: lessons for structure-based drug design. Chembiochem *3*, 246–249.

Sulsky, R., Magnin, D.R., Huang, Y., Simpkins, L., Taunk, P., Patel, M., Zhu, Y., Stouch, T.R., Bassolino-Klimas, D., Parker, R., et al. (2007). Potent and selective biphenyl azole inhibitors of adipocyte fatty acid binding protein (aFABP). Bioorg. Med. Chem. Lett. *17*, 3511–3515.

Tang, C., Schweiters, C.D., and Clore, G.M. (2007). Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. Nature *449*, 1078–1082.

Taylor, R.D., Jewsbury, P.J., and Essex, J.W. (2002). A review of protein-small molecule docking methods. J. Comput. Aided Mol. Des. *16*, 151–166.

Teague, S.J. (2003). Implications of protein flexibility for drug discovery. Nat. Rev. Drug Discov. *2*, 527–541.

Tembe, B.L., and McCammon, J.A. (1984). Ligand-receptor interactions. Comput. Chem. *8*, 281–283.

Thomson, J.A., and Perni, R.B. (2006). Hepatitis C virus NS3-4A protease inhibitors: countering viral subversion in vitro and showing promise in the clinic. Curr. Opin. Drug Discov. Devel. *9*, 606–617.

Tirado-Rives, J., and Jorgensen, W.L. (2006). Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. J. Med. Chem. *49*, 5880–5884.

Uytterhoeven, K., Sponer, J., and Meervelt, L.V. (2002). Two 1:1 binding modes for distamycin in the minor groove of d(GGCCAATTGG). Eur. J. Biochem. *269*, 2868–2877.

Vajpai, N., Strauss, A., Cowan-Jacob, S.W., Manley, P.W., Crzesiek, S., and Jahnke, W. (2008). Solution conformations and dynamics of ABL kinase-inhibitor complexes determined by NMR substantiate the different binding modes of imatinib/nilotinib and dasatinib. J. Biol. Chem. *283*, 18292–18302.

Volkman, B.F., Lipson, D., Wemmer, D.E., and Kern, D. (2001). Two-state allosteric behavior in a single-domain signaling protein. Science *291*, 2429–2433.

von Itzstein, M., Wu, W.Y., Kok, G.B., Pegg, M.S., Dyason, J.C., Jin, B., Van Phan, T., Smythe, M.L., White, H.F., Oliver, S.W., et al. (1993). Rational design of potent sialidase-based inhibitors of influenza virus replication. Nature *363*, 418–423.

Vorobjev, Y.N., and Hermans, J. (1999). ES/IS: estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. Biophys. Chem. *78*, 195–205.

Wang, J., Deng, Y., and Roux, B. (2006). Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. Biophys. J. *91*, 2798–2814.

Warren, G.L., Andrews, C.W., Capelli, A.M., Clarke, B., LaLonde, J., Lambert, M.H., Lindvall, M., Nevins, N., Semus, S.F., Sender, S., et al. (2006). A critial assessment of docking programs and scoring functions. J. Med. Chem. *49*, 5912–5931.

Warren, G., and Perola, E. (2008). Field of extremes: if you re-refine them, they will come down. Paper presented at CUP IX, March 16–19, 2008, Santa Fe, NM. Available online at http://www.eyesopen.com/about/events/cups-2008/pdfs-CUP/CUP9-Field-of-Extremes.pdf.

Warshel, A., Sussman, F., and King, G. (1986). Free energy of charges in solvated proteins: microscopic calculations using a reversible charging process. Biochemistry *25*, 8368–8372.

Wei, B.Q., Weaver, L.H., Ferrari, A.M., Matthews, B.W., and Shoichet, B.K. (2004). Testing a flexible-receptor docking algorithm in a model binding site. J. Mol. Biol. *337*, 1161–1182.

Weisberg, E., Manley, P.W., Cowan-Jacob, S.W., Hochhaus, A., and Griffin, J.D. (2007). Second generation inhibitors of BCR-ABL for the treatment of imatinib-resistant chronic myeloid leukaemia. Nat. Rev. Cancer *7*, 345–356.

Wong, C.F., and McCammon, J.A. (1986). Dynamics and design of enzymes and inhibitors. J. Am. Chem. Soc. *108*, 3830–3832.

Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. *293*, 321–331.

Zhang, L., and Hermans, J. (1996). Hydrophilicity of cavities in proteins. Proteins Struct. Funct. Genet. *24*, 433–438.

Zhang, X.J., Wozniak, J.A., and Matthews, B.W. (1995). Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. J. Mol. Biol. *250*, 527–552.

Zhou, G., Ferrer, M., Chopra, R., Kapoor, T.M., Strassmaier, T., Weissenhorn, W., Skehel, J.J., Oprian, D., Schreiber, S.L., Harrison, S.C., and Wiley, D.C. (2000). The structure of an HIV-1 specific cell entry inhibitor in complex with the HIV-1 gp41 trimeric core. Bioorg. Med. Chem. *8*, 2219–2227.

Highly infrared reflective nickel doped ZrO_2 from first principles simulation

Shizhong Yang[1,2], S.M Guo[3], Guang-Lin Zhao[1], and Ebrahim Khosravi[2]
[1]*Physics Department, Southern University and A & M College, Baton Rouge, Louisiana 70813*
[2]*Department of Computer Science, Southern University and A & M College, Baton Rouge, Louisiana 70813*
[3]*Department of Mechanic Engineering, Louisiana State University, Baton Rouge, Louisiana 70803*

Abstract

First principles（or *ab initio*） density-functional-theory (DFT) with projected augmented wave (PAW) method simulations were performed to calculate the electronic structures and optical properties of 25% nickel (Ni) doped cubic ZrO_2 crystals. We implemented two *ab initio* DFT application methods to the ZrO_2 ceramic elastic constant, structure stability, and optical property calculation. The Cerply-Alder type local density approximation（LDA） models show that the interstitial and substituting nickel doped ZrO_2 structures are metastable, through the elastic stability analysis, while the structure of 25% Ni doped at substitute site with a Zr vacancy is in a stable state. The reflectivity at different directions is evaluated by calculating the real and imaginary part of the dielectric constants. The result shows that the reflectivity of Ni doped ZrO_2 crystal, with a Zr vacancy structure, varies from 59% to 80% at (001) ~ (111) directions respectively. In comparison, the reflectivity of pure ZrO_2 is only 18% in the infrared wavelength range. The high reflectivity of 25% Ni doped ZrO_2 structure, with a Zr vacancy, is caused by the unique doped crystal structure and the associated vacancy charge state in this configuration.

I. INTRODUCTION

Zirconia (ZrO_2) stabilized by rare earth elements like Y^{3+} is often used as thermal barrier coatings (TBC)[1, 2] in gas turbines, and rocket engines, due to its superior properties such as low thermal conductivity, chemically inert, and high corrosion resistance. Yttria-stabilized zirconia is also used as the solid electrolytes for solid oxide fuel cells. The stabilized ZrO_2 has a cubic fluorite structure at room temperature. Zirconia based TBCs are transparent or translucent to radiation wavelength in the 0.3μm~5μm range[3]. Because more than 90% of radiation is within this range at typical gas turbine temperatures of 1700~2000 ºK, to effectively reduce thermal radiation transport through TBC systems, thus to improve the thermal insulation function, researches have been carried out with emphasis to increase the photon scattering within the coating and to improve the coating's reflectivity property. Increasing the density of scattering defects, such as micro-cracks and pores within the coating, has been reported as an efficient way of reducing thermal radiation [4].

In solid oxide fuel cell (SOFC) application, Ni is usually used as an effective electron transport channel, while in TBC application Ni is added with Cr, Al, and Y as a buffer layer to form good binding and match the coefficient of thermal expansion between substrate and the ceramic top layer, mainly ZrO_2. An earlier theoretical simulation of carbon doped ZrO_2 containing interstitial or substitution carbon and zirconium vacancy was reported by Ivanovskii *et al*.[5]. *Ab initio* DFT method was successfully applied in ceramic elastic constant calculations [6, 7, 8], including cubic[6] and monoclinic ZrO_2[8]. However to the best of the authors' knowledge, no experimental and theoretical study has been conducted on the stabilized high concentration Ni

doped ZrO_2. In this report, we applied the stress-strain elastic constant calculation method, analyzed the three Ni doped ZrO_2 systems and concluded that the structure of 25% Ni doped at substitute site with a Zr vacancy is in a most stable structure. We also implemented an effective reflectivity (reflection coefficient) calculation method to the above Ni doped ZrO_2 system and our results show that at (111) of the doped ZrO_2, the reflectivity in infrared range reaches 80%.

II. COMPUTATIONAL METHODS

In this report, *ab initio* DFT with the projector augmented wave (PAW) plane-wave method [9, 10], relativistic effect was considered, was used to calculate the elastic constant and electronic properties of ZrO_2 composites. The local density approximation (LDA) and Ceperley-Alder potential (quantum Monte Carlo based) [11] was used in this study. For O atoms, the 2s and 2p electrons were described as valence, for Zr the 4s, 4p, 4d, and 5s electrons were treated as valence, whereas for Ni the 3d and 4s electrons were treated as valence. The remaining electrons were kept in a frozen core. The calculated total energies converged to a value less than 1meV while using the plane wave energy cutoff of 450 eV. With this setting and a 12 x 12 x 12 *k*-space Monkhost grid, the lattice constant of ZrO_2 crystal is found to be a = 5.062 Å, in excellent agreement with the value 5.07 Å published in JCPDS file 7-337 for cubic zirconia. All of our calculations are based on the same potential, energy cutoff, energy and residue force convergence criteria as those used in the lattice constant calculations.

(a) **(b)**

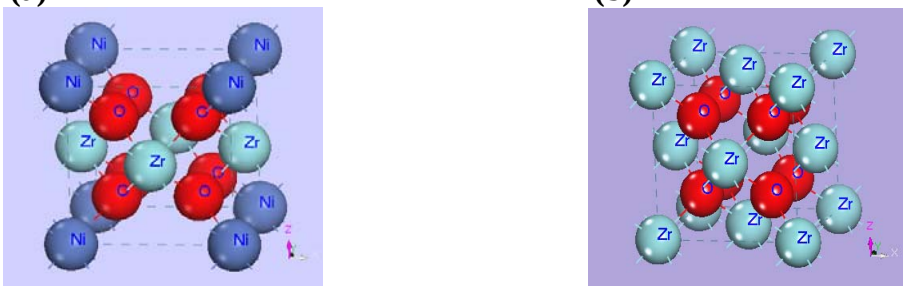


Figure 1. (a). The Ni doped at a substitution site with a Zr vacancy; and (b) pure ZrO_2 unit cell. The blue sphere stands for Ni atoms, the light green sphere represents the Zr atoms, and the red sphere stands for oxygen atoms.

The unit cell models are shown in Fig. 1. In all of the electronic calculations presented in this report, the *k*-space sampling uses 12 x 12 x 12 Monkhost grid. The electronic energy convergence value was set to be 10^{-5} eV and the residue force was set to be 1meV/Å. The unit cell shape, size, and atomic coordinates of the systems were relaxed in the optimization process. To save computer time, symmetry was considered in all of the elastic constant calculations.

There are two *ab initio* methods to calculate the elastic constants. One is the fitting method by fitting the total energies with respect to related strains near the ground state energy. The other is a straight forward method by analyze the stress-strain relations. This method can be traced back to the work of Nielson and Martin[12]. According to Hooks law, the stress component σ_i (i=1~6) is in linear dependency of the applied strain ε_j (j=1~6) under a small deformation:

$$\sigma_i = C_{ij}\,\varepsilon_j \qquad (1)$$

The linear elastic constants C_{ij} form a 6 × 6 symmetric matrix, having 27 different components. Due to symmetry, a cubic crystal has only three different symmetry elements (C_{11},

C_{12} and C_{44}), each of which represents three equal elastic constants ($C_{11}=C_{22}=C_{33}$; $C_{12}=C_{23}=C_{31}$; $C_{44}=C_{55}=C_{66}$)[13]. Properties such as the bulk modulus and shear modulus can be computed from the values of C_{ij}. Methods to determine the elastic constants from first principles usually involve setting the strain to a finite value, re-optimizing any free parameters and calculating the stress. By carefully choosing the applied deformation, elastic constants can then be determined. The deformation value needs to be tested before final calculation is done. Too big deformation may cause nonlinear effect, while too small deformation could also induce large force error. In our Ni doped ZrO_2 calculation, we limit our deformation maximum to be 0.01. The stress-strain elastic constant calculation method allows us to predict elastic constants for new materials, materials where experimental data do not exist, and to resolve discrepancies between contradictory experimental results.

Under the same setting, after the structure optimization, the optical dielectric tensor was calculated by Fermi golden rule using PAW method. The imaginary part (ε_2) was calculated first, then the real part was deduced by Kramers-Kronig transform. The reflectivity *R* was calculated by

$$R = \left[\frac{1-N}{1+N}\right]^2, \tag{2}$$

where N is the reflective index which can be evaluated by:

$$N^2 = \varepsilon_1 + i\varepsilon_2, \tag{3}$$

in which ε_1 and ε_2 are the real and imaginary parts of the complex dielectric constant. The spin polarization was considered in all of the calculations.

IIIa. RESULTS OF CALCULATED ELECTRONIC STRUCTURE

To validate the simulation software, the lattice constant and bulk modulus of pure ZrO_2 was calculated first. Excellent agreement was found between the predicted ZrO_2 lattice constant and the corresponding experimental results. Based on the settings used in the validation lattice calculation, further calculations were performed on the elastic constants for pure ZrO_2, Ni interstitial and substitution site doped ZrO_2, and Ni substitution site doped with a Zr vacancy ZrO_2 crystal. To have a stable cubic structure, the following conditions[13] should be held for each test case,

$$C_{44}>0,\ C_{11}>|C_{12}|,\ \text{and}\ C_{11}+2C_{12}>0 \tag{4}$$

The calculation results show that the C_{44} of both Ni interstitial and substitution site doped structures are negative, i.e. they are meta-stable structures. Thus only pure and Ni substitution site doped with a Zr vacancy ZrO_2 crystals (the term "Ni doped ZrO_2" is used for this case hereafter) were reported in Table I for elastic constant values.

The optimized lattice constants of Ni doped ZrO_2 are a=b=5.27 Å, c=4.20 Å, that are 4% larger and 17% smaller than that of the pure ZrO_2 crystal value of 5.07 Å. The calculated total magnetic moment is very small (less than 0.005 μ_B). It can be seen from Table I that the Voigt bulk modulus of experiment is 212.33 GPa, which is close to our result of 226.89 GPa in the simulation. The C_{11}, C_{33}, and C_{44} of Ni doped ZrO_2 decrease tremendously compared to their pure ZrO_2 values, i.e. from 532.15, 532.15, and 55.70 GPa to 139.31, 128.26, and 0.23 GPa respectively. While the C_{66} value only has a minor decrease, from 55.70 GPa to 38.03 GPa,

comparing to the above mentioned large variations. The calculated Bader charge[14] results of both systems are listed in Table II.

Table I. Elastic constants of pure and Ni doped ZrO_2 (in GPa). B_v stands for Voigt bulk modulus, while S_v stands for Voigt shear modulus. Temperature unit is in K, Y_2O_3 is in mol% ratio.

System	Temperature	Y_2O_3	C_{11}	C_{12}	C_{13}	C_{33}	C_{44}	C_{66}	B_v	S_v	Ref
Experiment	300	8%	413	112	112	413	61	61	212.33	96.80	(Ref. [15])
	300	8%	394	91	91	394	56	56	190		(Ref. [16])
	300	18%	375	75	75	375	64	64			(Ref. [17])
	300	15%	475	144	144	475	61	61			(Ref. [18])
	300	8.1%	402	95	95	402	56	56			(Ref. [19])
Theory	0	0%	500±100	90±20	90±20	500±100					(Ref. [6])
	0	0%	222	61	61	222	54	54	115		(Ref. [20])
	0	0%	455	64	64	455	63	63			(Ref. [21])
Pure ZrO_2	0	0%	532.15	74.24	74.24	532.15	55.70	55.70	226.89	125.00	this work
Ni doped ZrO_2	0	0%	139.31	133.27	75.36	128.26	0.23	38.03	108.32	15.89	this work

Table II. Bader charge of pure and Ni doped ZrO_2. (in unit of electron charge |e|)

System	Zr	O	Ni
Pure ZrO_2	2.64	-1.32	N/A
Ni doped ZrO_2	2.58	-0.81~-0.85	1.45

Figure 2 shows the electron density of states (DOS) of s and p orbitals of Zr and O for pure and Ni doped ZrO_2. The Fermi level is set at 0.0eV. The p_z component of Zr in Fig. 2(a) and 2(b) has shown a down-shift near -2eV. The p_y intensity near 2eV increases slightly. Comparing the p_z DOS of O atoms in Fig. 2(c) and 2(d), it is clear that in pure ZrO_2, p_z extends from -6eV to -0.5 eV below Fermi level with two major peaks at -5.3 eV and -1.2eV, while in Zr doped ZrO_2, p_z state of O atom extends from about -7.2 eV up to 2eV above the Fermi level with a small peak at 1.3 eV. Similarly p_y and p_x also show band extension from below Fermi level to 1eV above and from -7 eV to -6 eV. In Fig. 3(a) and 3(b), the Zr d_{xy}, d_{yz}, and d_{xz} are mainly located below the Fermi level in both pure and Ni doped ZrO_2 with extension to +2eV and -7 eV. The Zr d_{z^2} and d_{x^2} of Ni doped ZrO_2 are also more extended, comparing to the pure ZrO_2 case with the peaks sitting at 3.5eV and 4 eV. Thus it is clear that the p orbitals of O are hybridized with p and d_{xy}, d_{yz}, and d_{xz} orbitals of Zr with energy range extended to higher and lower energy scales.

Fig. 3(c) shows that the symmetric nature of the spin up and down states of Ni d orbitals makes the total magnetic moment zero. The d_{z^2} and d_{x^2} orbitals of Ni atom are below Fermi level which is different from that of Zr d_{z^2} and d_{x^2} where both are above the Fermi level, while the d_{xy}, d_{yz}, and d_{xz} orbitals are extended to both higher and lower energy scales similar to that in Zr.

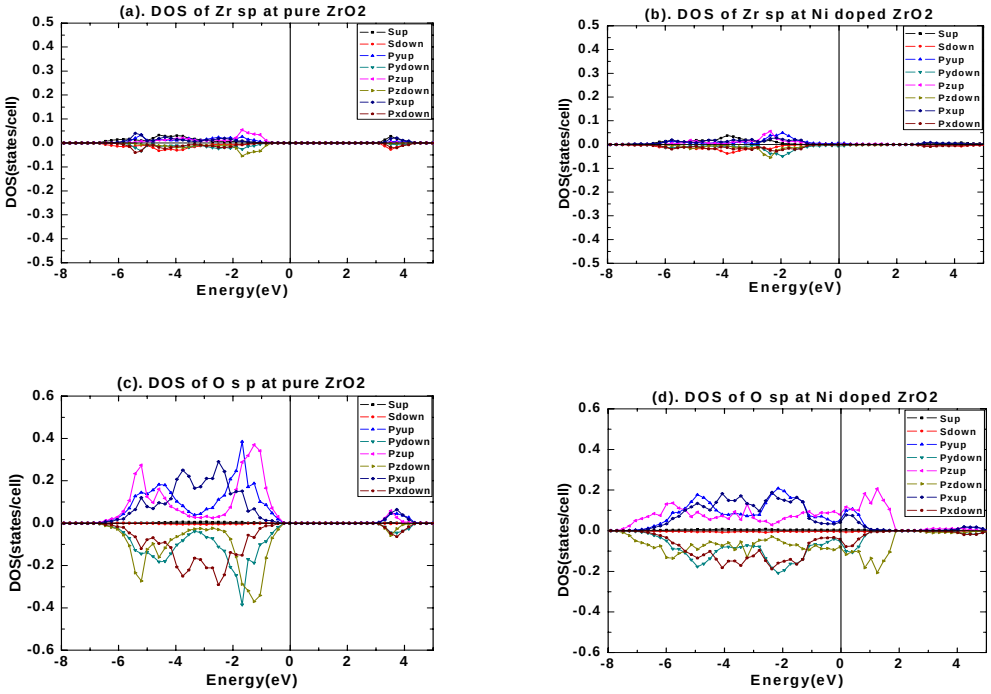
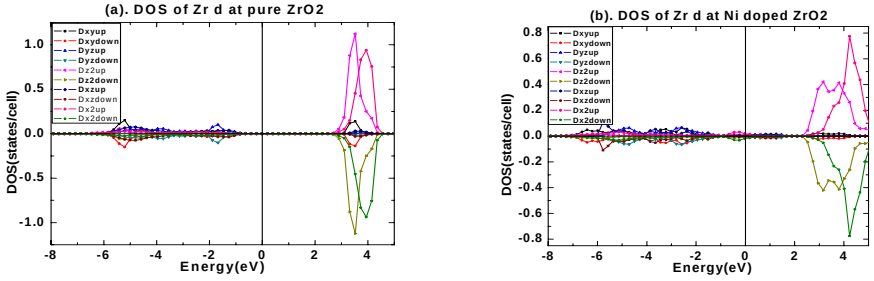


Figure 2. The partial DOS of s and p states of Ni doped and pure ZrO_2. The black square stands for s states with spin up, the red circle stands for s states with spin down; blue upward triangle for p_y electron with spin up, light green downward triangle for p_y down, pink leftward triangle for p_z up, gold rightward triangle for p_z down, cyan diamond for p_x up, while the brown pentagon for the p_x with a down spin.



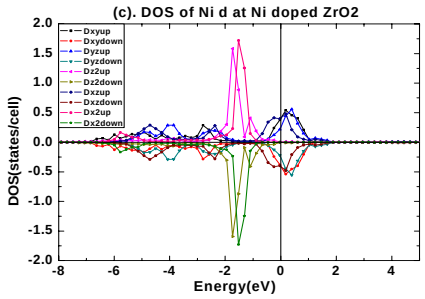


Figure 3. The partial DOS of d states in Ni doped and pure ZrO_2. The black square stands for d_{xy} states with spin up, the red circle stands for d_{xy} states with spin down; blue upward triangle for d_{yz} with spin up, light green downward triangle for d_{yz} down, pink leftward triangle for d_{z^2} up, gold rightward triangle for d_{z^2} down, cyan diamond for d_{xz} up, brown pentagon for d_{xz} down, light red hexagon for d_{x^2} up, while green star for d_{x^2} down.

III b. OPTICAL PROPERTY RESULTS

The electromagnetic reflectivity properties of pure and Ni doped ZrO_2 are shown in Fig. 4(a) and 4(b). The reflectivity data of pure ZrO_2 at (010) direction was ignored in Fig. 4(a) since it has the same curve as in (001) direction. From Fig. 4 (a) and (b), it can be seen that the reflectivity is increased from (001), (011), to (111) direction for both pure and Ni doped ZrO_2 crystals. It can also be seen from Fig. 4(a) and 4(b) that the reflectivity of Ni doped ZrO_2 increased by 3.4, 4.1, 2.6, and 2.3 times in comparison to pure ZrO_2 cases along the (001), (010), (011), and (111) polarization directions respectively.

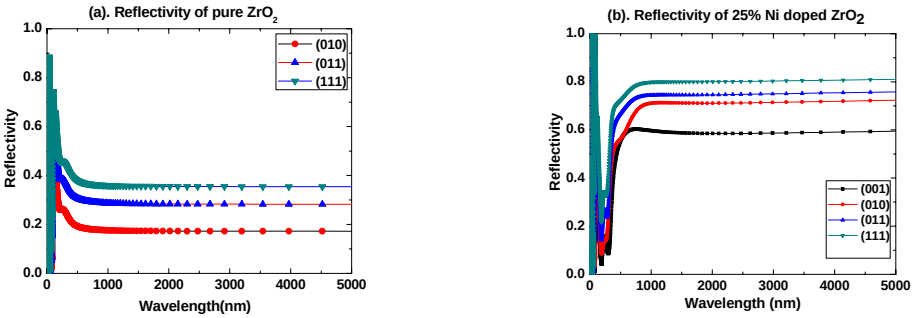


Figure 4. (a). The reflectivity of pure ZrO_2. (b). The reflectivity of 25% Ni doped ZrO_2. The black squares stand for the reflectivity in (001) direction. The red circles stand for the reflectivity in (010) direction. The up blue triangles stand for the reflectivity in (011) direction. The down triangles (in dark cyan color) stand for the reflectivity in (111) direction.

IV. DISCUSSION AND CONCLUSIONS

From Table I, we can see that most experimental elastic constant data of cubic ZrO_2 were measured at near 8% Y doped and room or higher temperatures while our calculation for the cubic ZrO_2 contains 0% Y at zero temperature, we would expect some differences between experimental data and our calculation results. Actually our calculated C_{12}, C_{44} results are close to the data at Reference 16. The C_{11} is off the experimental data but is in the range of former *ab*

initio DFT calculation presented in Reference 6. From Table I we can also see that among the methods of *ab initio* DFT, muffin-tin(MT) approximation with free electron gas pair potential, and lattice dynamics, the *ab initio* DFT method is the best one which gives good agreement with the corresponding experimental data. This is due to the two body pair potential approximation used in the MT method and the empirical data included in the lattice dynamics method, while in *ab initio* DFT method, no empirical data was used.

From Fig. 3(c) and Fig. 2(d), it is clear that near the Fermi level, the Ni d_{z^2} and d_{x^2} orbitals hybridize with O p_y, and p_x orbitals, and thus form chemical σ bonding, as can be seen that these Ni d and O p bands reach their DOS peaks in the energy region of -2.0 ev ~ -1.5 eV. The Ni d_{xy}, d_{yz}, and d_{xz} orbitals, on the other hand, hybridized with O p_y, p_z, and p_x orbitals at π bonded states in the energy range of both below and above Fermi level. As shown in Fig. 2 and 3, the p orbitals of O are hybridized with Zr p orbitals and d_{xy}, d_{yz}, and d_{xz} orbitals. An important effect of the Ni doping is that, as can be seen from Fig. 2(d) and 3(c), both O and Ni atom form states above and close to the Fermi level. These states can effectively serve as recombination and scattering centers when electromagnetic waves enter the system. The total magnetic moment of Ni atom is zero due to the symmetric spin up and down DOS of Ni d orbital as discussed in the last section.

Table II listed the Bader charge of Zr, O, and Ni in pure and Ni doped ZrO_2. The Bader charge of Zr cation decreases slightly from 2.64|e| in pure ZrO_2 to 2.58|e| (loses less than 0.1|e| charge) in Ni doped ZrO_2 whereas O anion increased from -1.32|e| to -0.81~-0.85|e| with an average of 0.5|e| charge gain. The substitution Ni cation has a 1.45|e| which is about one electron charge less than Zr cation. O anion charges changed due to the less charge of Ni and the Zr vacancy formation. The larger charge difference of O is consistent with the large shift and extension of DOS shown in Fig. 2(c) and 2(d).

The high Ni doping concentration and Zr vacancy cause the formation of large scale localized dipoles around vacancies. The local electric field is stronger at (111) direction than that of (011) and (001) directions as from the above Ni, Zr and O Bader charge analysis. In each unit cell there are two opposite dipoles pointing from the center of $O^{-0.8}$ to the center of $Ni^{+1.45}$ (vacancy). The electromagnetic waves thus are interacted strongly with these high density localized dipoles.

In summary, we implemented two *ab initio* DFT application methods to the ZrO_2 ceramic calculation: (1) elastic constant calculation and application to the structure stability analysis; (2) the reflectivity along special directions of the related systems. They can be essential to the systems where experimental data is not available or cannot be acquired under current conditions. In the DFT based simulations on a Ni doped ZrO_2 system, the lattice constant of pure ZrO_2 is examined first, which compares very well with the experiment data. Then the elastic constants for a number of cases are calculated, which include the pure ZrO_2, 25% Ni interstitial and substitution site doped cubic ZrO_2 crystals, and 25% Ni substitute site doped cubic ZrO_2 crystal with a Zr vacancy. The elastic constant results demonstrate that only pure ZrO_2 and 25% Ni substitution site doped cubic ZrO_2 crystal with a Zr vacancy are stable structures. Thus the reflectivity calculations are performed on these two systems only. The reflectivity calculations clearly show that the high concentration Ni doped ZrO_2 increased the reflectivity values by 3.4, 4.1, 2.6, and 2.3 times in (001), (010), (011), and (111) polarization directions respectively, with a maximal 80% reflectivity at (111) direction. This may have potential applications in high efficient TBC material design and synthesizing.

Acknowledgement: This work is supported by NASA-BoR LaSPACE DART-2 program (subcontract No. 28538) and LONI institute. The authors thank Dr. J. I. Eldridge and his group for helpful discussion.

References:

[1]. D. A. Litton, N. E. Ulion, M. F. Trubelja, M. J. Maloney, and S. G. Warrier, United States Patent No. 6,730,422 (2004).
[2]. R. Gadow and G. Schaefer, United States Patent No. 6,998,064 (2006).
[3]. J. I. Eldridge, C. M. Spuckler, and K. W. Street, Infrared Radiative Properties of Yttria-Stabilized Zirconia Thermal Barrier Coatings, 26th Annual Conference of Composites, Advanced Ceramics, Materials and Structures: B, (Cocoa Beach, Florida, 2002, American Ceramic Society, Westerville, OH), p 417.
[4]. S. Gu, T. J. Lu, D. D. Hass, H. N. G. Wadley, Acta Mater. **49**, 2539 (2001).
[5]. A. L. Ivanovskii, S. V. Okatov, and G. P. Shveikin, Inorganic Materials **36**, 1121 (2000).
[6]. H. J. F. Jansen, Phys. Rev. B **43**, 7267 (1991).
[7]. H. Yao, L. Ouyang, and W. Y. Ching, J. Am. Ceram. Soc. **90**, 3194 (2007).
[8]. M. Iuga, G. S. Neumann, and J. Meinhardt, Eur. Phys. J. B **58**, 127 (2007).
[9]. P.E. Blöchl, Phys. Rev. B **50**, 17953 (1994).
[10]. G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).
[11]. D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).
[12]. H. Nielsen and R. M. Martin, Phys. Rev. Lett. **50**, 697 (1983).
[13]. F. Nye, Physical properties of crystals, Oxford, Clarendon, 1957.
[14]. R. F. W. Bader, in Atoms in Molecules -- A Quantum Theory, Oxford University Press, Oxford, 1990.
[15]. D. E. Ellis, K. Mundim, V. P. Dravid, and J. W. Rylander, pp350-64 in Computer Aided Design of High-Temperature Materials, Oxford University Press, Oxford, U.K. 1999.
[16]. T. Hailing and G. A. Saunders, J. Mater. Sci. Lett. **1**, 416 (1982).
[17]. S. Shin and M. Ishigame, Phys. Rev. B **34**, 8875 (1986).
[18]. D. W. Liu, C. H. Perry, A. A. Feinberg, and R. Currat, Phys. Rev. B **36**, 9212 (1987).
[19]. H. M. Kandil, J. D. Greiner, and J. F. Smith, J. Am. Ceram. Soc. **67**, 341 (1984).
[20]. L. L. Boyer and B. M. Klein, J. Am. Ceram. Soc. **68**, 278 (1985).
[21]. A. P. Mirgorodsky, M. B. Smirnov, and P. E. Quintard, Phys. Rev. B **55**, 19 (1997).

Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas

Neerja Karnani, Christopher Taylor, Ankit Malhotra and Anindya Dutta

Genome Res. 2007 17: 865-876
Access the most recent version at doi:10.1101/gr.5427007

Letter

Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas

Neerja Karnani,[1] Christopher Taylor,[1,2] Ankit Malhotra,[1,2] and Anindya Dutta[1,3]

[1]*Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia 22908, USA;*
[2]*Department of Computer Science, University of Virginia, Charlottesville, Virginia 22908, USA*

In eukaryotes, accurate control of replication time is required for the efficient completion of S phase and maintenance of genome stability. We present a high-resolution genome-tiling array-based profile of replication timing for ~1% of the human genome studied by The ENCODE Project Consortium. Twenty percent of the investigated segments replicate asynchronously (pan-S). These areas are rich in genes and CpG islands, features they share with early-replicating loci. Interphase FISH showed that pan-S replication is a consequence of interallelic variation in replication time and is not an artifact derived from a specific cell cycle synchronization method or from aneuploidy. The interallelic variation in replication time is likely due to interallelic variation in chromatin environment, because while the early- or late-replicating areas were exclusively enriched in activating or repressing histone modifications, respectively, the pan-S areas had both types of histone modification. The replication profile of the chromosomes identified contiguous chromosomal segments of hundreds of kilobases separated by smaller segments where the replication time underwent an acute transition. Close examination of one such segment demonstrated that the delay of replication time was accompanied by a decrease in level of gene expression and appearance of repressive chromatin marks, suggesting that the transition segments are boundary elements separating chromosomal domains with different chromatin environments.

[Supplemental material is available online at www.genome.org.]

Although all the DNA in a eukaryotic cell replicates during the S phase of cell cycle, there is a great variability in the actual point in S phase when a given chromosomal segment replicates. Segments are known to reproducibly replicate early or late in S phase, and it is generally believed that this is determined by the time at which the origins in a segment fire. All origins of replication are licensed with MCM proteins by the time S phase begins (Bell and Dutta 2002), and yet, once conditions in the cell change to favor the firing of the origins, all origins do not fire at the same time. In situ labeling techniques and other methods have led to some general principles determining the time of replication of a segment in S phase (for review, see MacAlpine and Bell 2005). Early-replicating segments are generally enriched in euchromatin, while late-replicating segments are enriched in heterochromatin. Some loci that are selectively expressed in specialized cells (e.g., immunoglobulin, beta-globin, or neural-associated genes) show a change in time of replication from late-S phase in undifferentiated, nonexpressing cells to early-S phase after differentiation (Simon et al. 2001; Zhou et al. 2002; Perry et al. 2004). The correspondence between the activation of chromatin at differentiation-induced genes with the advancement in replication time also suggests that the chromatin environment dictates time of replication (Bickmore and Carothers 1995; Rountree et al. 2000; Demeret et al. 2001).

The completion of many genomic sequences and the advent of genome-tiling microarrays provided an opportunity to correlate gene expression or chromatin structure with time of replication at a much finer resolution. DNA replicated at specific intervals in S phase were hybridized to genome-tiling microarrays to determine the exact time in S phase when specific genes replicate. Early experiments in model organisms like *Saccharomyces cerevisiae* and *Drosophila melanogaster* confirmed many of the principles outlined above (Raghuraman et al. 2001; Schubeler et al. 2002; MacAlpine et al. 2004).

Extending this method of analysis to human cells, specifically to Chromosomes 21 and 22, we confirmed that similar principles dictate time of replication in human chromosomes (Jeon et al. 2005). We made the surprising observation that almost 60% of the chromosomal probes studied gave a replication signal at multiple times in S phase, described as a pan-S-phase pattern of replication. While asynchrony of replication between alleles of a given gene would give rise to a pan-S-phase pattern of replication, it seemed highly unlikely to us that 60% of the human chromosomes would show such asynchrony. In addition, it was unclear whether the pan-S-phase replication was an artifact of cells losing their synchrony of progression through the cell cycle, of the thymidine-aphidicolin method of cell cycle synchronization, or of the aneuploidy inherent in HeLa cells.

The ENCODE region encompasses 44 segments covering ~1% of the human genome on which multiple groups are applying different techniques to find the best methods to annotate the human genome (The ENCODE Project Consortium 2004). We measured the replication time for this region and used the data to improve our method of computing the replication profile of chromosomal segments. The improvements in our algorithm decreased the pan-S replication pattern to ~20% of the segments interrogated. We confirmed the prediction of pan-S replication

[3]Corresponding author.
E-mail ad8q@virginia.edu; fax (434) 924-5069.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.5427007.
Freely available online through the *Genome Research* Open Access option.

by an independent method of assessing replication time: interphase FISH. The results demonstrate that pan-S-phase replication is a real pattern of replication that cannot be explained by artifacts derived from microarray platform, methods of cell cycle synchronization, or aneuploidy of cells. Instead, pan-S-phase replication is a reflection of asynchrony of replication between alleles in a given cell, suggesting that differences in the chromatin environment of two alleles can be seen in up to 20% of the human genome in some cells.

Finally, using the high-definition temporal profile of replication over the ENCODE areas, we identified adjoining chromosomal segments of a few hundred kilobases each with differing times of replication. Hypothesizing that these areas are "replication domains," we demonstrate for one such region that the adjoining domains have different levels of gene expression and activating and repressing marks on histones. We believe that the replication domains correspond to chromosomal domains separated by boundary elements.

Results

Replication timing analyses of 1% genome using synchronized HeLa cells

HeLa cells were synchronized at G_1/S by thymidine-aphidicolin block. After release from the block, cells were pulsed with bromodeoxyuridine (BrdU) at every 2-h interval of S-phase and genomic DNA isolated. In all, five time intervals (0–2, 2–4, 4–6, 6–8, and 8–10 h) representing 10 h of the entire S phase were collected. The BrdU-incorporated heavy/light (H/L) DNA was purified using a CsCl density gradient as described earlier (Jeon et al. 2005). Purified DNA from each time interval was hybridized to the high-density genome-tiling Affymetrix array comprising unique 25-mer oligonucleotides in the ENCODE-selected chromosomal loci covering 1% of the human genome (~30 Mb) (see Methods for details of the ENCODE regions).

Segregation of chromosomal regions into temporally specific and pan-S replicating segments

Probes that replicated in a discrete interval in S phase were called temporally specific, while probes that replicated at multiple intervals in S phase were called temporally nonspecific. The Methods and Supplemental Table 1 contain examples of the specificity classification. For the Affymetrix ENCODE array, 26.115% of the probes were temporally nonspecific.

In order to classify chromosomal segments as temporally specific or asynchronously replicating (pan-S), a 10-kb sliding window was passed along the chromosome and each window defined as replicating in a pan-S manner if >60% of the probes in that window are temporally nonspecific (see Methods for details). Thus, by ensuring that the majority of contiguous probes in a given segment replicate in a temporally nonspecific manner, we eliminate artifacts from cross-hybridization or from poor probe hybridization. Since the estimated average speed of a replication fork is ~1 kb/min, isolated segments <10 kb (<10 min) that appeared to replicate in a nonspecific manner were significantly below the resolution of the 2-h sampling method. Such segments (<0.2% of the ENCODE region) were therefore eliminated from our calculations. After these corrections, ~20% of the ENCODE area replicated in a pan-S-phase pattern as determined by a base-pair count (Fig. 1A), while the remaining 80% shows a temporally distinct profile. Individual chromosomal segments showing these patterns are presented below.

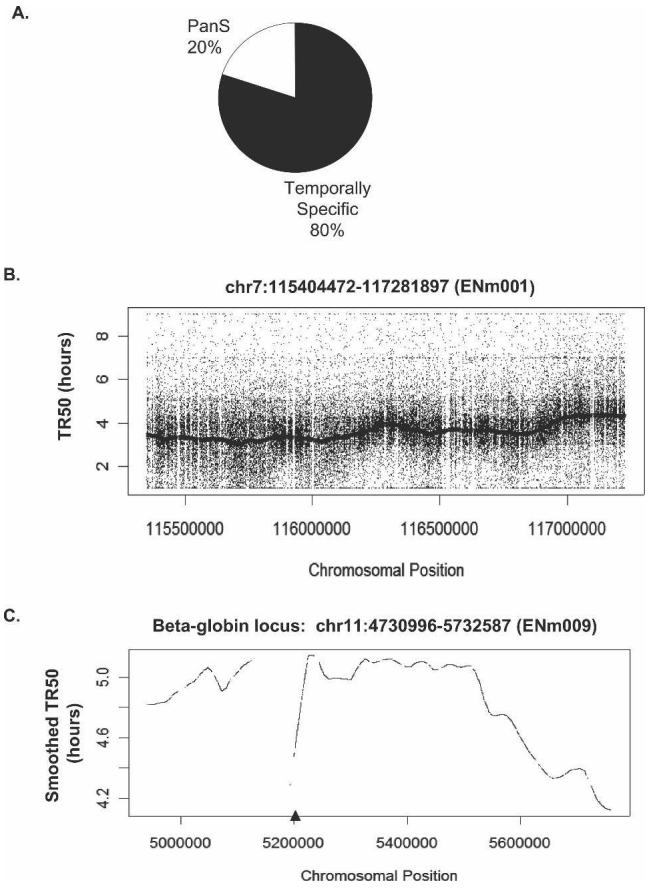


Figure 1. Temporal profile of replication of chromosomal segments. (*A*) Temporally specific versus pan-S distribution of replication for 1% of the human genome investigated in this study. (*B*) Raw TR50 data with a smoothed TR50 curve overlaid from the 1.9-Mb region on Chromosome 7. According to the ENCODE Consortium nomenclature, this chromosomal segment is referred to as ENm001 (http://hgwdev.cse.ucsc.edu/ENCODE/encode.hg17.html). (*C*) Smoothed TR50 data from the 1-Mb beta-globin locus (ENm009) on Chromosome 11. The lowest point in each valley indicates a site that is replicated before its adjoining segments and thus is likely to contain origins of replication. The gaps in the TR50 plots indicate the presence of repeats. In order to minimize cross-hybridization of oligonucleotides, repeat regions of the genome are not spotted on the tiling arrays. The triangle on the *X*-axis indicates the position of the known beta-globin origin.

Continuous TR50 profile along the length of a chromosomal segment

The time at which a temporally specific probe replicates to 50% (TR50) is calculated by summing the replication signal over the five time points and linearly interpolating the time when 50% of the total signal was reached. Supplemental Table 1 gives examples of TR50 calculation for several probes. Plotting the TR50 values for specific probes against the linear coordinate of the probe on the chromosome gives a view of the replication profile of the chromosome. Because the raw TR50 data are noisy, Lowess smoothing for all temporally specific probes in a 60-kb window was performed to ascertain the trends in the replication pattern along the length of the chromosome. Figure 1B shows the raw TR50 data and a smoothed TR50 curve for a 1.9-Mb segment of Chromosome 7 (ENm001). By averaging over relatively long segments of DNA, the smoothed curve corrects for scatter created by

differences in probe hybridization efficiency or from cross-hybridization of a few errant probes and is very useful for comparing the time of replication of adjoining segments of DNA. Figure 1C shows another example of a smoothed TR50 plot from a 1-Mb region of Chromosome 11 containing the beta-globin locus. The late replication of this segment in HeLa cells agrees with previous findings that the beta-globin locus replicates late in S phase in nonerythroid cells (Epner et al. 1988; Dhar et al. 1989). The TR50 profiles for all the 43 chromosomal loci can be viewed using the UVa DNA Rep TR50 track at the http://hgwdev.cse.ucsc.edu/ENCODE/encode.hg17.html site. Temporal profiles from 12 of these regions are shown in Supplemental Figure 1.

Local minima of the TR50 curve show areas that replicate earlier than the flanking regions and thus are likely to contain origins of replication, as has been shown previously in *S. cerevisiae* (Raghuraman et al. 2001). Only one previously validated origin of replication lies in the ENCODE area near the large stretch of repeat sequences (chr11: 5124929–5193780) within the beta-globin locus (Kitsberg et al. 1993; Aladjem et al. 1998; Wang et al. 2004). The repeat sequences near the beta-globin gene were not represented on the microarray, causing a gap in the TR50 profile (Fig. 1C). However, the TR50 profile of the regions immediately adjoining these repeats clearly suggests that a minimum in the TR50 profile is located somewhere at or near these repeats, indicating the presence of an origin of replication at this site. Thus, the hundreds of minima in the TR50 profile are likely to be at or near origins of replication.

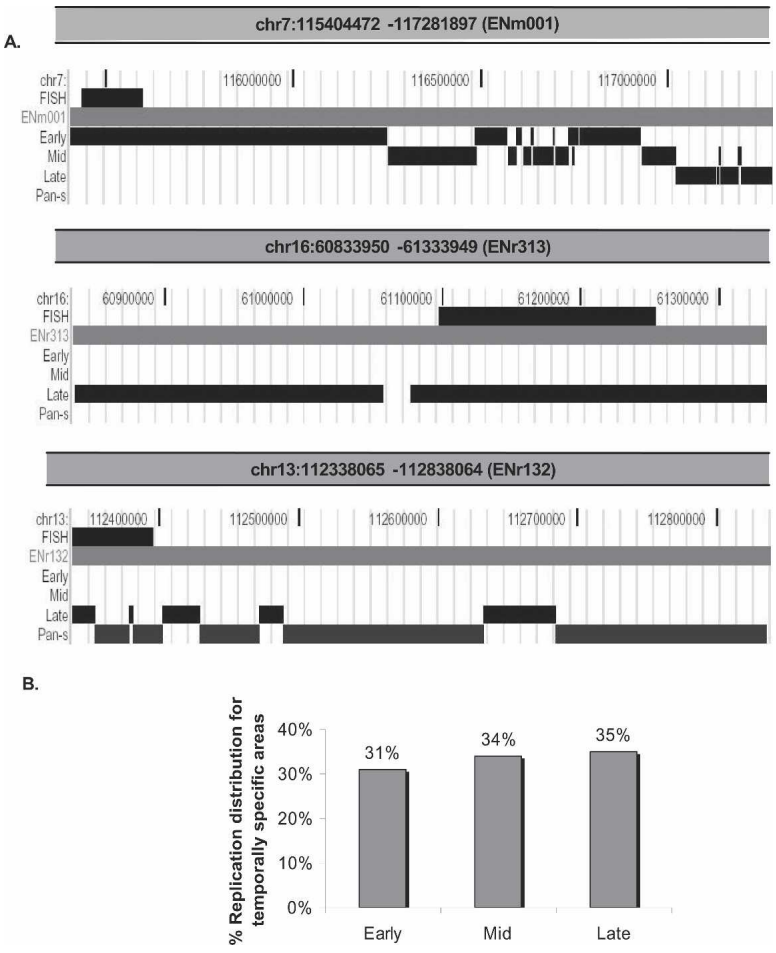


Figure 2. Segregation of chromosomal segments with temporally specific and temporally nonspecific pattern of replication. (*A*) On the basis of TR50, temporally specific regions are further segregated and displayed as three tracks (UCSC Genome Browser), early-, mid-, or late-replicating, while chromosomal regions undergoing temporally nonspecific replication are highlighted under the pan-S track. The three panels in this figure show segregation of replication timing for three chromosomal segments. (*Top* panel) The 1.9-Mb region (ENm001) of Chromosome 7; (*middle* and *bottom* panels) examples of two 500-kb chromosomal segments from Chromosomes 16 (ENr313) and 13 (ENr132) that underwent late and pan-S replication, respectively. The FISH track in all the three panels refers to the chromosomal positions of BAC clones selected for the interphase FISH experiment shown in Figures 3 and 4. (*B*) Percent of temporally specific chromosomal segments replicating in early-, middle-, or late-S phase.

Segregation of temporally specific regions into early-, mid-, and late-S replicating regions

The smoothed TR50 profile suffers from a compression of the *Y*-axis values due to the smoothing operation; thus we do not get an accurate estimate of the time of replication of a given segment from the profile. We therefore processed the TR50 data to define discrete segments with early-, mid-, and late-S-phase replication in addition to the pan-S-phase replication patterns described above. A temporally specific region is classified into early, mid, or late replication based on the average TR50 of the temporally specific probes within a 10-kb window. TR50 cutoffs of 3.4 h (for early- to mid-S transition) and 3.9 h (for mid- to late-S transition) are used.

The top panel of Figure 2A shows the segregation of ENm001 after these analyses. Tracks representing segments that replicate in early-, mid-, late-, or pan-S-phase, respectively, are indicated. Since ENm001 is an early-replicating region, only a very small region shows up in the late-replication track. The general trend of the right portion of the region replicating later can be seen in the transition from a solid early-replication track into mid-replicating regions as we move left to right. The tracks are nonoverlapping at the base-pair level, and the apparent overlap in certain places is due to the low resolution of the UCSC Browser snapshot required to fit the whole region into a figure.

The second and third panels of Figure 2A show similar segmentation of 500-kb chromosomal regions from Chromosomes 16 (ENr313) and 13 (ENr132), respectively. ENr313 replicates late, while ENr132 shows a pan-S pattern of replication. TR50 segmentation profiles for 12 regions are shown in Supplemental Figure 1. Profiles for all the 43 regions can be viewed in the UVa DNA Rep Seg track at http://hgwdev.cse.ucsc.edu/ENCODE/encode.hg17.html site. Eighty percent of the ENCODE area replicates in a temporally specific interval (Fig. 1A). Within the spe-

cific regions, 31% segregates into early-, 34% into mid-, and 35% into late-S-phase replicating patterns (Fig. 2B).

Validation of replication time by interphase FISH

To check the temporal profile of replication generated by the microarray data, we used interphase FISH as an independent method for determining replication time. Although labor-intensive, this method has the additional advantage in that the large sizes of the probes reduce errors from poor signal strength and cross-hybridization. Ten BAC clones of 48–187 kb each (details in Supplemental Table 2) were selected to validate the microarray data for 10 segments from nine ENCODE areas: three each with early and pan-S-phase and four with late-S-phase patterns of replication. The positions of BAC clones used in Figure 3 and Figure 4 are highlighted in Figure 2A. HeLa cells were synchronized and harvested at 2-h intervals during S phase, and BAC clones were labeled and hybridized to denatured interphase nuclei. A single hybridization signal (visible as a dot under the microscope) indicates one copy of the targeted DNA. ENm001 showed 2 dots/cell in G_1 (0 h) and 4 dots/cell in G_2, while the remaining eight regions had 3 dots/cell in G_1 and 6 dots/cell in G_2 because of the aneuploidy of HeLa cells. The percent replication of a probed segment in each time interval in S phase was determined by counting the increments in dots/cell during that interval, where 100% replication means that the number of dots/cell is twice the G_1 value.

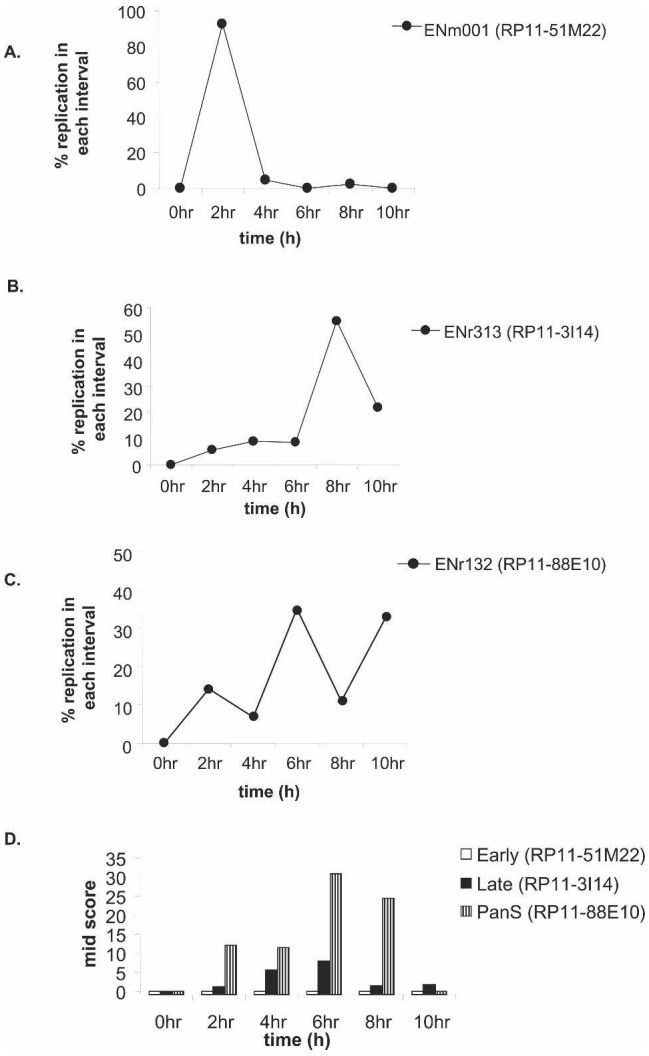


Figure 3. Interphase FISH for validating replication timing in HeLa. (*A–C*) Synchronously progressing HeLa cells were hybridized to fluorescence-labeled probes of BAC clone DNA RP11-51M22, RP11-3I14 (for early- and late-replicating areas, respectively) and RP11-88E10 (for pan-S pattern of replication). The chromosomal locations of these BACs are highlighted in Figure 2A. The percent replication at each interval of S phase is plotted against time in S phase. (*D*) The interallelic variation in replication for FISH data observed for each of the BAC clones mentioned above was determined by calculating the Mid-Score (detailed in Results and Supplemental Material).

The RP11-51M22 probe shows that this region of ENm001 replicates early with the complete doubling of all signal in the first 2 h of S phase (Fig. 3A). For a late-replicating region, RP11-3I14 from ENr313, the increase in dot number was maximum in the last 4 h of S phase (Fig. 3B). RP11-88E10 from the ENr132 region indicated that significant replication occurred in multiple time intervals (Fig. 3C), consistent with the pan-S replication detected in the microarrays (Fig. 2A). All the 10 segments tested by FISH reproduced the microarray data for time of replication (see Supplemental Table 2 for details).

Pan-S replication is due to interallelic variation in time of replication

To ascertain whether the pan-S-phase pattern of replication was due to intercellular or interallelic variation in replication time, we calculated the percent nuclei in mid-replication. The Mid-Score for a time point is defined as the percentage of cells in mid-replication, having replicated one, but not all alleles, for a given probe. Thus cells in mid-replication will have 3 dots/cell for ENm001, and 4 or 5 dots/cell for ENr313 or ENr132. Segments that replicate synchronously in a narrow interval of S phase are expected to have a very narrow temporal window with a high Mid-Score (a more detailed explanation is in Supplemental Fig. 2). ENm001 (early replicating) had no time point with a high Mid-Score, while ENr313 (late replicating) had only two time points with a Mid-Score of 5.6% and 7.9%, indicating that all the alleles replicated in a narrow time window (Fig. 3D).

If pan-S-phase replication is due to intercellular variation in time of replication of the chromosomal segment, the two alleles in a cell will still replicate simultaneously so that the window of time when a cell is caught in mid-replication will remain short. Mid-Scores would be low or elevated for only a tightly restricted time interval (Supplemental Fig. 2). However, the ENr132 region (pan-S-phase replication) showed four time points with high Mid-Scores (i.e., 12.1, 11.4, 30.7, and 24.5) (Fig. 3D), suggesting that there was significant asynchrony in the time of replication of the alleles in a given cell. Thus the asynchrony in replication seen in the pan-S-phase pattern of replication is due to interallelic variation in replication time.

Pan-S-phase pattern of replication is not due to thymidine-aphidicolin block

We next investigated whether pan-S-phase replication was caused by the prolonged arrest in S phase that is inherent to the thymidine-aphidicolin double-block method of synchronization of cells in the cell cycle. HeLa cells were synchronized in mitosis using nocodazole and released. The time of replication was determined by interphase FISH for five regions (Fig. 4B): three that were temporally monophasic and two that had a pan-S-phase

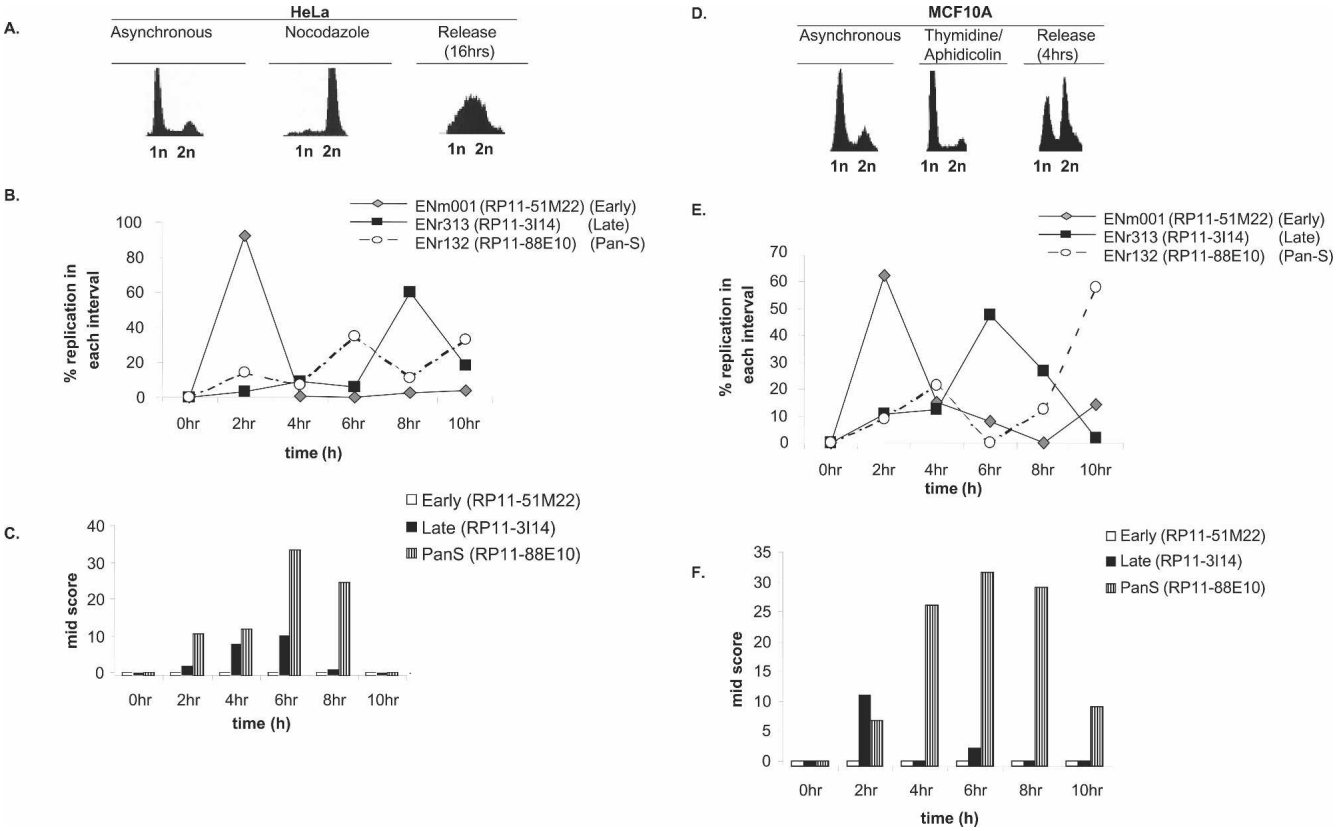


Figure 4. Pan-S replication pattern is independent of cell synchronization method and aneuploidy. (*A*) HeLa cells blocked (by nocodazole) and released from mitosis followed by FACS for DNA content. (*B,C*) Interphase FISH was performed with HeLa cells synchronized with nocodazole and released. The *X*-axis represents time in S phase such that 0 h = 12 h post-release from the nocodazole block. The rest is as in Figure 3. (*D*) MCF10A cells released from a G_1/S block with thymidine/aphidicolin followed by FACS for DNA content. (*E,F*) Interphase FISH with MCF10A cells synchronously progressing through S phase to determine the replication profile and Mid-Score with the chromosomal segments mentioned in Figure 3.

pattern of replication. The temporally specific segments still replicated in the expected time frames despite the different method of synchronization (Fig. 4B; Supplemental Table 2). Most important, both the pan-S-phase regions continued to replicate at multiple times in S phase (Fig. 4B; Supplemental Table 2), suggesting that pan-S-phase replication was not an artifact of the synchronization method. The observed asynchrony in replication was due to interallelic variation as determined by the wide time interval when the cells displayed high Mid-Scores (Fig. 4C).

Pan-S pattern of replication is not restricted to aneuploid HeLa cells

We wanted to rule out the possibility that the pan-S-phase pattern of replication is seen only in aneuploid cancer cells like HeLa. To address this, we repeated the interphase FISH experiments with MCF10A, a breast epithelial cell line derived from fibrocystic breast disease that is near diploid and nonmalignant (Fig. 4D). The area covered by probe RP11-88E10 (a region with pan-S-phase replication in HeLa cells) replicated at two time intervals (Fig. 4E). The first peak at 4 h corresponded with the time interval during which the Mid-Score increased (Fig. 4F). The Mid-Score remained high until the 10-h time interval, when the second peak of replication was observed, indicating a significant time lapse in the replication of two alleles. Therefore, pan-S-phase replication is also seen in MCF10A cells and is not unique to HeLa cells. Replication of RP11-51M22 (early) and RP11-3I14 (late) was also consistent with that seen in HeLa cells. FISH analyses for two more regions in MCF10A are detailed in Supplemental Table 2.

Correlation of TR50 profile with genome sequence features

The replication timing for the 43 ENCODE regions were correlated against genome sequence features such as AT content, CpG islands, and gene density. AT content was computed using a 10-kb sliding window and plotted against the smoothed TR50 curve. A transition from low to high AT content is evident for early- to late-replicating regions (Fig. 5A). The Spearman rank correlation coefficient calculated from the plot was 0.257, suggesting a moderate positive correlation. The Pearson correlation coefficient was 0.252, also indicating a moderate positive correlation. Computation of AT content at a window size of 1 kb gave a lower correlation coefficient (0.19).

DNA methylation is an important epigenetic marker (Jones and Takai 2001), with differential DNA methylation between alleles leading to monoallelic gene expression, interallelic differences in the chromatin, and asynchronous replication (Simon et al. 1999; Rountree et al. 2001; Fournier et al. 2002; Jiang et al. 2004; Fuks 2005). Since the Mid-Score calculations above suggested that the pan-S areas demonstrated interallelic differences in replication, we wondered whether the pan-S replicating segments were enriched in CpG islands and thus potentially suscep-

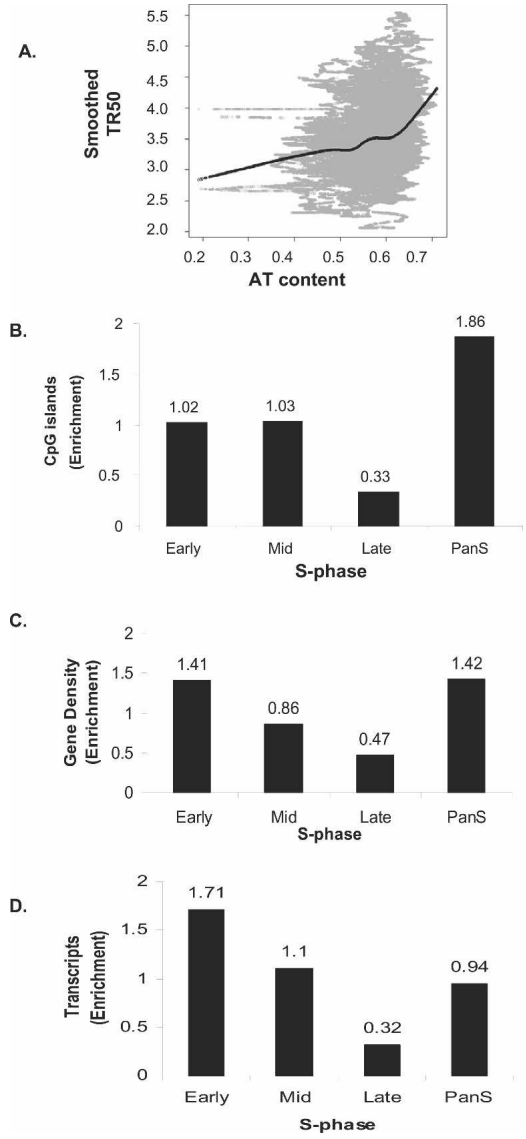


Figure 5. Correlation between replication time and genomic features. (*A*) Plot of smoothed TR50 against AT content in a 10-kb sliding window. Lowess smoothed curve done at $f = 0.3$ (fraction of the data included in the running local fit) is overlaid in black to show the general trend. (*B–D*) Histograms showing distribution of (*B*) CpG islands, (*C*) gene density, and (*D*) transcripts (HeLa cells) against temporal segregation of replication.

tible to regulation by differential DNA methylation. Indeed, the pan-S-phase regions showed the maximum enrichment (1.86) of CpG islands (Fig. 5B).

We next compared the replication time of a segment with its gene density. In the chromosomal areas where replication was temporally specific, a threefold higher enrichment of genes was found in regions replicating early compared to those replicating late (Fig. 5C). Interestingly, the pan-S-phase regions had gene content (enrichment = 1.42) (Fig. 5C), comparable to early-replicating chromosomal segments (enrichment = 1.41), consistent with the idea that these regions could have replicated early if not for interallelic variation in chromatin structure that resulted in a subset of the alleles replicating late and producing a pan-S-phase pattern of replication.

Early-replicating regions are highly transcribed

Active transcription of genes is associated with euchromatin and may be expected to correlate with early replication. Total RNA was prepared from logarithmically growing HeLa cells and hybridized to an Affymetrix HG-U133 Plus 2.0 array to measure the level of expression of genes in different chromosomal segments. Early-replicating segments have 5.34-fold higher transcription over the late-replicating regions (Fig. 5D). The pan-S regions had an intermediate level of gene expression, consistent with the idea that all alleles of the genes in these segments are not in favorable chromatin and are not uniformly well expressed.

TR50 profile on one chromosomal segment defines chromosomal domains

The global correlations described above are consistent with the hypothesis that early-replicating regions are usually gene dense and contain actively transcribed genes. The fine resolution of replication profile possible with the genome-tiling arrays allowed us to closely examine how such correlations hold up across contiguous stretches of chromosomes. Intriguingly, the TR50 profile of some regions revealed the presence of neighboring chromosomal segments with acute transitions in replication time. For example, in ENm005, an ~366-kb (Chr21: 33119705–33486048) late-replicating stretch was bracketed by two early-replicating areas (Fig. 6A). Dual color interphase FISH was performed to confirm the transition in replication time from early to late in two neighboring segments of ENm005 (Fig. 6B). BAC clones (separated by ~355 kb) from the early (RP11-54F16) and late (RP11-79D9) replicating areas confirmed that the two DNA segments, indeed, replicated in two different intervals of S phase (Fig. 6B).

The replication dissimilarities between the adjoining domains correlated with dissimilarities in gene expression and gene density (Fig. 6C). The late-replicating island was both gene-poor and transcriptionally less active compared to the adjoining early-replicating chromosomal segments.

These observations suggested the existence of two chromatin environments in a contiguous stretch of a chromosome separated by some type of boundary element. Since histone modifications distinguish euchromatin from heterochromatin, we decided to confirm the existence of two chromatin environments in this locus in HeLa cells by performing a chromatin immunoprecipitation (ChIP) assay for the active and inactive chromatin marks. H3 Lys4 methylation is specific for active chromatin at active promoters (Bernstein et al. 2005). We therefore selected nine genes, two in the late-replicating region (*OLIG1* and *OLIG2*) and seven in the adjoining early-replicating chromosomal segments (*C21orf119*, *SYNJ1*, *C21orf66*, *IFNAR1*, *GART*, *ITSN1*, and *ATP5O*) and designed primers to amplify unique 100–300-bp fragments from the 2-kb sequences upstream of the genes (see details for primers in Supplemental Table 3). Chromatin immunoprecipitation (ChIP) and amplification of these promoters revealed that all seven genes in the early-replicating segments were positive for H3 lysine 4 (H3K4) methylation, while the two embedded in the late-replicating environment (005HM4 and 005HM5) lacked this modification (Fig. 6D). Conversely, ChIP for markers of repressed chromatin, H3 lysine 9 (H3K9) dimethylation and association of HP1α, showed that the two promoters in the late-replicating domain were in repressed chromatin. Five out of the

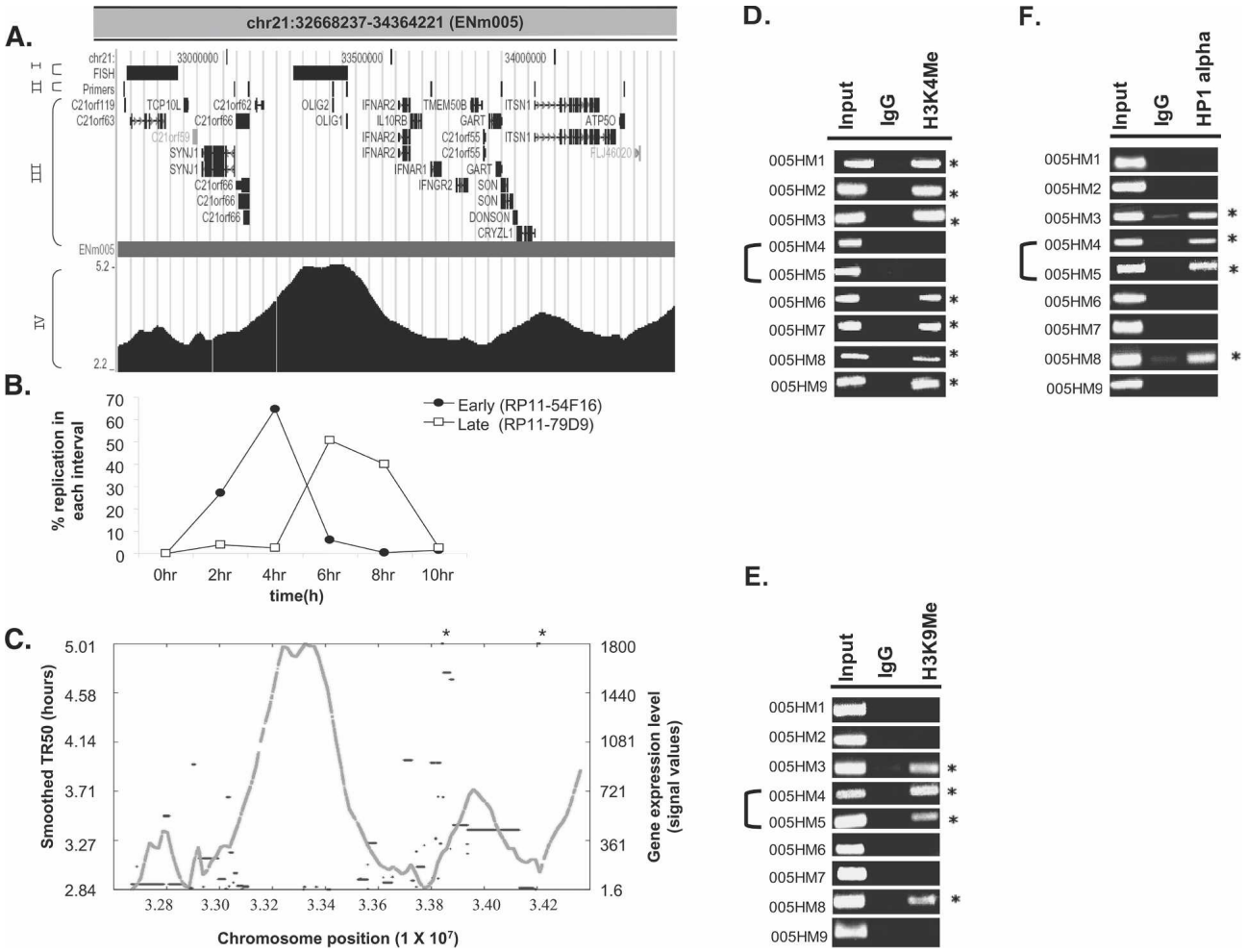


Figure 6. Replication profile demarcates chromosomal domains. (*A*) UCSC Genome Browser display of a 1.7-Mb region from Chromosome 21 (ENm005). This Browser picture highlights four tracks (I–IV): (I) FISH: chromosomal location of BAC clones (RP11-54F16 and RP11-79D9, from *left* to *right*) selected for the interphase FISH experiment shown in *B*; (II) Primers: chromosomal locations of the primers (005HM1–9, *left* to *right*) selected for ChIP assay to ascertain the histone modifications and HP1α-binding sites shown in *D–F*; (III) RefSeq: positions of all the genes in this chromosomal segment; and (IV) the contiguous TR50 profile. (*B*) Dual color FISH was performed with HeLa cells synchronously progressing through S phase. RP11-54F16 (from early-replicating area on *left*) was labeled with spectrum red dUTP, while RP11-79D9 (from late-replicating area) was labeled with spectrum green dUTP. Dual color FISH with these two BAC clones ascertained the replication time of the two regions of Chromosome 21 set 355 kb apart. (*C*) Plot of smoothed TR50 (*Y*-axis on *left*, gray) against level of transcription of genes (*Y*-axis on *right*, black). The two asterisk marks represent transcripts whose transcription levels exceeded the *Y*-axis limit (i.e., 2346 and 10,010 for *left* and *right* asterisks, respectively). (*D–F*) ChIP-PCR assay across ENm005 region (see Supplemental Table 3 for primers). PCR was performed on DNA chromatin immunoprecipitated with antibodies against methylated histones (H3 Lys4 and H3 Lys9) and HP1α. (Input) DNA control before immunoprecipitation; (IgG) ChIP with rabbit IgG was negative control for nonspecific precipitation. Forty cycles of PCR were performed for H3 Lys4 and HP1α and 30 cycles for H3 K9 di-Me. The asterisks refer to primer pairs that gave positive ChIP signal for the indicated antibodies relative to the IgG negative control. 005HM4 and 005HM5 were from the late-replicating island in ENm005.

seven promoters in the early-replicating segments were negative for markers of repressed chromatin, while the other two were positive (Fig. 6E,F).

Therefore, the island of late-replicating DNA represents a specific chromosomal domain with all the features of heterochromatin: low gene density, low gene expression, lack of activating chromatin marks, and presence of repressive chromatin marks. The rapid transition of the features of heterochromatin in this late-replicating island to those of euchromatin in the flanking areas suggests that the chromosome may be divided into discrete domains with different chromatin features. In addition, the existence of such discrete adjoining domains with different chromatin structure suggests the presence of boundary elements that prevent the spread of euchromatin from the neighboring areas to this island of heterochromatin.

Pan-S segments contain markers for both active and repressed chromatin

The interallelic variation in replication time observed in pan-S replicating segments predicts that one allele will be in active chromatin and another in repressed chromatin, leading us to test whether pan-S replicating segments are enriched in both types of marks. ENr132 contained extensive stretches with the pan-S replication pattern with a few interspersed segments that were exclusively late replicating. The two promoters in the pan-S repli-

cating area, 132HM1 and 132HM2, were positive by ChIP for both the activating histone modification (H3K4 methylation) and repressive histone modification (H3K9 dimethylation) and a marker for heterochromatin (HP1) (Fig. 7). In contrast, 132HM3, from a late-S replicating segment only carried the repressed chromatin marks and not the activating histone modification. Therefore, combining the data in Figure 6 and Figure 7, three out of three late-replicating promoters were exclusively in repressed chromatin, and five out of seven early-replicating promoters were exclusively in activated chromatin. In contrast, the promoter from the pan-S replicating segment carried marks of both active and repressed chromatin, consistent with the pan-S replication pattern arising from interallelic variation in the chromatin environment.

Discussion

Since the ENCODE project specifically selected the target 1% of the genome to be broadly representative of the whole genome based on criteria like gene density and sequence conservation, we expect that the lessons learned from these high-resolution replication time profiles can be extended to the entire genome. The pan-S-phase pattern of replication; the correlation of replication time with chromatin modifications, gene expression, and AT content; and the significance of chromosomal domains and boundary elements revealed by our studies are discussed here.

We still identify regions that replicate in multiple times in S phase in mammalian cell lines (pan-S replication pattern). Since the genome-based studies of replication in *S. cerevisiae* were executed only in haploid strains, they were not expected to identify regions with interallelic difference in time of replication (Raghuraman et al. 2001). Genome-based studies of replication in diploid organisms were also unsuitable to identify this pattern because of the study design (MacAlpine et al. 2004; Woodfine et al. 2004). In those studies, the time of replication was assessed by determining the ratio of DNA content for a locus in late-S (or G_2) cells compared to G_1 cells. In such experiments, segments showing replication in both early- and late-S phase would appear to replicate in mid-S phase, and the pan-S pattern would be missed. In contrast, the sampling of cells in multiple intervals in S phase and the use of a more sensitive method of detecting replication dependent on a positive selection for BrdU-labeled DNA enabled us to identify chromosomal segments that replicate in multiple intervals in S phase.

In this study, 20% of the studied genome appeared to replicate asynchronously, a value that is one-third that of our previous analysis on Chromosomes 21 and 22 (Jeon et al. 2005). This difference is due to an important refinement in the method of analysis in the present study. In the previous work, the hybridization data from genome-tiling arrays was analyzed by the standard Affymetrix GTRANS software to generate a track that showed when the replication signal from a given time point was significantly enriched over signal obtained from DNA replicated for the entire duration of S phase. Although this method provided an intuitive belief for replication timing, not surprisingly, replication signal was not only seen in the time period when the locus replicated but lower levels of signal were seen in adjoining time intervals. The presence of a signal in multiple time tracks led us to overestimate that nearly 60% of sequences showed a pan-S replication pattern (Supplemental Figure 3, ENm001). In contrast, in this study, we segregate probes into those that are temporally synchronous versus temporally asynchronous by quantitative criteria that take into account the spillage of replication signal into adjoining time points. In addition, only large contiguous DNA segments (≥10 kb) containing >60% of probes with asynchronous replication signals are classified as pan-S. This prevents mis-calling as pan-S short stretches where low signal strength or cross-hybridization from isolated probes give an apparent replication signal in multiple intervals in S phase. As is evident from the comparison of the two methods in one segment (Supplemental Fig. 3), the present method gives a more conservative estimate of segments that

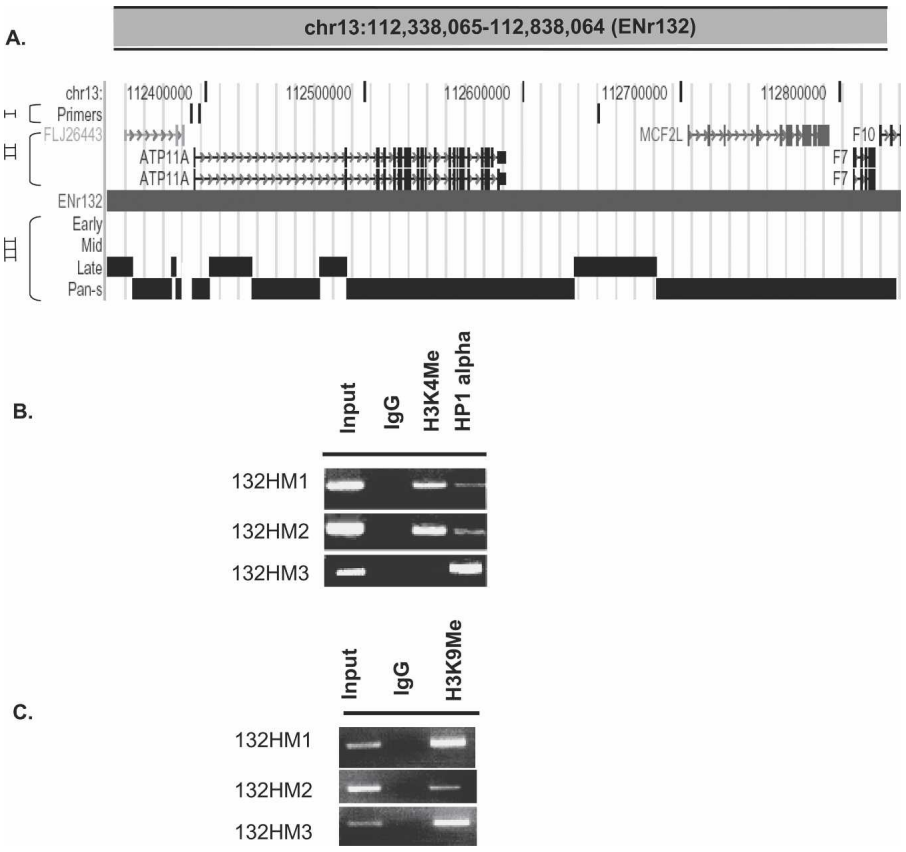


Figure 7. Both active and repressive chromatin marks are present in a pan-S segment. (*A*) UCSC Genome Browser display of a 500-kb region from Chromosome 13 (ENr132). This Browser picture highlights three tracks: (I) Primers: ChIP-PCR primers (132HM1–3, *left* to *right*) to study histone modifications and HP1α-binding sites; (II) RefSeq: positions of all the genes in this chromosomal segment; and (III) the temporal segregation of replication data. (*B,C*) ChIP-PCR assay across ENr132 region (see Supplemental Table 3 for primers) against methylated histones (H3 Lys4 and H3 Lys9) and HP1α (as indicated).

replicate at multiple times in S phase. Because microarray-based profiling of replication is a relatively new approach, we also validated the time and pattern of replication for some of the segments by a completely independent method, interphase FISH. The confirmation of all three pan-S regions as replicating asynchronously adds to the confidence that ~20% of the chromosomal segments in HeLa cells, indeed, show this unexpected pattern of replication.

All 10 regions tested by interphase FISH (including the temporally specific regions) reproduced the time of replication estimated by the microarray-based replication profile. In addition, the time of replication for five of five tested chromosomal regions remained unaltered when a different cell cycle block method was used in HeLa cells. Interphase FISH allowed us to check the time of replication of the same five regions in another cell line, MCF10A, where we found replication times of 3/5 chromosomal segments to match that of HeLa. The differences at the other two loci are likely due to differences in the chromatin environment of these loci in the two cell lines. Since MCF10A cells are near-diploid and untransformed, the detection of pan-S-phase replication in these cells indicates that pan-S replication is not an artifact arising exclusively from the aneuploidy or the transformed state of HeLa cells. It is, of course, entirely possible that aneuploidy or cell transformation increases the fraction of the genome that shows pan-S replication.

Since FISH-based methods analyze replication in the context of individual nuclei, the Mid-Scores showed that the asynchrony in replication time was due to interallelic difference in replication. Homologous alleles usually replicate synchronously in S phase, but there are some notable exceptions to this general rule. In humans, examples of such exceptions include monoallelically expressed genes such as those imprinted depending on parent of origin (Simon et al. 1999), genes encoding olfactory receptors (Chess et al. 1994), genes on the female X-chromosome (Avner and Heard 2001; Boumil and Lee 2001), and immunoglobulin and T-cell receptor genes (Mostoslavsky et al. 2001). We will test in the future whether all the pan-S segments express all their genes monoallelically. The interallelic asynchrony in replication in the pan-S segments suggests that one allele is in euchromatin and the other in heterochromatin. Consistent with this, pan-S areas are unique in being enriched in both activating and repressive marks (Fig. 7), with the different marks residing presumably in the two different alleles.

Since the HeLa cell line is of female origin (XX), the inactivation of one of the X-chromosomes predicts that segments from the X-chromosome should replicate in a pan-S manner, unless the long passage and aneuploidy of these cells have disrupted such inactivation. There are two regions from the X-chromosome included under ENCODE (Supplemental Fig. 4). The 1.2-Mb ENm006 region had three areas of pan-S replication (126 kb, 62 kb, and 10 kb), one of which contained the Glucose-6-phosphate dehydrogenase (*G6PD*) gene, which is known to be transcriptionally repressed on the inactivated X-chromosome and delayed in replication compared to its active counterpart (Hansen et al. 1996). The second region, ENr324 (ChrX: 122,507,850–123,007,849), contained no pan-S replicating segments. Thus the survey of the X-chromosome fragments for pan-S replication gave mixed results. The lack of pan-S replication over the entire stretch of X-chromosome in HeLa cells could not only be due to transformation and long-term culture affecting inactivation, but also because ENm006 and ENr324 contain blocks of genes that normally escape X-chromosome inactivation, similar to many reported X-linked genes (Chang et al. 1990; Disteche 1995; Miller et al. 1995; Carrel et al. 1996; Vermeesch et al. 1997).

Correlation of gene expression with time of replication in eukaryotes has produced contradictory results. In *S. cerevisiae*, the expression of genes did not correlate with their time of replication in S phase. In contrast, in cultured *Drosophila* cells, there was a positive correlation between early replication and gene transcription (MacAlpine et al. 2004). In mammalian cells, housekeeping genes like *Hprt*, histones, beta-tubulin, actin, and rDNA are ubiquitously expressed and replicated in the first half of S phase. On the other hand, tissue-specific genes such as those coding for Factor IX, fibronectin, and myosin heavy chain replicate late in the cell lines not expressing them (Holmquist et al. 1982; Iqbal et al. 1984; Goldman 1988). The previous study from our laboratory on human Chromosomes 21 and 22 also showed a positive correlation between early replication and gene expression, but the results could have been improperly skewed because of atypical features of the two small chromosomes. The positive correlation between early replication and gene expression in this study is likely to be generally true throughout the genome, because it was obtained with a distributed set of segments that together are representative of the entire genome.

The association of early replication with gene expression suggests that there are consistent differences in chromatin environment between the early- and late-replicating segments. Cytological studies have shown spatial differences in nuclear staining for both activating and silencing histone modification marks, and these spatial differences in histone modification are correlated with differences in replication time (Wu et al. 2005). ChIP for histone modification marks reported here strengthens the correlation at a finer resolution: early replication and gene expression correlate with euchromatin, and conversely, late replication correlates with low gene expression and heterochromatin. These results are confirmed in a wider study that correlates our replication time data with ChIP-on-chip data for histone modifications done by the ENCODE Consortium (The ENCODE Project Consortium 2007).

Interestingly, the finer resolution offered by genome-tiling microarrays identified chromosomal segments with acute transitions in replication timing. For one particular segment (ENm005) (Fig. 6), the replication time transition was confirmed by interphase FISH and appeared to correlate with transitions in both gene expression and chromatin modifications: a late-replicating island of 355 kb had repressive chromatin marks and low gene expression. The genes *OLIG1* and *OLIG2* in this island are known to be expressed during development of oligodendrocyte (OL) lineage (Jakovcevski and Zecevic 2005), and thus the island is expected to become early replicating in oligodendrocytes.

Identification of transition zones separating chromosomal domains is an interesting outcome from the replication profiles. Thirty-one genes of biomedical significance (including 10 oncogenes/tumor suppressor genes on 11q and 21q) reside in or near replication-timing transition regions (Watanabe et al. 2002). The mechanism by which a boundary is maintained between euchromatin and heterochromatin around these transition zones is not understood. At the major histocompatibility complex (MHC), loci replication timing switches precisely where there is a transition in the GC% content and is associated with nuclear scaffold attachment regions (Tenzen et al. 1997). A similar transition in GC content in the neurofibromatosis 1 (*NF1*) gene demarcates early replicating from late-replicating chromatin, and a stalled replication fork was observed in this transition region (Schmeg-

ner et al. 2005). The sites of replication time switch identified by our method will likely lead to the identification of more such transition zones, and we are interested in determining in the future whether such zones cause replication forks to slow down or stall, whether they contain nuclear scaffold attachment regions, and whether they act as boundary elements responsible for keeping adjoining chromatin domains separate from each other.

In humans, the R and G chromosomal bands have been linked to both gene density and AT/GC content. G bands are AT-rich, while the R bands are more GC-rich. GC-rich regions are not only enriched in genes but specifically in expressed genes (Saccone et al. 1993; Caron et al. 2001; Lander et al. 2001; Versteeg et al. 2003). The moderately positive correlation between AT content and TR50 (0.26 at 10-kb, 0.19 at 1-kb resolution) suggests a trend favoring an increase in AT content as we move from early- to late-replicating chromosomal segments. This observation is also in concordance with our previous study on Chromosomes 21 and 22 (Jeon et al. 2005). The correlation increases as the computation is done at larger scales, suggesting that the influence of AT content on TR50 occurs at scales greater than tens of kilobases. Consistent with this, replication-timing studies done at 1-Mb resolution show an even stronger positive correlation with AT content (Woodfine et al. 2004).

Finally, the smoothed TR50 profile suggests locations of origins of replication at local minima and positions of replication fork termination at local maxima. Replication speed can also be estimated based on the slope of the smoothed TR50 profile at a given locus. These possibilities will be explored in our future work.

Methods

Cell culture, synchronization, and FACS analysis

HeLa and MCF10A cells were cultured as per standard growth conditions. For synchronous progression through S phase, HeLa and MCF10A cell lines were arrested by thymidine-aphidicolin block and released as described earlier (Jeon et al. 2005). For nocodazole block, HeLa cells at 60% confluency were treated with 40 ng/mL nocodazole for 10 h. This was followed by selection of cells blocked in mitosis by mitotic shake-off. These cells were washed three times with PBS and released into fresh medium for 12 h to reach the 0-h point when they enter S phase. Cells harvested at indicated time points of S phase were either used for FISH or fixed in 70% ethanol and stained with propidium iodide (PI) for FACS by standard methods.

Newly replicated DNA (H/L DNA) purification

Synchronously released cells were labeled with 100 µM BrdU for the indicated time interval; 10 ~ 30 150-mm plates of cells were used to purify H/L DNA from each time point as described earlier (Jeon et al. 2005).

Microarray hybridization

To generate replication time profiles, ENCODE01-Forward (P/N 900543; Affymetrix) tiling arrays were used. These arrays are designed to study the pilot ENCODE regions of DNA, comprised of 30 Mb of DNA, or ~1% of the human genome. These pilot regions were selected by a committee of the National Human Genome Research Institute (NHGRI). Half of the content on the ENCODE01 Array was manually selected by the NHGRI committee, while the remaining 50% was randomly selected (The ENCODE Project Consortium 2004). A total of 14.82 Mb of sequence constituted the manually selected regions and included 14 targets ranging in size from 500 kb to 2 Mb. The randomly selected content includes 30 500-kb regions selected based on gene density and level of nonexonic conservation.

Nonrepetitive, 25-mer oligonucleotide probe pairs (Perfect Match, PM; Mis-Match control, MM) spaced at intervals of ~22 bp as measured from the central nucleotide were spotted on arrays. Heavy/light DNA from each time point was fragmented to 50–100 bp by DNase I digestion, end-labeled with biotin-ddATP using terminal transferase, and hybridized to the arrays as per the manufacturer's protocol (Affymetrix). Each microarray was scanned and analyzed for signal intensities using a GeneChIP Scanner 3000 and GeneChIP Operating Software (GCOS; Affymetrix). Two biological and one technical replicates were hybridized to ascertain the reproducibility of array hybridizations. The primary data in the form of .cel files can be accessed at http://www.cs.virginia.edu/~cmt5n/Rawtimepoints/. All the primary and processed data have been generated using the hg17 (NCBI Build 35, May 2004) build of the Human genome assembly.

The replication signal for each probe was calculated as PM − MM. If the difference was negative, then the signal was assigned a value of 0. For a given probe on the array, we have five replication signals, one from each time point. Each probe is classified to be replicating either in a temporally specific or nonspecific (asynchronous) manner as follows. Probes were temporally specific if the signal in any one time point was at least twice the signal of each of the other four time points. To accommodate the possibility that a temporally specific replication signal could span two adjacent time points, probes were also called specific if the sum of any two adjacent time points was at least three times the signal of each of the other three time points. Probes that do not satisfy either of the criteria above are designated as temporally nonspecific. Supplemental Table 1 gives some examples of the specificity classification. For the Affymetrix ENCODE array, we classified 26.115% of the probes as temporally nonspecific in their pattern of replication.

For studying gene expression, RNA was extracted from logarithmically growing HeLa cells by using an RNeasy Kit (QIAGEN) and hybridized to the Human HG-U133 Plus 2.0 array (containing ~38,500 genes) as described in the Affymetrix GeneChIP protocol (Affymetrix). Each microarray was scanned, visualized, and analyzed for the level of each individual transcript using a GeneChIP Scanner 3000 and GeneChIP Operating Software (GCOS; Affymetrix). Transcript signal was mapped against the chromosome coordinates (probe-by-probe basis) using the HG-U133A_2 Annotations, CSV provided by the manufacturer (Affymetrix).

Segregation of temporally specific and pan-S replicating segments

To segregate broad regions of replication, a sliding window of 10 kb was passed along each chromosomal segment, calculating the percentage of temporally nonspecific probes within the window. A pan-S region is begun when the percentage exceeds 60% and continues until it drops below the 60% threshold minus a given tolerance (e.g., 10% for our analysis). The tolerance is introduced in order to avoid thrashing between nonspecific and specific regions. Once the percentage drops below "threshold tolerance" (e.g., 50% for our settings of threshold and tolerance), the current pan-S region ends and a temporally specific region is started. The temporally specific region is continued until the percentage again rises above the threshold. In this manner, moving along the chromosome, broad regions of replication are segregated into temporally specific or pan-S classes.

The tolerance parameter, which helps us avoid thrashing between the two classes, introduces a directional bias into the segregation algorithm. As we move from lower chromosomal positions to higher chromosomal positions, the percentage must exceed 60% in order to begin a pan-S region. But the pan-S region does not end until the percentage drops below 50%. In order to correct for this directional bias, we perform two passes of the algorithm. One pass moves the window toward higher chromosomal positions, while the other pass moves the window toward lower chromosomal positions. Then we merge the two passes into a single segregation, which no longer suffers from a directional bias.

Interphase fluorescence in situ hybridization

Cells in S phase were harvested at indicated time points and incubated in pre-warmed 75 mM KCl solution for 15 min at 37°C to prepare nuclei. These nuclei were fixed in 3:1 (v/v) methanol/glacial acetic acid and mounted on a slide. A nick translation kit and SpectrumGreen dUTP/Spectrum Red dUTP (Vysis Inc.) were used for labeling the probe. Hybridization was carried out in a humidified chamber for 16 h at 37°C as described in the Vysis FISH protocol (Vysis Inc.). Slides were washed with 0.4× SSC/0.3% NP-40 for 2 min at 73°C and again with 2× SSC/0.1% NP-40 solution for 1 min at room temperature. Chromosomal DNA was counterstained with DAPI (VECTASHIELD Mounting Medium; Vector Labratories) and visualized with a Nikon Microphot.SA fluorescent microscope equipped with a DAPI filter, FITC, and a TRITC cube set (for Spectrum Green and Spectrum red fluors, respectively). Images were digitally obtained with a Nikon UFX-DX camera and Spot version 3.5.4 software. All the BAC clones were purchased from Children's Hospital Oakland Research Institute.

The number of dots was visually counted in ~100 cells at each time interval, and the number of dots/cell was calculated; 100% replication (in G_2 cells) was when the increase in the number of dots/cell equaled the number of dots/cell observed in G_1. After determining the dots/cell at 0, 2, 4, 6, 8, and 10 h of S phase, for each interval (e.g., 0–2 h, 2–4 h, etc.), we calculated the increase in dots/cell during that interval and converted it to the percent of replication.

Correlation of TR50 with genome features

CpG island annotations were obtained from the UCSC Genome Browser Web site (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=64765488&c=chr16&g=cpgIsland). The density of CpG islands was calculated for all the chromosomal regions in each of the replication segments, that is, early-, mid-, late-, and pan-S. For the CpG islands that overlapped two temporal segments, the number of bases in the CpG island were counted and a 60% cutoff was used to assign it a specific temporal classification.

For determining gene density, we used the annotated genes under the Refseq database from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=64796361&c=chr16&g=refGene).

Enrichment in each of the replication segments (early-, mid-, late-, and pan-S) within a given data set (CpG islands, gene density, and transcripts) was calculated as follows. The number of elements from the data set whose majority base count fell into early segments was calculated. This was divided by the total number of elements in the data set to get a ratio of early-replicating elements. This ratio was divided by the ratio of early segments to all segments to give the enrichment ratio. Hence, a value of 1.0 would indicate that the data set was distributed in early segments as was expected by chance, while a value of 2.0 would indicate twice as many as expected by chance. Enrichment of the mid-, late-, and pan-S replicating regions was calculated similarly.

Chromatin immunoprecipitation

A chromatin immunoprecipitation assay was performed as per the protocol described at http://www.upstate.com with a variation in the sonication step. Samples were sonicated using a Branson microtip (3.2 mm) and Fisher Model 500 Sonic Dismembrator. Eight cycles of 15-sec pulse at 50% amplitude and 45 sec of cooling on ice were done to disrupt the cells. The antibodies used for ChIP were for identifying sites of histone 3 Lys 4 mono-, di-, and trimethylation (H3K4 Me), histone 3 Lys9 dimethylation (H3K9 di-Me), and HP1α. These antibodies were purchased from Upstate (Anti-H3K4 Me; 05-791 and H3K9 Anti-Me; 07-441) and Abcam (Anti- HP1α; ab9057). To determine the ChIP signal for H3K9 di-Me, 4 µL of ChIP DNA were first amplified in a linear range (14 cycles) using the WGA2 kit from Sigma-Aldrich and cleaned up by the QIAGEN PCR clean-up kit. Two microliters of this purified DNA was used as template for ChIP assay with primers. To rule out any amplification bias, three independent amplifications were performed and PCR with primers repeated with each of these template preparations. As a negative control, ChIP DNA from an IgG sample was amplified in a similar way. The details on primers used for ENm005 and ENr132 regions are provided in Supplemental Table 3.

Acknowledgments

This work was supported by National Institutes of Health Grant HG003157 (to A.D.). We thank members of the Dutta laboratory for helpful discussions.

References

Aladjem, M.I., Rodewald, L.W., Kolman, J.L., and Wahl, G.M. 1998. Genetic dissection of a mammalian replicator in the human beta-globin locus. *Science* **281:** 1005–1009.

Avner, P. and Heard, E. 2001. X-Chromosome inactivation: Counting, choice and initiation. *Nat. Rev. Genet.* **2:** 59–67.

Bell, S.P. and Dutta, A. 2002. DNA replication in eukaryotic cells. *Annu. Rev. Biochem.* **71:** 333–374.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas III, E.J., Gingeras, T.R., et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120:** 169–181.

Bickmore, W.A. and Carothers, A.D. 1995. Factors affecting the timing and imprinting of replication on a mammalian chromosome. *J. Cell Sci.* **108:** 2801–2809.

Boumil, R.M. and Lee, J.T. 2001. Forty years of decoding the silence in X-chromosome inactivation. *Hum. Mol. Genet.* **10:** 2225–2232.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., et al. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289–1292.

Carrel, L., Clemson, C.M., Dunn, J.M., Miller, A.P., Hunt, P.A., Lawrence, J.B., and Willard, H.F. 1996. X inactivation analysis and DNA methylation studies of the ubiquitin activating enzyme E1 and PCTAIRE-1 genes in human and mouse. *Hum. Mol. Genet.* **5:** 391–401.

Chang, P.L., Mueller, O.T., Lafrenie, R.M., Varey, P.A., Rosa, N.E., Davidson, R.G., Henry, W.M., and Shows, T.B. 1990. The human arylsulfatase-C isoenzymes: Two distinct genes that escape from X inactivation. *Am. J. Hum. Genet.* **46:** 729–737.

Chess, A., Simon, I., Cedar, H., and Axel, R. 1994. Allelic inactivation regulates olfactory receptor gene expression. *Cell* **78:** 823–834.

Demeret, C., Vassetzky, Y., and Mechali, M. 2001. Chromatin remodelling and DNA replication: From nucleosomes to loop domains. *Oncogene* **20:** 3086–3093.

Dhar, V., Skoultchi, A.I., and Schildkraut, C.L. 1989. Activation and repression of a beta-globin gene in cell hybrids is accompanied by a

shift in its temporal replication. *Mol. Cell. Biol.* **9:** 3524–3532.
Disteche, C.M. 1995. Escape from X inactivation in human and mouse. *Trends Genet.* **11:** 17–22.
The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.
The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* (in press).
Epner, E., Forrester, W.C., and Groudine, M. 1988. Asynchronous DNA replication within the human beta-globin gene locus. *Proc. Natl. Acad. Sci.* **85:** 8081–8085.
Fournier, C., Goto, Y., Ballestar, E., Delaval, K., Hever, A.M., Esteller, M., and Feil, R. 2002. Allele-specific histone lysine methylation marks regulatory regions at imprinted mouse genes. *EMBO J.* **21:** 6560–6570.
Fuks, F. 2005. DNA methylation and histone modifications: Teaming up to silence genes. *Curr. Opin. Genet. Dev.* **15:** 490–495.
Goldman, M.A. 1988. The chromatin domain as a unit of gene regulation. *Bioessays* **9:** 50–55.
Hansen, R.S., Canfield, T.K., Fjeld, A.D., and Gartler, S.M. 1996. Role of late replication timing in the silencing of X-linked genes. *Hum. Mol. Genet.* **5:** 1345–1353.
Holmquist, G., Gray, M., Porter, T., and Jordan, J. 1982. Characterization of Giemsa dark- and light-band DNA. *Cell* **31:** 121–129.
Iqbal, M.A., Plumb, M., Stein, J., Stein, G., and Schildkraut, C.L. 1984. Coordinate replication of members of the multigene family of core and H1 human histone genes. *Proc. Natl. Acad. Sci.* **81:** 7723–7727.
Jakovcevski, I. and Zecevic, N. 2005. Olig transcription factors are expressed in oligodendrocyte and neuronal cells in human fetal CNS. *J. Neurosci.* **25:** 10064–10073.
Jeon, Y., Bekiranov, S., Karnani, N., Kapranov, P., Ghosh, S., MacAlpine, D., Lee, C., Hwang, D.S., Gingeras, T.R., and Dutta, A. 2005. Temporal profile of replication of human chromosomes. *Proc. Natl. Acad. Sci.* **102:** 6419–6424.
Jiang, Y.H., Bressler, J., and Beaudet, A.L. 2004. Epigenetics and human disease. *Annu. Rev. Genomics Hum. Genet.* **5:** 479–510.
Jones, P.A. and Takai, D. 2001. The role of DNA methylation in mammalian epigenetics. *Science* **293:** 1068–1070.
Kitsberg, D., Selig, S., Keshet, I., and Cedar, H. 1993. Replication structure of the human beta-globin gene domain. *Nature* **366:** 588–590.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
MacAlpine, D.M. and Bell, S.P. 2005. A genomic view of eukaryotic DNA replication. *Chromosome Res.* **13:** 309–326.
MacAlpine, D.M., Rodriguez, H.K., and Bell, S.P. 2004. Coordination of replication and transcription along a *Drosophila* chromosome. *Genes & Dev.* **18:** 3094–3105.
Miller, A.P., Gustashaw, K., Wolff, D.J., Rider, S.H., Monaco, A.P., Eble, B., Schlessinger, D., Gorski, J.L., van Ommen, G.J., Weissenbach, J., et al. 1995. Three genes that escape X chromosome inactivation are clustered within a 6 Mb YAC contig and STS map in Xp11.21–p11.22. *Hum. Mol. Genet.* **4:** 731–739.
Mostoslavsky, R., Singh, N., Tenzen, T., Goldmit, M., Gabay, C., Elizur, S., Qi, P., Reubinoff, B.E., Chess, A., Cedar, H., et al. 2001. Asynchronous replication and allelic exclusion in the immune system. *Nature* **414:** 221–225.
Perry, P., Sauer, S., Billon, N., Richardson, W.D., Spivakov, M., Warnes, G., Livesey, F.J., Merkenschlager, M., Fisher, A.G., and Azuara, V. 2004. A dynamic switch in the replication timing of key regulator genes in embryonic stem cells upon neural induction. *Cell Cycle* **3:** 1645–1650.
Raghuraman, M.K., Winzeler, E.A., Collingwood, D., Hunt, S., Wodicka, L., Conway, A., Lockhart, D.J., Davis, R.W., Brewer, B.J., and Fangman, W.L. 2001. Replication dynamics of the yeast genome. *Science* **294:** 115–121.
Rountree, M.R., Bachman, K.E., and Baylin, S.B. 2000. DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. *Nat. Genet.* **25:** 269–277.
Rountree, M.R., Bachman, K.E., Herman, J.G., and Baylin, S.B. 2001. DNA methylation, chromatin inheritance, and cancer. *Oncogene* **20:** 3156–3165.
Saccone, S., De Sario, A., Wiegant, J., Raap, A.K., Della Valle, G., and Bernardi, G. 1993. Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci.* **90:** 11929–11933.
Schmegner, C., Berger, A., Vogel, W., Hameister, H., and Assum, G. 2005. An isochore transition zone in the NF1 gene region is a conserved landmark of chromosome structure and function. *Genomics* **86:** 439–445.
Schubeler, D., Scalzo, D., Kooperberg, C., van Steensel, B., Delrow, J., and Groudine, M. 2002. Genome-wide DNA replication profile for *Drosophila melanogaster*: A link between transcription and replication timing. *Nat. Genet.* **32:** 438–442.
Simon, I., Tenzen, T., Reubinoff, B.E., Hillman, D., McCarrey, J.R., and Cedar, H. 1999. Asynchronous replication of imprinted genes is established in the gametes and maintained during development. *Nature* **401:** 929–932.
Simon, I., Tenzen, T., Mostoslavsky, R., Fibach, E., Lande, L., Milot, E., Gribnau, J., Grosveld, F., Fraser, P., and Cedar, H. 2001. Developmental regulation of DNA replication timing at the human beta globin locus. *EMBO J.* **20:** 6150–6157.
Tenzen, T., Yamagata, T., Fukagawa, T., Sugaya, K., Ando, A., Inoko, H., Gojobori, T., Fujiyama, A., Okumura, K., and Ikemura, T. 1997. Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex. *Mol. Cell. Biol.* **17:** 4043–4050.
Vermeesch, J.R., Petit, P., Kermouni, A., Renauld, J.C., Van Den Berghe, H., and Marynen, P. 1997. The IL-9 receptor gene, located in the Xq/Yq pseudoautosomal region, has an autosomal origin, escapes X inactivation and is expressed from the Y. *Hum. Mol. Genet.* **6:** 1–8.
Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J., and van Kampen, A.H. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13:** 1998–2004.
Wang, L., Lin, C.M., Brooks, S., Cimbora, D., Groudine, M., and Aladjem, M.I. 2004. The human beta-globin replication initiation region consists of two modular independent replicators. *Mol. Cell. Biol.* **24:** 3373–3386.
Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y., and Ikemura, T. 2002. Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: Disease-related genes in timing-switch regions. *Hum. Mol. Genet.* **11:** 13–21.
Woodfine, K., Fiegler, H., Beare, D.M., Collins, J.E., McCann, O.T., Young, B.D., Debernardi, S., Mott, R., Dunham, I., and Carter, N.P. 2004. Replication timing of the human genome. *Hum. Mol. Genet.* **13:** 191–202.
Wu, R., Terry, A.V., Singh, P.B., and Gilbert, D.M. 2005. Differential subnuclear localization and replication timing of histone H3 lysine 9 methylation states. *Mol. Biol. Cell* **16:** 2872–2881.
Zhou, J., Ermakova, O.V., Riblet, R., Birshtein, B.K., and Schildkraut, C.L. 2002. Replication and subnuclear location dynamics of the immunoglobulin heavy-chain locus in B-lineage cells. *Mol. Cell. Biol.* **22:** 4876–4889.

Received April 21, 2006; accepted in revised form October 30, 2006.

To appear in the 3rd International Conference on Bioinformatics and Biomedical Engineering (ICBBE 2009).

An Information Theoretic Viewpoint on Haplotype Reconstruction from SNP Fragments

Huimin Chen
Department of Electrical Engineering
University of New Orleans
New Orleans, LA 70148, USA
Email: hchen2@uno.edu

Zhiyu Zhao
Department of Computer Science
University of New Orleans
New Orleans, LA 70148, USA
Email: sylvia@cs.uno.edu

***Abstract*—The problem of haplotype reconstruction based on aligned single nucleotide polymorphism (SNP) fragments is formulated as decoding over a discrete memoryless channel. An information theoretic view point is used to illustrate the deficiency of parsimonious models such as minimum error correction model. A new computational model with genotype information is proposed and a low complexity reconstruction algorithm for this model is shown to guarantee the desired reconstruction rate by increasing the number of SNP fragments sequentially. The advantage of using genotype information is quantified by exploiting a simplified statistical model for haplotype sequences.**

Index Terms: Haplotype inference, error correction, sequential test, genotype.

I. Introduction

Single nucleotide polymorphism (SNP) is a locus in the DNA sequence where an alternation of the nucleotide from other members of the same species occurs at a considerable frequency. Alleles of a set of linked genetic markers located on a single DNA sequence is called a haplotype. For diploid organisms, such as human, haplotypes come in pairs, where the two haplotypes in the pair are not necessarily identical. Each pair of haplotypes can also be combined to form a genotype. When a pair of alleles at an SNP site is made of identical type (either being both wild, denoted by 0, or both mutant, denoted by 1), the site is called homozygous site. Otherwise, it is called heterozygous site. It is generally believed that while haplotypes contain more crucial information than the individual SNPs in disease association studies, it is substantially more difficult to determine haplotypes than to determine genotypes or individual SNPs through experiments. Owing to its potential in genomic study, haplotype inference has drawn significant attention from both statistical and computational societies.

Computational methods for haplotype inference can be largely treated in two different categories. One concerns with obtaining compatible haplotypes from the genotype samples in a population [7]. The pioneering work by Clark [2] demonstrated effective haplotype inference from genotypes by a parsimony based method. Later, various statistical and Bayesian methods have been proposed [9], [1], [13], [5] for large genotype data sets as well as large number of subjects under different assumptions on the underlying biology systems. The other uses short gnome fragments with SNPs coming from DNA shotgun sequencing or some other resequencing procedure to reconstruct the haplotypes directly. A particular problem called single individual haplotyping is to reconstruct a pair of long haplotype sequences from short fragments of SNPs. If SNP fragments are well aligned, then the problem becomes how to partition these SNP fragments into two sets and use each set to determine a haplotype sequence. Suppose that there are m SNP fragments from a pair of chromosomes and the length of the corresponding haplotypes is n. Define an $m \times n$ SNP matrix M whose entry m_{ij} has value 0, 1 or –. The symbol – means a missing or skipped base, which is called a gap. Let $\mathbf{h}$ be a haplotype sequence whose entry can be either 0 or 1. Let $\Theta = (M_1, M_2)$ be a partition of of M that divides the rows of M into two disjoint sets. The haplotype reconstruction problem can be written in terms of maximum likelihood estimation

$$(\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \hat{\Theta}) = \arg \max_{\mathbf{h}_1, \mathbf{h}_2, \Theta} \Lambda(M_1|\mathbf{h}_1)\Lambda(M_2|\mathbf{h}_2).$$

Alternatively, one can entail minimum error correction [11], minimal entropy [4], minimum fragment removal, minimum snip removal or longest haplotype reconstruction [12] as the optimality criterion in reconstructing the pair of haplotype sequences without the need of likelihood function model $\Lambda(\cdot)$. Most of these criteria are inherently based on parsimony and combinatorial in nature — finding exact optimal solution is often an NP hard problem [6]. This paper provides a different viewpoint on the haplotype reconstruction from M. The problem is treated as decoding with noisy observations over a discrete memoryless channel. It is shown that parsimony based criterion, albeit computationally demanding, does not necessarily yield the best reconstruction rate. There exists a significant gap between the error correction capability with SNP fragments and the theoretical limit given by the channel capacity. To improve the error correction rate, we propose a new computational model with genotype information available for haplotype reconstruction. The resulting algorithm has low complexity and is shown to guarantee the desired error rate by increasing the number of SNP fragments sequentially. The advantage of using genotype information is quantified by exploiting a simplified statistical model for haplotype sequences.

The rest of the paper is organized as follows. Section II formulates the haplotype reconstruction problem from information theoretic perspective. Section III introduces a low complexity sequential algorithm using genotype information. Section IV compares the reconstruction rate between the cases without and with genotype information and quantifies performance gain. Concluding remarks are in Section V.

II. Haplotype Reconstruction As A Decoding Problem

For $x, y \in \{0, 1, -\}$, we define the distance

$$d(x,y) = \begin{cases} 1 & \text{if } x \neq y, x \neq -, y \neq - \\ 0 & \text{otherwise} \end{cases}$$

which is an extension of Hamming distance defined for $x, y \in \{0, 1\}$. Similarly, we can define

$$d(\mathbf{h}_1, \mathbf{h}_2) = \sum_{i=1}^{n} d(h_{1i}, h_{2i}).$$

The dissimilarity between $\mathbf{h}_1$ and $\mathbf{h}_2$ is measured by

$$\beta = d(\mathbf{h}_1, \mathbf{h}_2)/n.$$

We assume that for any given bit of a haplotype $\mathbf{h}$, each symbol of the corresponding column of the SNP matrix is generated according to a probability distribution $P(y|x)$. Specifically, we can write $P(y|x)$ as a transition probability matrix given by

$P(y\|x)$	$x=0$	$x=1$
$y=0$	$1-\alpha_{01}-\alpha_{02}$	α_{10}
$y=1$	α_{01}	$1-\alpha_{10}-\alpha_{12}$
$y=-$	α_{02}	α_{12}

If we further assume that $\alpha_{01} = \alpha_{10} = e_1$ and $\alpha_{02} = \alpha_{12} = e_2$, then the SNP fragments can be viewed as a haplotype sequence being transmitted repetitively over a binary symmetric erasure channel [8]. Clearly, e_1 is the probability of bit flip and e_2 is the probability of bit erasure.

Claim 1: Assume that the SNP matrix M is obtained by transmitting m_1 times on each bit of $\mathbf{h}_1$ and m_2 times on each bit of $\mathbf{h}_2$ through the binary symmetric erasure channel ($m_1 + m_2 = m$). If $\beta > 4e_1$, then $P(\hat{\Theta} = \Theta) \to 1$ as $n \to \infty$.

Proof Sketch: Denote by $\mathbf{y}_i$ the i-th row of M_1 generated from $\mathbf{h}_1$ and $\mathbf{z}_j$ the j-th row of M_2 generated from $\mathbf{h}_2$. Asymptotically, we have

$$d(\mathbf{y}_i, \mathbf{h}_1)/n \sim \mathcal{N}(e_1, e_1(1-e_1)/n),$$
$$d(\mathbf{z}_j, \mathbf{h}_2)/n \sim \mathcal{N}(e_1, e_1(1-e_1)/n)$$

for $i = 1, ..., m_1$, $j = 1, ..., m_2$. Thus

$$\frac{d(\mathbf{y}_i, \mathbf{y}_j)}{n} \leq \frac{d(\mathbf{y}_i, \mathbf{h}_1)}{n} + \frac{d(\mathbf{y}_j, \mathbf{h}_1)}{n} \sim \mathcal{N}\left(2e_1, \frac{2e_1(1-e_1)}{n}\right).$$

On the other hand, we have

$$\begin{aligned} \frac{d(\mathbf{y}_i, \mathbf{z}_j)}{n} &\geq \frac{d(\mathbf{h}_1, \mathbf{h}_2)}{n} - \left[\frac{d(\mathbf{y}_i, \mathbf{h}_1)}{n} + \frac{d(\mathbf{z}_j, \mathbf{h}_2)}{n}\right] \\ &\sim \mathcal{N}\left(\beta - 2e_1, \frac{2e_1(1-e_1)}{n}\right). \end{aligned}$$

As $n \to \infty$, it is clear that

$$P\left(\frac{d(\mathbf{y}_i, \mathbf{y}_j)}{n} \leq 2e_1, \frac{d(\mathbf{y}_i, \mathbf{z}_k)}{n} \geq 2e_1\right) \to 1.$$

Note that $\forall i \neq j$ and $\forall k$, the condition for equality to hold has zero probability measure.

Claim 2: If $e_1 > 0$, then for any finite m_1, the maximum likelihood estimate of the haplotype sequence yields $P(\hat{\mathbf{h}}_1 \neq \mathbf{h}_1) > 0$ and $P(\hat{\mathbf{h}}_1 \neq \mathbf{h}_1) \to 1$ as $n \to \infty$.

Proof Sketch: Without loss of generality, we assume that the haplotype sequence is independent and identically distributed so that we can focus on the decoding of each bit of the haplotype sequence based on the corresponding column of M_1. Assume that k_1 0s and k_2 1s are observed with $m_1 - k_1 - k_2$ erasures in a particular column of M_1. The log-posterior ratio is

$$\log\left(\frac{P(x=0|k_1,k_2)}{P(x=1|k_1,k_2)}\right) = \log\left(\frac{P(x=0)}{P(x=1)}\right) + (k_1 - k_2)\log\left(\frac{1-e_1-e_2}{e_1}\right).$$

Assume equal prior probability and $1 - e_2 > 2e_1$, then the decision rule becomes declaring $x = 0$ when $k_1 - k_2 > 0$, i.e., $k_1 > k_2$ and $x = 1$ when $k_1 < k_2$. The probability of decision error is given by

$$P(\hat{x} \neq x) = \sum_{k_1 < k_2} \frac{m_1!}{k_1!k_2!(m_1-k_1-k_2)!}(q)^{k_1} e_1^{k_2} e_2^{m_1-k_1-k_2} > 0$$

where $q = 1 - e_1 - e_2$. Thus the probability of correctly decoding an n-bit haplotype is

$$P(\hat{\mathbf{h}}_1 = \mathbf{h}_1) = (1 - P(\hat{x} \neq x))^n \to 0$$

as $n \to \infty$.

Claim 3: If $C = \log_2 3 - H(e_1, e_2, 1-e_1-e_2) > 0$ and $m_1 > n/C$, then for large enough n, there exists a procedure to generate the SNP matrix such that $P(\hat{\mathbf{h}}_1 \neq \mathbf{h}_1)$ can be made arbitrarily small.

Proof Sketch: If we assume that a haplotype is an independent identically distributed stochastic sequence with each bit having a distribution $P(x)$, then for large n, there are about $2^{nH(x)}$ typical sequences. For each row of M, there are about $2^{nH(y)}$ typical sequences. Since there are only $2^{nH(x,y)}$ jointly typical sequences, we can see that not all pairs of typical haplotype and typical fragment sequence are jointly typical. For the discrete memoryless channel $P(y|x)$, the probability that any randomly chosen pair is jointly typical is about 2^{-nC} where the channel capacity is [3]

$$C = I(x;y) = \log_2 3 - H(e_1, e_2, 1-e_1-e_2).$$

Hence for a fixed fragment sequence, we can consider about 2^{nC} such pairs before we are likely to come across a jointly typical pair. This suggests that there are about 2^{nC} distinguishable haplotypes we can handle with M_1. From Shannon's channel coding theorem [3], if the communication rate

$n/m_1 < C$, then there exists a code book that by examining the joint typicality, the decoding error can be made arbitrarily small.

In summary, finding the correct partition Θ has overwhelmingly large probability asymptotically if the two haplotype sequences are dissimilar enough. However, since the standard SNP matrix M is constructed by repetition code which has zero rate asymptotically [8], there exists a significant gap between the information content conveyed by M and the actual information needed to make the decoding error probability arbitrarily small.

III. Efficient Haplotype Reconstruction with Genotype Information

Consider a long haplotype sequence of length n being measured by L aligned SNP fragments each of which has length $s = n/L$. For each SNP block with s columns, we assume that the genotype information is available. For a genotype $\mathbf{g} = [g_1\, g_2 \ldots g_s]$ without reading error, when the i-th SNP site is wild type homozygous, $g_i = 0$; when it is mutant type homozygous, $g_i = 1$; when it is heterozygous, $g_i = 2$. Clearly, a pair of identical haplotypes can not yield a genotype sequence with any site being 2 unless it is a reading error. If we screen each SNP block and observe 2 on certain sites, the information will be very useful to infer Θ. Thus we can view the genotype information as some form of the parity check [8].

With the help of genotype information, we can design an efficient haplotype reconstruction algorithm that controls the decoding error of each site by sequentially taking an additional SNP block. The algorithm runs in two stages.

Given initial SNP matrix $M = [B_1\, B_2 \ldots B_s]$ with known bit flip probability e_1, erasure probability e_2, and the corresponding genotype sequence $\{\mathbf{g}_1, ..., \mathbf{g}_s\}$.

Output a pair of haplotypes $\mathbf{h}_1$ and $\mathbf{h}_2$ with error probability for each site below a desired level e.

Algorithm

- Partition: Start with an arbitrary initial partition $\Theta = (M_1, M_2)$. Identify all sites where $g_i = 2$ and set $h^*_{1i} = 0$, $h^*_{2i} = 1$ if there are more 0s in the i-th column of M_1 than 1s in the i-th column of M_2. Otherwise, set $h^*_{1i} = 1$, $h^*_{2i} = 0$.
  For the remaining sites, set $h^*_{1j} = h^*_{2j} = g_j$. Use majority vote by the corresponding column vector in M_1 to decide h^*_{1j} and the corresponding column vector in M_2 to decide h^*_{2j} when g_j is a gap.
  Cluster M into two groups with initial centers given by $\mathbf{h}^*_1$ and $\mathbf{h}^*_2$ and generalized Hamming distance defined over each pair of row vectors in M. Standard k-means algorithm [8] will converge in less than 10 iterations in practice.
- Decoding: For each column l in B_i ($i = 1, ..., s$), do the following.
  If $g_{il} = 0$, then count the number of 0s k_1 and the number of 1s k_2 in the whole column l. Declare $h_{il1} = h_{il2} = 0$ if $k_1 - k_2 > c_0$; declare $h_{il1} = h_{il2} = 1$ if $k_2 - k_1 > c_0$; and request one more piece of SNP block appending to B_i if $|k_1 - k_2| \le c_0$.
  If $g_{il} = 1$, then count the number of 0s k_2 and the number of 1s k_1 in the whole column l. Declare $h_{il1} = h_{il2} = 1$ if $k_1 - k_2 > c_0$; declare $h_{il1} = h_{il2} = 0$ if $k_2 - k_1 > c_0$; and request one more piece of SNP block appending to B_i if $|k_1 - k_2| \le c_0$.
  If $g_{il} = 2$, then count the number of 0s k_{11} and the number of 1s k_{12} in the column l belonging to group 1 and count the number of 0s k_{21} and the number of 1s k_{22} in the column l belonging to group 2. Declare $h_{il1} = 0$, $h_{il2} = 1$ if $k_{11} - k_{12} > c_1$ or $k_{22} - k_{21} > c_1$; declare $h_{il1} = 1$, $h_{il2} = 0$ if $k_{12} - k_{11} > c_1$ or $k_{22} - k_{21} > c_1$; otherwise, request one more piece of SNP block appending to B_i.
  If g_{il} is a gap, then count the number of 0s k_{11} and the number of 1s k_{12} in the column l belonging to group 1 and count the number of 0s k_{21} and the number of 1s k_{22} in the column l belonging to group 2. Declare $h_{il1} = 0$ if $k_{11} - k_{12} > c_1$; declare $h_{il2} = 1$ if $k_{22} - k_{21} > c_1$; declare $h_{il1} = 1$ if $k_{12} - k_{11} > c_1$; $h_{il2} = 0$ if $k_{21} - k_{22} > c_1$; otherwise, request one more piece of SNP block appending to B_i.

Note that the above algorithm is applicable to the case without genotype information by setting g_{il} to be a gap $\forall l, i$.

Claim 4: Denote by p the probability that the genotype reading of an SNP site is correct. If $\beta > 4e_1$ or $\beta > 2(1-p)$ and

$$c_0 = \left\lceil \frac{\log\left(\frac{e(1-p)}{(1-e)p}\right)}{\log\left(\frac{(1-e_1-e_2)}{e_1}\right)} \right\rceil, c_1 = \left\lceil \frac{\log\left(\frac{e}{1-e}\right)}{\log\left(\frac{(1-e_1-e_2)}{e_1}\right)} \right\rceil,$$

then $P(\hat{h}_{jl} \neq h_{jl}) < e$, for $j = 1, 2$ and $l = 1, ..., n$, as $n \to \infty$.

Proof Sketch: Similar to the distance argument in the proof of Claim 1, we can show that $P(\hat{\Theta} = \Theta) \to 1$ as $n \to \infty$. Assuming conditional independence between M and $\{\mathbf{g}_1, ..., \mathbf{g}_s\}$, the decoding rule for each haplotype sequence is sequential probability ratio test (SPRT) with the upper and lower limit given by Wald's fundamental approximation [10]. Since the test statistic $(k_1 - k_2)$ is discrete, we have to increase the threshold to the nearest integer which in principle reduces the decoding error, i.e., $P(\hat{h}_{ilj} \neq h_{ilj}) < e$ with strict inequality $\forall i, l, j$. However, the expected number of samples to reach a decision will not be minimized with the exact error constraint e as the original SPRT, which requires randomized decision rule switching between the threshold c and $c - 1$ [10].

IV. Experiment on Simulation Data

We consider one chromosome data set used in [11] for haplotype reconstruction by minimum error correction (MEC) criterion. It contains a pair of haplotypes of length 95 after removing 8 missing sites. For the SNP matrix with fixed sample size $m = 40$, we generate random fragments with gap rate of fragments being 0.75. Among non-gap elements in the SNP matrix, the error rate is 0.3. The reconstruction rate given by

$$1 - \frac{\min\{r_{11} + r_{22}, r_{12} + r_{21}\}}{2n}$$

where $r_{ij} = d(\mathbf{h}_i, \hat{\mathbf{h}}_j)$, $i = 1, 2$ and $j = 1, 2$, is used to evaluate algorithm performance. The expected number of errors needs to be corrected is 285 while the MEC method using branch and bound algorithm in [11] only needs to correct 215 errors to make the reconstructed haplotypes compatible with the SNP matrix. The reported reconstruction rate is 0.705

[11] while the expected decoding error rate is 0.188. Thus the MEC method does not provide the best reconstruction rate. To apply the proposed sequential haplotype reconstruction algorithm, we set $L = 5$ thus $s = 19$. The initial sample size is 10 and we set $c = 2$. The expected number of SNP fragments used for reconstruction is 32.7 in 10 Monte Carlo runs. Ideally, the Wald's SPRT only needs 30 samples on average to reach a decision, which is smaller than the hypothesis test with fixed sample size $m = 40$. On the other hand, the reconstruction rate is 0.806, which is significantly higher than that by MEC method. It is also close to the expected decoding error rate even in this non-asymptotic regime.

Next, we consider that a genotype is associated with the SNP fragments with each SNP site having gap probability 0.25. For a non-gap genotype reading, the error probability is 0.1 ($p = 0.9$). We want to have $e < 0.01$ and end up with $c = 2$ as in the case without genotype information. In 10 Monte Carlo runs, the reconstruction rate becomes 0.998, which is higher than the desired rate. This is due to the conservative design of the SPRT procedure where the actual bit error rate is lower than the desired level. If one seeks minimum number of error corrections to make the reconstructed haplotype to be compatible with the modified genotype, then the reconstruction rate reduces to 0.924. Clearly, genotype information helps improve the reconstruction rate. However, correcting minimum number of errors does not lead to the optimal decoding due to the fact that typical errors will be unlikely minimal ones when n and e_1 are large enough [8]. Similar observations have also been confirmed using other parsimony based criteria. We conclude that parsimony based methods, being attractive mainly due to their general applicability without knowing the error rate of the underlying channel model, can be quite suboptimal compared with the best achievable reconstruction rate with the knowledge on the statistical model of the SNP fragments.

To quantify the improvement by using genotype information, we consider an idealized scenario where $p \to 1$ and every SNP site in the genotype is heterozygous. In this case, one can not infer any site of each haplotype purely from the genotype.

Claim 5: Assume that the decoding error rate using the proposed sequential algorithm is e for large n without genotype information. Then with genotype information under the idealized scenario, the decoding error rate is $O(e^2)$.

Proof Sketch: We assume that under both cases the inference on Θ is perfect. Without loss of generality, consider decoding the j-th site of the haplotypes where $h_{1j} = 0$ and $h_{2j} = 1$. Let k_{11} be the number of 0s of the j-th column in M_1 and k_{12} the number of 1s of the j-th column in M_1. Similarly, let k_{21} and k_{22} be the corresponding number of 0s and 1s in M_2, respectively. Without haplotype information, an error will occur either when $k_{11} - k_{12} < -c$ or when $k_{22} - k_{21} < -c$. Both events have been designed to have probability smaller than e. With haplotype information, an error will occur when both $k_{11} - k_{12} < -c$ and $k_{22} - k_{21} \leq c$ are true or both $k_{11} - k_{12} \leq c$ and $k_{22} - k_{21} < -c$ are true. Since under the same sample size, $k_{22} - k_{21} > c$ implies the correct decoding probability being at least $1 - e$, the event that $k_{22} - k_{21} \leq c$ is true prior to the stop time when $k_{11} - k_{12} < -c$ occurs has probability at most $O(e)$. Thus the overall decoding error rate is at most $O(e^2)$.

V. Conclusions

We have shown that haplotype reconstruction based on aligned SNP fragments can be treated as decoding over a discrete memoryless channel. There exists a nontrivial gap between the error correction capability by parsimony based methods and that given by the channel capacity. The parsimony based methods may not achieve the best error correction rate under such as channel model. In addition, we have shown how the genotype information can be useful to improve the haplotype reconstruction rate. A new sequential haplotype reconstruction algorithm with genotype information was proposed that guarantees the desired reconstruction rate with smaller expected number of SNP fragments than what is needed using fixed sample size. The advantage of using genotype information is quantified by exploiting a simplified statistical model with nearly perfect genotype information.

Acknowledgment

Stimulating discussions with Dr. Bin Fu of University of Texas – Pan American are gratefully acknowledged. H. Chen was supported in part by ARO W911NF-08-1-0409 and Louisiana Board of Regents NSF(2009)-PFUND-162.

References

[1] D. Brinza, and A. Zelikovsky, "2SNP: Scalable Phasing Method for Trios and Unrelated Individuals", *IEEE/ACM Trans. Computational Biology and Bioinformatics*, 5(2), pp. 313–318, 2008.

[2] A. Clark, "Inference of Haplotypes from PCR-Amplified Samples of Diploid Populations", *Mol. Bio. Evol.*, 7, pp. 111–122, 1990.

[3] T. M. Cover, and J. A. Thomas, "Elements of Information Theory", Wiley-Interscience, 1991.

[4] A. Gusev, I. Mandoiu, and B. Pasaniuc, "Highly Scalable Genotype Phasing by Entropy Minimization", *IEEE/ACM Trans. Computational Biology and Bioinformatics*, 5(2), pp. 252–261, 2008.

[5] K. Liang, and X. Wang, "A Deterministic Sequential Monte Carlo Method for Haplotype Inference", *IEEE Journal Selected Topics in Signal Processing*, 2(3), pp. 322–331, 2008.

[6] G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz, "SNPs Problems, Complexity and Algorithms", *LNCS*, 2161, pp. 182–193, 2001.

[7] S. Lin, D. J. Cutlera, M. E. Zwicka, and A. Chakravarti, "Haplotype Inference in Random Population Samples", *The American Journal of Human Genetics*, 71(5), pp. 1129–1137, Nov. 2002.

[8] D. MacKay, "Information Theory, Inference, and Learning Algorithms", Cambridge Press, 2003.

[9] T. Niu, Z. S. Qin, X. Xu, and J. S. Liu, "Bayesian Haplotype Inference for Multiple Linked Signal-Nucleotide Polymorphisms", *The American Journal of Human Genetics*, 70(1), pp. 157–169, 2002.

[10] A. Wald, *Sequential Analysis*, Wiley and Sons, New York, NY, 1947.

[11] R.-S. Wang, L.-Y. Wu, Z.-P. Li, and X.-S. Zhang, "Haplotype Reconstruction from SNP Fragments by Minimum Error Correction", *Bioinformatics*, 21, pp. 2456–2462, 2005.

[12] L. Wang, and Y. Xu, "Haplotype Inference by Maximum Parsimony", *Bioinformatics*, 19, pp. 1773–1780, 2003.

[13] E. P. Xing, M. I. Jordan, and R. Sharan, "Bayesian Haplotype Inference via the Dirichlet Process", *Journal of Computational Biology*, 14(3), pp. 267–284, 2007.

New Algorithm and Web Server for Finding Proteins with Similar 3D Structures

Zaixin Lu[1], Zhiyu Zhao[2], Sergio Garcia[1], Bin Fu[1]

[1]Department of Computer Science
University of Texas–Pan American, Edinburg, TX 78539, USA
lzaixin@broncs.utpa.edu, sergio.garcia@tecmonkeys.com, binfu@cs.panam.edu,
Phone: 956-381-3635, Fax: 956-384-5099

[2]Department of Computer Science
University of New Orleans, New Orleans, LA 70148, USA
zzha2@cs.uno.edu
Phone: 956-384-5099, Fax: 504-280 - 7228
Contact author: Bin Fu
Submitted to BIOCOMP'08
Key words: Protein 3D structures, Alignment, Searching

Abstract

We develop a new algorithm for finding similar protein structures in the protein databank. Our new algorithm is implemented to run in a single machine and it can return a set of protein structures that are similar to the given protein structure in a few minutes. In order to compare the performance with other web-servers, we develop a a symmetric difference model between the two sets of protein structures which are outputted by two different softwares when the same protein structure is given as input. It is compared with the similar tool of Secondary Structure Matching (SSM) at address http://www.ebi.ac.uk/msd-srv/ssm/ and shows competitive performance with SSM. Our algorithm has been implemented as a web-server at address: http://fpsa.cs.panam.edu/ for public access.

1 Introduction

Protein 3-D structures are widely believed to be related to their biological functions. It is particularly important to find out the structural similarity between different proteins. The number of protein structures is becoming larger and larger in the protein databank (PDB). As the current protein databank has more than 48,000 proteins and 100,000 chains, it is crucial to develop efficient computer algorithms and softwares to find similar protein structures from the protein databank.

A widely accepted idea of protein 3-D structure comparison is to align the C_α atoms in the protein backbones. A protein molecule is made up of one or more polypeptide backbones with specific side chains attached to them. The backbone of a protein is also known as a main chain which has a constant structure. The carbon atoms in the main chain that the side chains are attached to are known as alpha-carbons (C_α). A protein backbone is usually represented by one chain of C_α atoms with their 3-D coordinates. With this representation, the problem of protein 3-D structure alignment can be solved by finding out first the longest common sub chains between different protein backbone chains, then a proper rigid body transformation which translates and rotates one protein backbone chain so as to align it as close as possible to the other one. In recent years, various approaches on protein 3-D structure alignment have been presented (e.g. [5, 9, 11, 15, 14, 3, 6, 8, 17, 16]).

Protein query problem is to find the proteins with similar 3D structures in the protein databank, given a protein structure as the input. The input may be a protein structure, which is described in certain format (e.x. the PDB format), or a protein code in the protein databank. Algorithms for searching protein in the database for similar structures have been developed by multiple research groups [1, 7, 11, 12,

13, 15, 2, 4]. In particular, the methods developed in [1, 7, 11, 12, 13, 15, 2] belong to the hierarchical method.

It is easy to see that the algorithm for the pair-wise protein alignment in C_α atom level can be applied to find similar proteins from the protein database. This kind of algorithms is usually not fast enough to compare the input protein with all proteins in the databank. A natural approach is to exclude proteins that have greatly different structures from the input structure by using some simple method, and then apply more complicated algorithms to check the similarity with small number of proteins left.

In the hierarchical algorithm VAST [11], it builds a bipartite graph. Each node in one side of the graph is a pair of SSEs from the input protein, and each node in the other side of the graph is a pair of SSEs from the target protein. Connect two nodes between two sides if they can be aligned well. Their SSE alignment algorithm finds the maximal clique in the bipartite graph and extend it to C_α-atom level alignment by Gibbs sampling.

In this paper, we develop a practical algorithm for the protein query problem. Our main technical contribution is that we apply a new method to check the similarity for the secondary structures between two proteins. Our method for grouping the secondary structures is different from other protein secondary structure alignment algorithms like [11]. Our approach is based on finding the star which has a center of two pairs aligned secondary structures between two proteins. Add other pairs of secondary structures to the star if there exists a common rigid body transformation between the center and the new pair. Prune the star until it has a satisfactory RMSD. Finding a maximal star, which can be computed in linear time, is easier than a maximal clique. This method based on star was first used in our recent C_α-atom level alignment [18] and shows improvement over the existing alignment algorithms.

We found an efficient way to combine the secondary structure level alignment with the C_α-atom level alignment. The combination of two alignments are embedded into our protein query system so that it can find similar proteins in the protein databank of more than $100,000$ chains in a short time, and avoid missing similar structures.

The quality of protein query system is determined by how similar the list of output proteins to the input protein is . In order to compare the performance with other web-servers, we develop a model based on the symmetric difference between the two sets of protein structures which are outputted by two different softwares when the input is the same protein structure. It is compared with the similar tool of SSM and shows improved performance. It has been implemented as a web-server at address: http://fpsa.cs.panam.edu/.

2 Overview of Our Methods

Our algorithm has a series of filters. Given a protein 3D structure as input, the algorithm first exclude those proteins that have big difference in the number of C_α atoms in the protein backbone, the average distance from all C_α atoms of backbone to the center of protein backbone, or the statistics about the secondary structures. The second layer filter does the second structure sequence alignment. The third layer filter aligns the secondary 3D structures. The fourth layer filter uses a simplified version of our pair-wise protein 3D structure alignment algorithm developed by us recently [18], and does the protein 3D structure alignment.

3 Description of Algorithm

A straightforward method to find the similar structures in the protein databank is to apply a pairwise protein alignment software to check all of the protein structures saved in the database. Since the protein databank has a large number of protein structures, it would be very slow to check each structure carefully. Our algorithm has multiple phases to filter those structures that have weak similarity with the input protein structure in the early stage. When there are small number of candidate structures left, a more complicated pairwise algorithm is used in selecting the most similar protein structures.

3.1 Checking Off-line Information

We first reject those protein structures that have greatly different number of C_α atoms in the backbone, the structures that has large difference for the average distance from the C_α-atoms to the center of C_α-backbone, and the structures that have big difference in ratio of α-helix among all secondary structures entities (SSEs). The number of C_α atoms for all protein structures can be easily computed off-line. So are the average distance to the center, and the ratio of α-helix. This stage is very fast since those offline information is ready during the query and the decision can be made very quickly.

Let S_0 be the input protein structure. We often use the parameter Structure-list to represent a list of protein structures which will be selected by checking similar properties with input protein structure S_0. For protein structure S, define $C(S)$ to be the C_α-chain

of the backbone of S. Function Check-protein-size() checks if a target protein has a similar number of C_α atoms with the input protein S_0.

Check-protein-size(S_0, S)
Input: S_0 is the input protein structure, and S is another protein structure.
Output: true or false.
Begin
Let n_0 be the number of C_α atoms in $C(S_0)$.
Let n be the number of C_α atoms in $C(S)$.
If ($|n_0 - n| \leq$ empirical_value$\cdot \max(n_0, n)$)
then $true$.
return $false$.
End (of Check-protein-size)

For a list of points $p_1, \cdots, p_n$ in 3D space, its center is computed by $\frac{\sum_{i=1}^{n} p_i}{n}$. We have the function Check-average-distance-to-center() to check if the average distance from all C_α atoms to the center of $C_\alpha(S)$ is similar to that of $C_\alpha(S_0)$. If two structures are similar their average distances to center are also close.
Check-average-distance-to-center(S_0, S)
Input: S_0 is the input protein structure, and S is another protein structure.
Output: true or false.
Begin
Let c_0 be the center of $C(S_0)$.
Let d_0 be the average distance from the C_α atoms of $C(S_0)$ to c_0.
Let c be the center of $C(S)$.
Let d be the average distance from the C_α atoms of $C(S)$ to c.
If ($|d_0 - d| \leq$ empirical_value$\cdot \max(d_0, d)$)
Then $true$
Else return $false$.
End (of Check-average-distance-to-center)

The function Check-secondary-structure-statistics() is used to check the statistics information about the secondary structures such as the number of α-helixes and β sheets.
Check-secondary-structure-statistics(S_0, S)
Input: S_0 is the input protein structure, and S is another protein structure.
Output: true or false.
Begin
Let a_0 be the number of α-helixes in $C(S_0)$.
Let b_0 be the number of β-sheets in $C(S_0)$.
Let a be the number of α-helixes in $C(S)$.
Let b be the number of β-sheets in $C(S)$.
If ($|a_0 - a| \leq$ empirical_value$\cdot \max(a_0, a)$ and $|b_0 - b| \leq$ empirical_value$\cdot \max(b_0, b)$)
Then Return $true$.
Return $false$.
End (of Check-secondary-structure-statistics)

The function Select-via-offline-information() filters protein by the offline information. If the Structure-list is the list of proteins in the protein databank, there will be less than 20% proteins left after calling this function.
Select-via-offline-information(S_0, Structure-list)
Input: S_0 is the input protein structure, and Structure-list is the list of structures to be searched for similar proteins.
Output: a sublist of protein structures that each has similar average distance to the it center.
Begin
Let $L = \emptyset$.
For each protein structure S in L
Begin
If (Check-protein-size(S_0, S) and
Check-average-distance-to-center(S_0, S) and
Check-secondary-structure-statistics (S_0, S))
Then put S into L ($L = L \cup \{S\}$).
End (of For)
Return L.
End (of Select-via-offline-information)

3.2 Secondary Sequence Alignment

It has been observed that if two structures are similar, their secondary structure sequences can be well aligned at sequence level, where each secondary structure can be represented by either α for α helix or β for a β sheet. Using the α-β sequence alignment can reduce a large number of unrelated structures and greatly speed up the searching in the database.

The second structure sequences of all protein structures in the databank are extracted. For each protein, its secondary structure sequence has format $s_1 s_2 \cdots s_k$ such that each s_i contains the following information:

- α-β type.
- Number of C_α atoms in the secondary structure. Define $c_\alpha(s_i)$ be the number of C_α atoms in s_i.
- Two crucial points of the secondary structure.

Define the weight function such that $w_1(a, b) = \frac{\max(c_\alpha(a), c_\alpha(b))}{c_\alpha(a) + c_\alpha(b)}$ if the α-β type of a and b are different or one of them is a space, and $w_1(a, b) = \frac{|c_\alpha(a) - c_\alpha(b)|}{c_\alpha(a) + c_\alpha(b)}$ if the α-β type of a and b are the same.

An alignment of two sequences $a_1 \cdots a_n$ and $b_1 \cdots b_m$ of secondary structures is to add some

spaces, which is marked by '-', into both of them. The first sequence become $a'_1 \cdots a'_k$ and the second sequence becomes $b'_1 \cdots b'_k$, and for each $1 \le i \le k$, at least one of a_i and b_i is not a space. The total cost for the alignment that is from $a'_1 \cdots a'_k$ and $a'_1 \cdots a'_k$ is measured by $\sum_{i=1}^{k} w(a'_i, b'_i)$. The optimal alignment is to find the one with the least cost by the function $w()$.

Define $D(i, j)$ be the cost of an optimal alignment between $a_1 \cdots a_i$ and $b_1 \cdots b_j$. We have the following recursion, which implies $D(n, m)$ can be computed in $O(mn)$ time by dynamic programming method.

$$D(i,j) \quad = \quad \min \begin{cases} D(i-1, j-1) + w(a_i, a_j) \\ D(i-1, j) + w(a_i, -), \\ D(i, j-1) + w(-, b_j) \end{cases} \qquad (1)$$

Secondary-structure-sequence-alignment(S, S')

Input: S is the first protein structure, and S' is the second protein structure.

Output: an alignment for the secondary structure sequences between S and S'.

Begin

Let $s_1 s_2 \cdots s_k$ be the sequence of the first protein S.

Let $s'_1 s'_2 \cdots s'_{k'}$ be the sequence of the second protein S'.

Apply the dynamic programming with weight function $w()$.

Output the alignment with the best score.

End (of Secondary-structure-sequence-alignment)

We use function Select-via-secondary-structure-sequence() to select those proteins that have the secondary structure sequence to be well aligned with that of input protein structure S_0.

Select-via-secondary-structure-sequence(S_0, Structure-list)

Input: S_0 is the input protein structure, Structure-list is the list of structures to be searched for similar proteins.

Output: a sublist of protein structures that can be well aligned with S_0 according the secondary structure sequence alignment.

Begin

Let $L = \emptyset$.

For each protein structure S in the Structure-list

Begin

A =Secondary-structure-sequence-alignment(S_0, S).

If (alignment A is good enough) then put S into L.

End (of For)

return L.

End (of Select-via-secondary-structure-sequence)

3.3 3D Alignment for Secondary Structures

In this phase, we select those protein structures that have good geometric alignment by secondary structures. This phase is also fast since each protein has about 30 secondary structures in average. We just use two points to represent a secondary structure.

Build-Star$((s_1, s'_1), (s_2, s'_2), S, S')$

Input: S and S' are two protein structures, s_1 and s_2 are secondary structures in S, s'_1 and s'_2 are secondary structures in S', and there exists an rigid body alignment for (s_1, s'_1) and (s_2, s'_2).

Output: a star with center at $(s_1, s'_1), (s_2, s'_2)$.

Begin

Let Center=$\{(s_1, s'_1), (s_2, s'_2)\}$.

Let Star=Center.

For each pair secondary structure (s, s') between S and S'.

begin

If (there exists a rigid body transformation for center and (s, s')) Then Let Star=Star $\cup\{(s, s')\}$.

end (For)

Return Star.

End (of Build-Star)

The function Prune-star() deletes some pairs in a star until there exists an alignment with RMSD less than a threshold r.

Prune-star(Star, r)

Input: Star is a star of secondary structures, and r is a threshold.

Output: a new star of secondary structures that has rigid body alignment with RMSD no more than r.

Begin

While (RMSD(Star)$> r$)

Remove the pair (s, s') of Star that has the largest distance dist(s, s').

End (of Prune-star)

The function Secondary-structure-3D-alignment() align the 3D secondary structures between two protein structures S and S'. The method is based on building stars and pruning stars.

Secondary-structure-3D-alignment(S, S')

Input: S is a protein structure, S' is a protein structure.

Output: a 3D alignment between the secondary structures of S and S'.

Begin

L =Secondary-structure-sequence-alignment (S, S').

Best-star= $\emptyset$.

For each pair (s_1, s_2) of neighbor secondary structures in L

Begin

Star=Build-Star(s_1, s_2).

Star=Prune-star(Star, r)

if (size(best-star)< size(Star))

Then Best-star=Star.

End (of For)

Return best-star as an alignment.

End (of Secondary-structure-3D-alignment)

The function Select-via-secondary-structure-3D-alignment() selects those proteins that can be well aligned with S_0 by the Secondary-structure-3D-alignment function.

Select-via-secondary-structure-3D-alignment(S_0, Structure-list)

Input: S_0 is the input protein structure, and Structure-list is the list of structures to be searched for similar proteins.

Output: a sublist of protein structures that can be aligned with S_0 well.

Begin

Let $L = \emptyset$.

For each protein structure S in the Structure-list

Begin

A =Secondary-structure-3D-alignment (S_0, S).

If (alignment A is good enough)

Then Put S into L.

End (of For)

Return L.

End (of Select-via-secondary-structure-3D- alignment)

3.4 C_α-Atom Level Pair-wise Protein Alignment

In the bottom of our implementation, we find a suitable protein pairwise alignment algorithm which is first developed in our earlier work [18]. In this layer, the protein alignment algorithm should balance the speed and accuracy.

Pair-wise-alignment(S, S')

Input: S is the first protein structure, and S' is the second protein structure.

Output: an alignment between the C_α atoms in S and S'.

Begin

Let C be the C_α-atom chains S.

Let C' be the C_α-atom chains S'.

Find the similar local regions between two C_α-chains.

Let each local alignment be a node of a graph.

Add an edge between two nodes if they can share a common global rigid body transformation.

For each star in the graph

Begin

Prune those C_α-pairs until the RMSD is small enough.

End (of For)

Output the alignment with the largest number of C_α pairs.

End (of Pair-wise-alignment)

The function Select-via-atom-level-alignment() selects those proteins that can be well aligned with S_0 in the C_α atoms level.

Select-via-atom-level-alignment(S_0, Structure-list)

Input: S_0 is the input protein structure, and Structure-list is the list of structures to be searched for similar proteins.

Output: a sublist L of protein structures from Structure-list such that each protein structure in L has good C_α-atom alignment with S_0.

Begin

Let $L = \emptyset$.

For each protein structure S in the Structure-list

Begin

A =Pair-wise-alignment(S_0, S).

If (alignment A is good enough)

Then Put S into L.

End (of For)

Return L.

End (of Select-via-atom-level-alignment)

3.5 Combining Them Together

Now we put all of those layers of the algorithm together to form the entire algorithm. The first filter is based on the offline information, the second filter is based on the secondary structure sequence (α-β sequence) alignment, the third filter is based on the secondary structure 3D alignment, and the fourth filter is based on the C_α atoms alignment.

Search-proteins(S_0, Structure-list)

Input: S_0 is the input protein structure, and Structure-list is the list of structures to be searched for similar proteins.

Output: a list of proteins structures that are similar to S_0.

Begin

L_1 =Select-via-offline-information(S_0, Structure-list).

L_2 =Select-via-secondary-structure-sequence (S_0, L_2).

L_3 =Select-via-secondary-structure-3D-alignment(S_0, L_3).

L_4 =Select-via-atom-level-alignment(S_0, L_3).

Output L_4 as the list of proteins similar to S_0.

End (of Search-proteins)

4 Comparison with SSM

Our algorithm has been fully implemented and tested. It is available for public access at http://fpsa.cs.panam.edu/. It is running at a single machine, and will be supported by a cluster of machines soon.

4.1 Speed and Performance

We observed that given a protein structure as input for our software, it can output a list of protein structures that are similar to it among those available proteins in the protein databank.

The total number of protein chains is about $100,000$. After the top filter, which check the number of C_α atoms, the average distance to the center of C_α chain, and the ratio of SSEs, there are about $20,000$ structures left.

The second level filter aligns the secondary structure sequences and can filter 80% 90% structures from the output from the top layer. We often see that there are less than 2000 structures left after the second level filter.

The third level filter narrow down the number of structures to several hundreds in average. It depends on how many proteins are really similar to the input protein. The atom level alignment tool developed in [18] is efficient enough to align the input structures with those protein structures in several minutes.

4.2 Model of Comparison

Each web server outputs a list of proteins that are expected to have the maximal similarity with the input protein. When comparing with another web-server, we output the same number of results and check their symmetric difference. Assume that W_1 represents our software and W_2 represents another web-server. Given a protein structure p, $W_1(p)$ is the list of proteins similar to p by the server W_1. $W_2(p)$ is the list of proteins similar to p by the server W_2.

Let $W_1(p) - W_2(p)$ be the list of proteins that belong to $W_1(p)$ but not $W_2(p)$ and $W_2(p) - W_1(p)$ be the list of proteins that belong to $W_2(p)$ but not $W_1(p)$. The two lists are of the same length since we let $W_1(p)$ and $W_2(p)$ be of the same length.

According to [8], we use Q-score to measure the quality of alignment between two protein structures. The Q-score is defined by the formula below: $Q(p_1, p_2) = \frac{N^2_{align}}{(1+(RMSD/R_0)^2)N_1N_2}$, where N_{align} is the number of pairs of aligned C_α atoms, N_1 is the number of C_α-atoms in the protein p_1, N_2 is the number of C_α-atoms in the protein p_2, and R_0 is an empirical value (chosen at 3).

We select 88 proteins that are listed in [10] and belong to different categories such as $\alpha, \beta, \alpha/\beta$, and $\alpha+\beta$. Two Figures gives the comparison between SSM and our software based on the maximum and average Q-scores, respectively. The horizontal axis is the index of 80 protein and the vertical axis is the maximum Q-score value and average Q-score in two figures, respectively. Those proteins have the following names: 1cseI, 1dhr_, 1etu_, 1fx1_, 1paz_ 1pfkA, 1q21_, 1s01_, 1sbp_, 1sbt_, 1timA, 1treA, 1ula_, 1wayB, 2had_, 2liv_, 3gbp_, 5cpa_, 5p21_, 8abp_, 8atcA, 1ctf_, 1dnkA, 1eaf_, 1hsbA, 1ltsA, 1ltsD, 1ovb_, 1poc_, 1ppn_, 1rnd_, 1snc_, 1tfg_, 1tgsI, 2achA, 2bpa1, 2act_, 2sns_, 3il8_, 3rubS, 3sgbI, 3sicI, 4blmA, 4tms_, 9rnt_, 9rsaA. In our experiments, the eight proteins 1avhA, 1dnkA, 1ltsA, 1ovb_, 1pkfA, 1poc_, 2bpa1, and 3sicI have the same output between our software and SSM. Therefore, their are excluded in the results in Figure 1 and Figure 2.

Define $MaxQ_1(p) = \max\{Q(p, p')|p' \in W_1(p) - W_2(p)\}$. Define $MaxQ_2(p) = \max\{Q(p, p')|p' \in W_2(p) - W_1(p)\}$. The curve in the first figure is the function $MaxQ_1(p)$ (assume that a protein p and its index are the same). Each point (p, q) in the cure has the relationship $q = MaxQ_1(p)$. On the other hand, a point (p', q') for a dot in the first figure has the relationship $q' = MaxQ_2(p')$. Figure 1 shows the comparison between the missing in two softwares. Since most of the dots are below the curves, it indicates that SSM query tool has more serious missing problem than ours based on the maximum Q-score measure.

Define $AveQ_1(p) = \frac{\sum_{p' \in W_1(p) - W_2(p)} Q(p,p')}{|W_1(p) - W_2(p)|}$, where $|W_1(p) - W_2(p)|$ is the number of items in the set $W_1(p) - W_2(p)$. Define $AveQ_2(p) = \frac{\sum_{p' \in W_2(p) - W_1(p)} Q(p,p')}{|W_2(p) - W_1(p)|}$. The curve in the second fig-

ure is the function $AveQ_1(p)$ (assume that a protein p and its index are the same). Each point (p, q) in the cure has the relationship $q = AveQ_1(p)$. On the other hand, a point (p', q') for a dot in the second figure has the relationship $q' = AveQ_2(p')$. Figure 2 shows the comparison between the average missing in two softwares. Since most of the dots are below the curve, it indicates that SSM query tool has more serious missing problem than ours based on the average Q-score measure also.

5 Future Work

The algorithm only runs in a single machine. We are building a cluster of machine to support the protein query system. It will be ready soon. Our current C_α-atom level alignment algorithm is not fast enough. We try use a load balance way to speed up the server in a cluster of PCs.

References

[1] N. N. Alexandrov and D. Fischer. Analysis of topological and montopological structural similarities in the pdb: new examples from old structures. *Proteins*, 25:354–365, 1996.

[2] O. Camoglu, T. Kahveci, and A. K. Singh. Psi: Indexing protein structures for fast similarity search. In *Proceedings of Elventh International Conference on Intelligent Systems for Molecular Biology*, pages 81–83, 2003.

[3] L. P. Chew, K. Kedem, D. P. Huttenlocher, and J. Kleinberg. Fast detection of geometric substructure in proteins. *J. of Computational Biology*, 6(3-4):313–325, 1999.

[4] P.-H. Chi, G. Scott, and C.-R. Shyu. A fast protein structure retrieval system using image-based distance matrices and multidimensional index. In *Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering*, pages 522–532, 2004.

[5] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol*, 233:123–138, 1993.

[6] V. A. Ilyin, A. Abyzov, and C. M.Leslin. Structural alignment of proteins by a novel topofit method, as a superimposition of common volumes at a topomax point. *Protein Science*, 13:1865–1874, 2004.

[7] I. Koch, T. Lengauer, and E. Wanke. An algorithm for finding maximal common subtopolgies in a set of protein structures. *Journal of Computational Biology*, 3-2:289–306, 1996.

[8] E. Krissinel and K. Henrick. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst.*, D60:2256—2268, 2004.

[9] U. Lessel and D. Schomburg. Similarities between protein 3-d structures. *Protein Eng.*, 7(10):1175–87, 1994.

[10] W. Liu, F. Mao, L. Lai, and Y. Han. Protein fold recognition based on structural classification. *Acta Biophysica Sinica*, 15:126–136, 1999.

[11] T. Madej, J. F. Gibrat, and S. H. Bryant. Threading a database of protein cores. *Proteins*, 23:356–369, 1995.

[12] K. Mizguchi and N. Go. Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng.*, 8:353–362, 1995.

[13] S. D. Rufino and T. L. Blundell. Structure-based identification and clustering of protein families and superfamilies. *Journal of Comput. Aided. Mol. Dec.*, 233:123–138, 1994.

[14] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng.*, 11:739–747, 1998.

[15] A. P. Singh and D. L. Brutlag. Hierarchical protein superposition using both secondary structure and atomic representation. In *Proc. Intelligent Systems for Molecular Biology*, pages 284–293, 1997.

[16] J. Ye, R. Janardan, and S. Liu. Pairwise protein structure alignment based on an orientation-independent backbone representation. *Journal of Bioinformatincs and Computational Biology*, 4(2):699–717, 2005.

[17] Y. Ye and A. Godzik. Database searching by flexible protein structure alignment. *Protein Science*, 13(7):1841–1850, 2004.

[18] Z. Zhao and B. Fu. A flexible algorithm for pairwise protein structure alignment. In *Proceedings International Conference on Bioinformatics and Computational Biology 2007*, 2007.

Figure 1: Comparision of two softwares' Maximum Q−score of missing proteins

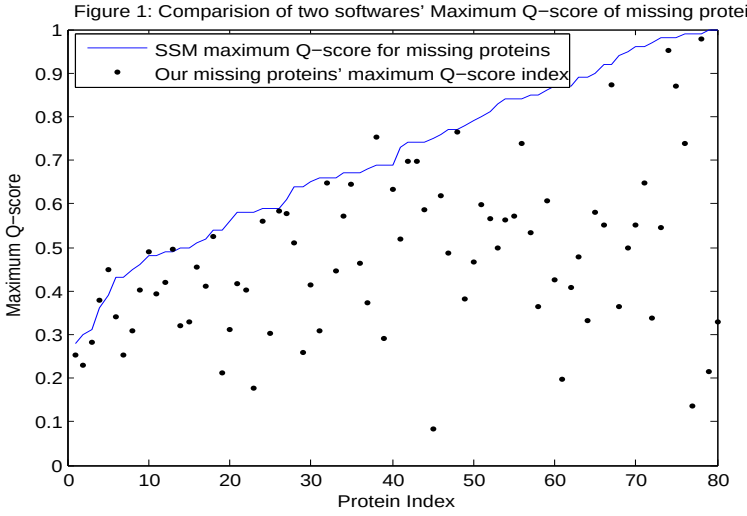
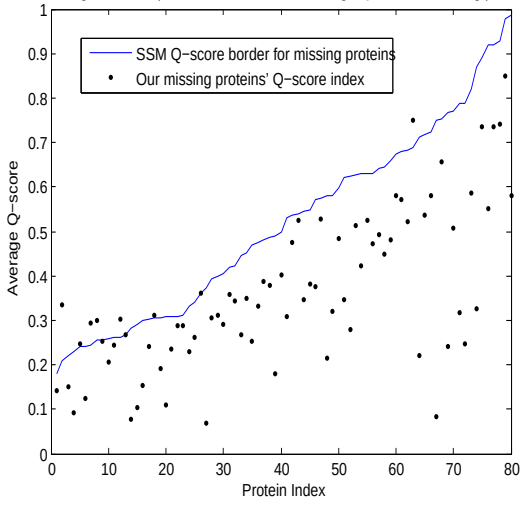


Figure 2: Comparision two softwares' average Q−score of missing proteins



Journal of Bioinformatics and Computational Biology *(to appear)*

Search Similar Protein Structures with Classification, Sequence and 3D Alignments*

Zaixin Lu

Department of Computer Science, University of Texas–Pan American, Address Edinburg, TX 78539, USA
lzaixin@broncs.utpa.edu

Zhiyu Zhao

Department of Computer Science, University of New Orleans, Address New Orleans, LA 70148, USA
sylvia@cs.uno.edu

Sergio Garcia

Department of Computer Science, University of Texas–Pan American, Address Edinburg, TX 78539, USA
sergio.garcia@tecmonkeys.com

Krishnakumar Krishnaswamy

Department of Computer Science, University of Texas–Pan American, Address Edinburg, TX 78539, USA
kkrishnasw@broncs.utpa.edu

Bin Fu

Department of Computer Science, University of Texas–Pan American, Address Edinburg, TX 78539, USA
binfu@cs.panam.edu

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

We have developed an algorithm and web tool to search similar protein structures in the PDB (Protein Data Bank). The algorithm is a combination of a series of methods including protein classification, geometric feature extraction, sequence alignment, and 3D structure alignment. Given a protein structure, the tool can efficiently discover similar structures from hundreds of thousands of structures stored in the PDB. Our experimental results show that it is more accurate than other well known protein search systems including PSI-BLAST, 3D-BLAST, and SSM in finding proteins that are structurally similar to the query protein, and its speed is also competitive with those sys-

*This research is supported by the National Science Foundation Early Career Award 0845376.

tems. The algorithm has been fully implemented and is accessible online at the address http://fpsa.cs.panam.edu/, which is supported by a cluster of computers.

Keywords: protein search; protein structure alignment.

1. Introduction

Searching for similarities between proteins has an important role in many biological and biomedical applications, such as disease diagnosis and drug design. It raises the need for tools that perform protein similarity searching to clarify the similarities between related or similar proteins. BLAST [2] (Eugene Myers., Stephen Altschul., 1990) is a classical sequence similarity search algorithm, and it has been widely used for searching proteins with similar amino acid sequences. In 1997, Altschul upgraded it to PSI-BLAST [3], which combines all closely related proteins into a general "profile" sequence as the query sequence for a search, and its accuracy performance has been improved. Unlike PSI-BLAST, most of the existing protein query tools search protein structures rather than protein amino acid sequences. It is because protein 3D structures are widely believed to be related to their biological functions, and in many cases, we cannot detect the similarity of two remotely homologous proteins by amino acid sequence comparison. Furthermore, during the evolution, protein 3D structures are more conserved than their amino acid sequences. The protein structure query problem is, given a protein structure as an input, to search for proteins with similar 3D structures in a protein structure database, such as the PDB [36] (Berman., 2000). The input may be a protein description file of a certain format (e.g. the PDB format) or a protein index in a certain database.

The major difficulty of searching for protein structure similarities is that the sizes of protein structure databases are growing rapidly while protein structure comparison algorithms are relatively slow. A widely accepted idea about the protein 3D structure comparison is to align the alpha-carbon (C_α) atoms in the protein backbones. In recent years, various approaches on the protein 3D structure alignment have been presented (e.g. [5, 7, 11, 13, 15, 20, 27, 28, 30, 31, 34]). However, the existing protein alignment algorithms are too computationally expensive for doing against-all alignments in a large protein structure database such as PDB. As of September 2008, there have been over 53,000 proteins, including more than 120,000 chains in the PDB. There can be multiple chains in a protein molecule and a protein chain is a basic unit for the protein structure comparison.

In the past, algorithms for searching similar protein structures have been developed by multiple research groups [1–4, 6, 8, 11, 14, 16, 18–21, 25, 27–29]. As we know, DALI [11] (Holm. and Sander., 1993) and CE [27] (Shindyalov. and Bourne., 1998) are two classical pairwise comparison methods. This kind of methods can provide us structure alignment results of high quality. However, their protein search programs have very slow response (or: very high response time) based on our experience and the evaluation in [4] and [18]. To improve the search speed, many methods have been designed to reduce the query time. Baker and Dauter (2004) developed SSM [5] which

uses Secondary Structure Match for the pairwise structure comparison; it reduces the comparison time a lot. In Addition, linear encoding has been applied to protein structural database searches recently. For instance, 3D-BLAST [29] developed by Yang and Tung (2006), can improve the comparison speed thousands times as the speed of CE and DALI. Similar methods include ProtDex2 [4], Sarst [18], and TopScan [16]. These methods improve the time performance greatly, and also have acceptable performance in accuracy. Due to the large size of the database and the high complexity of pair-wise protein structure alignment algorithms, a natural approach is to use some simple methods to exclude proteins that have greatly different structures with the input structure, and then apply more complicated algorithms to check the similarities with a small number of proteins left. Moreover, in our knowledge, a great number of proteins in the PDB are structurally similar. Therefore, driven by the need of a fast and accurate searching approach, we choose to first classify proteins into groups and select only one representative from each group of similar proteins. At present, several classification methods (e.g. [9, 12, 22–24, 26]) have been proposed, and their classified databases are available on the internet. They classify protein chains based on sequential, structural or functional similarities and aid the understanding of evolutionary relationships among them. However, the above classified databases cannot be used directly in a 3D structural similarities search system, because the proteins in the same group of those databases are not structurally similar enough and this will affect the query accuracy. Furthermore, the number of protein structures in the PDB increases rapidly while the above databases only issue their new version annually or even longer.

In this paper, we propose a practical approach for the protein structure query problem. In the approach, we first classify all proteins in the PDB into different groups such that each group only contains proteins with similar structures. Then, when a protein chain is specified, we combine geometric feature extraction, sequence alignment and structure alignment algorithms to search over the classified database, where each group has one structure to serve as a representative. In addition, our web tool has been implemented in a cluster of computers in order to increase its time performance, and is accessible at address http://fpsa.cs.panam.edu/ by the public. It can find similar proteins in the PDB (of more than 120,000 protein chains) in a short time (always less than one minute), and avoid missing similar structures. In our experiments, some exciting results have been observed when comparing our query tool with other well known protein search engines including PSI-BLAST, 3D-BLAST and SSM. The experimental results show that our tool is more accurate than other systems in finding proteins that are structurally similar to the query protein, and its speed is also competitive with them.

Our paper is organized as follows: Section 2 describes an outline of our protein search algorithm; aiming at accelerating the search process, Section 3 presents an offline classification for all the structures in the PDB; Section 4 further describes a geometric filter that consists of a series of layers to do geometric comparison between the candidate proteins and the query protein; Section 5 focuses on a sequence filter

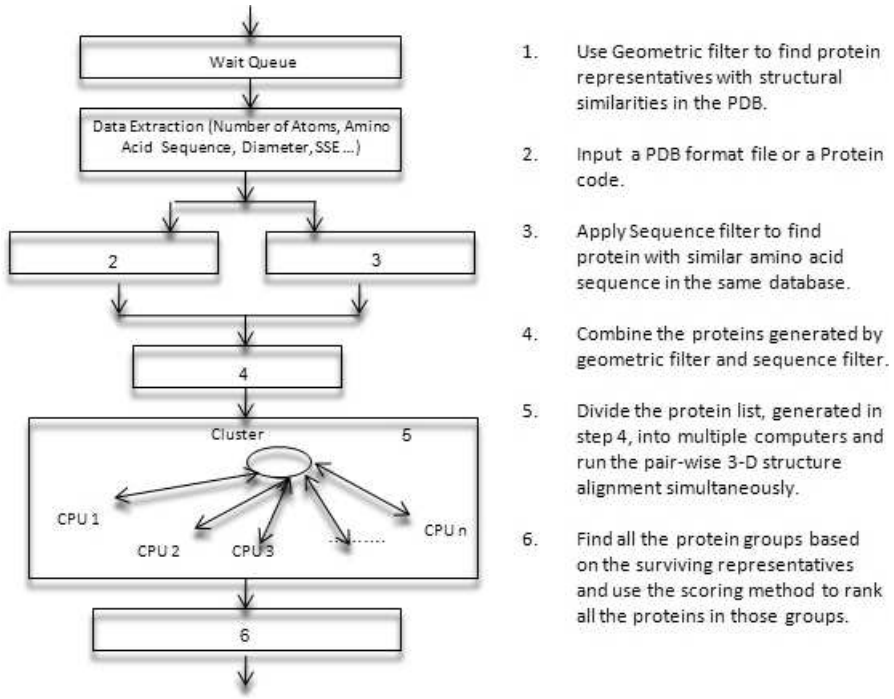


Fig. 1. System Process Diagram

based on the BLAST, a classical sequence similarity search method, to discover those proteins that have similar sequences and have been excluded by the geometric filter; Section 6 gives a brief introduction to our 3D structure comparison algorithm; Section 7 shows our distributed computer implementation; Section 8 discusses a score that is used for ranking the proteins to be sent to the output list; Section 9 proposes an evaluation model for the search performance with our experimental results, and compares our tool with other well known systems.

2. Outline of Our Approach

Our method is a combination of sequence alignment and geometric alignment. Its speed is improved by grouping proteins with similar structures and using one representative structure for each group. A brief overview about our method is as follows:

1. An offline phase partitions the protein 3D structures in the database into groups so that each group contains those proteins with similar structures. One representative structure is selected from each group.

2. When an input protein is given, use the BLAST algorithm to search proteins with similar amino acid sequences and put them into list L_1.

3.Use several layers of geometric filters to check all the representative proteins among the classified protein database and exclude dissimilar proteins. Put all those similar representative proteins into list L_2.

4. Put all the representative proteins whose groups are related to proteins in L_1 into L_3, and let list $L_4 = L_3 - L_2$ (which is the proteins in L_3, but not in L_2).

5. Use a simplified version of our pair-wise 3D structure alignment algorithm to check structural similarities between the input protein 3D structure and each structure in L_2 and L_4. Output those groups whose representatives are structurally similar to the input protein. This is the most time consuming part; therefore, it is implemented in a cluster of computers.

The first filter, called "sequence filter", is based on the sequence alignment of BLAST, and the second filter, called "geometric filter", is based on some simple geometric comparisons. A system process flow is shown in Figure 1.

3. An Offline Classification

As the protein structure database now contains more than 120,000 protein chains, and this number is increasing quickly, it is very time-consuming to do the all-against-all pair-wise alignment for the classification. So we first use our geometric filter to divide the entire protein database into small groups based on protein sizes, α-helix/SSE (secondary structure entity) ratios, and secondary structure sequence similarities, since it is easier to do the partition based on all-against-all pair-wise alignment in a smaller collection of proteins.

Two basic measures are often used for comparing the protein structural similarity, the alignment length and the RMSD value. The Q-score, proposed in [5] for measuring the similarity between two protein chains, and is defined by the formula : $Q(p_1, p_2) = \frac{N_{align}^2}{(1+(RMSD/R_0)^2)N_1N_2}$, where N_{align} is the number of pairs of aligned C_α atoms, N_1 is the number of C_α-atoms in the protein p_1, N_2 is the number of C_α-atoms in the protein p_2, and R_0 is an empirical value (chosen at 3). It was found that different servers agree reasonably well on this score. Here, we believe that the proteins with Q-score higher than 0.8 are very similar, and use our algorithm proposed in [32,33], to calculate the Q-score between the related proteins.

1. Partition the PDB roughly: In order to reduce the computational time, we first use the geometric filter to roughly partition the protein structures in the PDB into a small number of groups such that each group contains hundreds or thousands of structures.

2. Partition the groups: Use our pair-wise alignment algorithm [33] and an earlier version of this search engine [19] to check the structural similarity among the protein chains in each group generated in step one. If all the protein chains are similar, we just keep this group; otherwise, the group is partitioned into several new groups so that each group only contains proteins with similar structures.

3. Select the representatives: For each group, one protein chain is elected as the representative. In order to get the center of each group, we align each protein with

all other proteins in that group and calculate the sum of the Q-scores. A protein structure with the highest sum of the Q-score is selected as the representative.

4. Merge the groups: Check the similarities among those representatives and merge the groups into a new group if their representatives are very similar. Repeat step 3 for the new group.

5. Put in new proteins: For new structures, each of them is added to an existing classification group if our search engine can find such a group that contains similar structures with the new protein. Otherwise, a new group is created for it.

Our database synchronizes itself with the PDB and currently all the protein structures are partitioned into about 20,000 groups. This is much less than the number of structures in the PDB and can greatly reduce the number of structure comparisons.

4. Geometric Filter

Our geometric filter algorithm has multiple phases to reject protein structures that are dissimilar to the input protein structure. When there is a small number of candidate structures left, those candidates are assigned into multiple computers through a network and a relatively complicated pair-wise 3D structure alignment algorithm runs on those computers simultaneously to check the candidate proteins for structural similarity in a more accurate sense.

First, we reject those protein structures that have greatly different numbers of C_α atoms in their backbones, structures that have greatly different radii (for each protein we define a radius as an average distance from its C_α-atoms to the centroid of its backbone), and structures that have greatly different α-helix/SSE (secondary structure entity) ratios.

Second, it has been observed that, if two structures are similar, their secondary structure sequences can be well aligned at the sequence level. Therefore, we have designed an efficient sequence alignment algorithm based on dynamic programming to check secondary structure sequence similarities and reject protein structures with dissimilar sequences.

In the third layer, we select those protein structures that have good secondary structure alignments in the geometric sense. Therefore, some geometric alignments at the secondary structure level are designed and performed based on the steps below:

1. Let $s_1 s_2 \cdots s_k$ be the secondary structure sequence of the first protein S. 2. Let $s'_1 s'_2 \cdots s'_{k'}$ be the secondary structure sequence of the second protein S'. 3. Do the sequence alignment between $s_1 s_2 \cdots s_k$ and $s'_1 s'_2 \cdots s'_{k'}$, put each matched pair (s_i, s'_j) into L. 4. Use the algorithm Build-Star, which was designed in our earlier work [19], to build a Star for each pair (s_i, s'_j) in L. 5. Find the best Star as the secondary structure 3D alignment result of protein structures S and S'.

5. Sequence Filter

Our geometric filter is efficient in keeping those structures globally similar to the input protein and it speeds up the query greatly. But it also decreases the sensitivity of the search. For example, structures with remarkably different sizes (compared to the query structure) will be removed by the geometric filter. As we have observed, the geometric filter misses some good matches in the following two cases, especially when the searching is performed in multiple protein domains.

1. The first layer of the geometric filter misses a protein structure that is similar to a substructure in the input structure or contains a substructure similar to the input. Figure 2 is an example, where the first protein is much bigger than the second, but they partially match well.

2. Because our second and third layers of the geometric filter check the secondary structure similarities, the accuracy of secondary structure information is important for the final result. We use the Dictionary of Protein Secondary Structure (DSSP [35]) to extract secondary structure information. It works well in most cases; however, occasionally it brings us trouble. For instance, both proteins 132l:A and 1e8l:A have 129 C_α atoms. We used our alignment algorithm to compare their structures and found that 128 C_α atom pairs match well with RMSD=1.52. Thus, proteins 132l:A and 1e8l:A are structurally similar, but their secondary structures (as defined by DSSP) are $\alpha\alpha\beta\beta\alpha\alpha$ and $\alpha\alpha\alpha\alpha\alpha$ respectively. So, it is possible that the second layer of the geometric filter, which checks the secondary structure sequence similarity, or the third layer of the geometric filter, which does the secondary structure 3D alignment, rejects this kind of proteins, and introduces missing problems.

To avoid the above missing problems, relaxing the threshold of the filters may be a possible solution, but it will increase the use of the pair-wise alignment algorithm and make query time much longer. The proteins with similar amino acid sequences often have common 3D structures, and some well known sequence alignment algorithms are very fast. We use BLAST, one of the widely used sequence search algorithms, to find proteins for amino acid sequence similarities to compensate our geometric filter. Since proteins with weak sequential similarity may have strong structural similarity, we keep the proteins found by BLAST with E-value less than 1.0 for the pair-wise 3D alignment in the next phase.

In our experiment, BLAST often gives us hundreds of proteins with similar amino acid sequences immediately, and some of them (always less than 10) are not in the candidate list generated by the geometric filter. We often see that the sequence filter does find some protein structures that are missed by the geometric filter. For example, when doing query for protein 1hil:C, protein 1a14:L in Figure 2 was rejected by the geometric filter due to diameter and size comparisons. Its E-value calculated by BLAST is $1e^{-54}$, which shows that amino acid sequences between the two proteins 1hil:C and 1a14:L are very similar. Therefore protein 1a14:L, which is previously missed by the geometric filter, is found by the sequence filter. They are indeed structurally similar based on our structure alignment algorithm. As a

second example, when querying protein 132l:A, protein 1e8l:A was rejected by the geometric filter because of its different secondary structure sequence. However, its sequence is almost the same as that of protein 132l:A, and this is discovered by the sequence filter. Therefore, we believe that the geometric filter and the sequence filter well compensate each other. Our experiments show that combining the two filters reduces the missing problem without spending much time.

6. 3D Structure Comparison

In the bottom of our implementation, we use a suitable protein pair-wise structure comparison algorithm to check the surviving structures carefully and select proteins with similar structures. Our comparison algorithm, a simplified version of the method [33], first searches for a set of local alignments. Each local alignment consists of a series of consecutive C_α atom pairs in the backbones of two proteins. It then organizes the local alignments into a graph with each local alignment being a vertex. The connectivity between vertices is determined by the consistency relationship between local alignments. Two local alignments are said to be consistent if they share a common rigid body transformation. With this graph representation, a global alignment is an optimal group of local alignments sharing a common rigid body transformation. However, grouping mutually consistent local alignments is equivalent to finding cliques in a graph, which is an NP-complete problem. We have simplified the problem as looking for "stars" rather than cliques in a graph. A star is a set of vertices including a center and all other vertices that are connected to the center vertex. Since any clique must be included in some star, it reduces the computational complexity to $O(n^2)$, where n is the number of vertices in the graph. The next sub-phase works on those stars one by one. For each star, combine all of the pairs of matched points in each local alignment and look for an alignment to align as many as possible matched points. Delete the pair that has the worst violation (largest distance between its two matched points from two backbones), repeat the deletion unit until we obtain a global transformation with sufficient accuracy (small RMSD).

7. Distributed Computing

In the last step of implementation we assign the surviving proteins to multiple computers to perform the pair-wise structure alignments simultaneously. However, if not scheduled properly, a distributed system can decrease the overall reliability of computations because the unavailability of a node can result in disruption of other nodes. Instead of just evenly assigning proteins to the nodes of our cluster, at the beginning, the front node assigns a small number of proteins to each available machine of the cluster, then whenever a machine is free (i.e. it has completed its task), the front node will calculate the speed of that machine and send a certain number of proteins to it. This procedure is repeated until all the computation tasks have been completed.

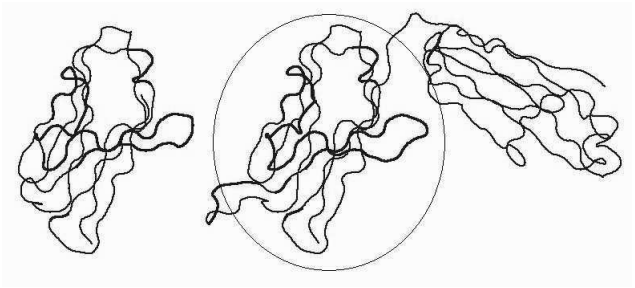


Fig. 2. 3D Structures of Protein 1hil:C and 1a14:L

1hil:C (the big one) has 217 C_α atoms, and 1a14:L (the small one) has 104 C_α atoms only, but it can be well superimposed with the left part of 1hil:C and their $RMSD$ is only 0.8.

8. Structure Similarity Score

How to rank output proteins as of strong to weak similarities is also an important factor of the protein query performance; it is better to generate a list of proteins ordered by similarity scores rather than to just output the results, and many scores have been proposed to rank results for a protein search system. Example scores include Z-score, P-score and E-value. Because the search engine of our web tool is aimed for finding proteins with similar structures, we want to use a protein structure related score to do the ranking. As mentioned previously, Q-score, which takes both $RMSD$ and alignment length into account, is a good score for measuring the overall similarity between two proteins. In other words, if the Q-score is high, two proteins can be superimposed well. However, based on our observations, when a protein is partially similar to another one, as shown in Figure 2, Q-score does not work well enough to reflect this kind of similarity. Therefore, we use a modified Q-score: N-score $= \frac{alignmentLengh}{[1+(\frac{RMSD}{3.5})^2]\times min(backboneLength1, backboneLength2)}$ as a criterion to rank the proteins. This score shows good performance in our experiment.

9. Experiments and Comparisons with Other Systems

In this section, we show the experimental results for our system implementation and its comparisons with other similar systems which are accessible online. The quality evaluation of protein search is based on its accuracy, miss ratio, and speed.

9.1. *Evaluation of accuracy*

We know that the quality of a protein query system is determined by how similar the list of output proteins is to the input protein, how to rank the output proteins and how many similar proteins are missing. In order to compare our performance with that of other web-servers, we select the newest version of SCOP [23] (1.73) as the target database and 87 query proteins, which were selected by Liu et. all [17] to

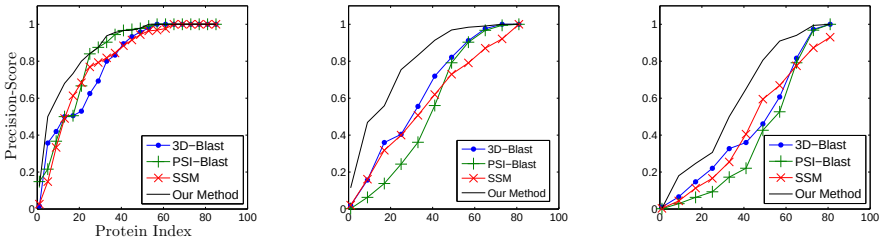


Fig. 3. Precision Curves of Multiple Methods

The left curves are the precisions of multiple methods at the SCOP domain level. The middle curves are the precisions of them at family level. The right curves are the precisions of them at super family level.

do their folding structural classification experiment. There are 25 all-α proteins, 22 all-β proteins, 24 α/β proteins, and 28 α+β proteins, totally 99 proteins in their test set, but only 87 of them are still present in the current PDB. We use these different categories of proteins to do the queries on our search engine and also on 3D-BLAST [29], PSI-BLAST [3], and SSM [5]. As we know the ProtDex2 [4], Sarst [18], and TopScan [16] are also efficient protein search systems, however they have not updated their database for a long time.

In the SCOP database, proteins in the same domain are the most similar proteins, and then are the proteins that belong to the same family and super-family. People also classify structurally similar proteins into the same fold. Therefore, we develop the following model to check the qualities of an output list:

1. A protein is regarded as "relevant" if it belongs to the same SCOP classification unit (domain, family, super family or fold) as the query protein does. 2. Let N be the total number of relevant proteins. 3. Let n_1 be the number of relevant proteins in the top N proteins of the output list. 4. Let n_2 be the number of relevant proteins in all the output proteins. 5. Precision-Score = $\frac{n_1}{N}$. 6. Missing-Rate = $\frac{N-n_2}{N}$.

Therefore, the Precision-Score is between 0 and 1, and the quality of a ranked output list is directly based on it. The Missing-Rate is also between 0 and 1, and the missing problem of a search engine is in relation to it.

According to the 87 query results, our structural similarity search method shows better performance both in finding similar proteins at the SCOP domain and family levels and in finding remote homologies. The precision curves of four methods look highly similar at SCOP domain level. It is reasonable because experts prefer to use strong sequence similarities to classify the SCOP domains and proteins with similar sequences always have very similar structures. At family level, our method is the most accurate one. 3D-BLAST which is based on linear encoding algorithm is the second. Nevertheless, all the similarity search methods including ours have problem detecting related proteins at super family level; a super family in SCOP is much larger than a family or a domain and experts classify proteins into the same super

Table 1. Statistics on the experimental results

Number of valid cases	3D-BLAST 87	PSI-BLAST 87	SSM 87	Our Method 87
average precision-domain level	79.07%	83.74%	80.01%	87.69%
average precision-family level	67.65%	59.90%	61.42%	81.56%
average precision-superfamily level	49.69%	43.56%	48.04%	63.81%
average missing rate-domain level	13.56%	14.31%	15.85%	3.02%
average missing rate-family level	31.12%	39.75%	38.21%	17.15%
average missing rate-superfamily level	49.24%	56.29%	51.83%	35.33%
serious missing problem-domain level	8(9.1%)	10(11.4%)	11(12.6%)	0(0.0%)
serious missing problem-family level	26(29.8%)	36(41.3%)	31(35.6%)	8(9.1%)
serious missing problem-superfamily level	50(57.4%)	52(59.7%)	42(48.2%)	32(36.7%)
100% precision-domain level	34(39.0%)	33(37.9%)	26(29.8%)	37(42.5%)
100% precision-family level	21(24.1%)	15(17.2%)	8(9.2%)	22(25.2%)
100% precision-superfamily level	11(12.6%)	10(11.4%)	6(6.9%)	14(16.1%)
0% missing rate-domain level	54(62.0%)	44(50.5%)	35(40.2%)	75(86.2%)
0% missing rate-family level	25(28.7%)	17(19.5%)	9(10.3%)	28(32.1%)
0% missing rate-superfamily level	15(17.2%)	11(12.6%)	6(6.8%)	17(19.5%)

family based on remote homologies; therefore, it is more complicated, and hence, the precisions of all four methods are not high: about 36 percent of our query results have serious missing problems (missing rate greater than 50%), and other methods have more serious missing problems than ours at this level.

9.2. *Performance of N-score*

In order to assess the reliability of N-score, we do the statistics of two common metrics, precision and recall, for various N-scores at superfamily and family levels. Precision is defined as n/N and recall is defined as n/T, where n is the number of true proteins (from the same family or superfamily) in the result list, N is the total number of proteins in the result list, and T is the total number of proteins in the family or superfamily of the input protein. According to the data in Table 2, when N-score is higher than 0.5, the average precision is greater than 90% for both levels, and their recalls are higher than 67.88% and 49.29% respectively.

Table 2. Statistics on the Reliability of scores

N-Score	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
avg.recall(%)-family	19.82	24.82	38.17	50.21	67.88	78.99	89.67	93.47
avg.precision(%)-family	99.78	98.82	97.86	95.21	91.55	81.75	64.15	43.63
avg.recall(%)-superfamily	14.33	18.23	28.04	36.85	49.29	62.24	73.86	82.42
avg.precision(%)-superfamily	99.91	99.47	99.30	99.14	97.75	94.05.15	80.70	58.29

9.3. *Experiment on incomplete structures*

Due to some unexpected reasons, sometimes experts cannot get all the coordinates of residues for a protein, therefore about 23 percent of proteins in the PDB are incomplete (based on our statistics, more than 25,000 protein chains in the current PDB have missing residues). It is difficult to do the structural similarities search

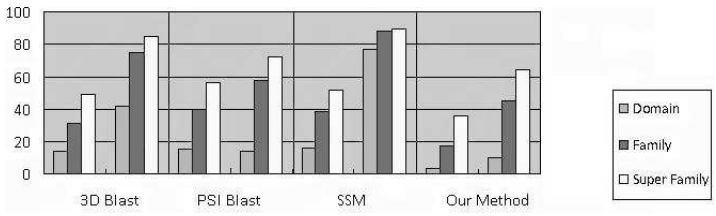


Fig. 4. Average Missing Rate of General Proteins and Incomplete Proteins.

For the four methods, the left columns of each method are the average missing rates for general protein queries and the right columns are the average missing rates for incomplete protein queries.

for a protein with lots of missing residues. To test the performance on incomplete structures, we have repeated the same experiment by using proteins whose backbones have a large number of missing residues. There are 38 protein chains in the PDB with missing residues ratio higher than 5 percent, and 20 have been selected by us as the query proteins because the other 18 of the 38 proteins are not in the current SCOP (1.73) database. As the result shown in Figure 4, our method has a lower missing rate than others for incomplete proteins, and in contrast to the result of the previous experiment, PSI-BLAST is affected less than others because it does not care the missing coordinates of residues. The tool 3D-BLAST, which uses liner encoding method, and SSM, which is based on pair-wise alignment, do have weak performance for queries with incomplete proteins.

9.4. *Improvement with classification*

In this paper, the goal of classification technologies is not only to advance the query speed, but also to obtain better results. Classification technologies have been widely used in many database search and management tools to speed up the search; however, because the protein query problem is more complex and difficult than a general search, we hold that a good classification database can improve the result accuracy. We have attempted to use our classification procedure to refine the results of 3D-BLAST, PSI-BLAST and SSM, and it does have improved their results. The refinement was performed in the following steps: 1. Let list L_1 be a pre-refined output list (it can be the protein list outputted by PSI-BLAST, 3D-BLAST or SSM). 2. Put all the proteins whose groups are related to the proteins in L_1 into L_2 (a protein and a group is "related" if the protein is in the group). 3. Use our protein structural similarity scoring method to rank proteins in L_2 and output the ranked proteins as the refined result list. 4. Let S_1 and S_2 be the Precision-Score and Missing-Rate of the pre-refined result list separately, and S'_1 and S'_2 be the Precision-Score and Missing-Rate of the refined result list. 5. The precision increment = S'_1 - S_1 and the missing rate decrement = S_2 - S'_2. As shown in Table 3, at family level, 45, 46, and 64 queries in the 87 cases of 3D-BLAST, PSI-BLAST, and

SSM, respectively, are improved by using our classification refinement procedure; the percentages are 51.72%, 52.87%, and 73.56%, respectively. The average precision increment is around 10% for all the three methods, with the maximum increment being more than 50%. Additionally, the data in Table 1 show that 3D-BLAST has 21 queries with 100% precision at family level, PSI-BLAST has 15 and SSM has 8. Since it is impossible to improve the results of those cases, without considering those queries, only few results are unimproved in our experiment. Therefore, the experiment shows that classification technologies help obtain better results.

Table 3. Statistics on the refinement results

	3D-BLAST	PSI-BLAST	SSM
number of cases	87	87	87
cases with higher precision-family level	45(51.72%)	46(52.87%)	64(73.56%)
cases with lower missing rate-family level	42(48.28%)	47(54.02%)	64(73.56%)
maximum precision increment-family level	55.12%	62.95%	82.60%
maximum missing rate decrement-family level	64.28%	62.95%	82.61%
average precision increment-family level	8.82%	8.67%	11.84%
average missing rate decrement-family level	9.92%	8.54%	11.89%

9.5. *Evaluation of speed*

We have used the web servers of 3D-BLAST, PSI-BLAST and SSM to do the 87 queries on the entire PDB database and following is a list of the query times. The web servers PSI-BLAST, SSM, 3D-BLAST, and our web server have average query time of 21, 26, 49, and 32 seconds respectively. In addition, our results contain an alignment length and an $RMSD$ value for every output protein. Although PSI-BLAST and 3D-BLAST do not have these data, they are the most important measures for comparing protein structural similarities. The web-server 3D-BLAST is the slowest one with an average query time of 49 seconds. In our knowledge, other search engines such as CE [27] and DALI [10] which are based on one-against-all pair-wise alignment algorithms need hours to days to complete the queries.

10. Conclusions

We have developed an efficient protein structural similarity search tool by a combination of sequence alignment, geometric alignment and structure classification. The sequence filter and geometric filter can compensate well in excluding dissimilar protein structures and the classified database does make the query much faster. The tool can return a list of similar proteins with the input protein in a short time. Our experiment result shows that it is more accurate than other well known systems in finding proteins that are structurally similar. However, the experiment result also shows that all the methods, including ours, have weakness in finding remote homologies. An interesting research in the future is to develop an efficient filter technology to detect proteins with weak structural similarities.

11. Acknowledgments

We are grateful to anonymous referees for their helpful comments.

References

1. N. N. Alexandrov and D. Fischer. Analysis of topological and montopological structural similarities in the pdb: new examples from old structures. *Proteins*, 25:354–365, 1996.
2. S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 3:403–410, 1990.
3. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 17:3389–3402, 1997.
4. Z. Aung and K. L. Tan. Rapid 3D protein structure database searching using information retrieval techniques. *Bioinformatics*, 20(7):1045–1052, 2004.
5. E. N. Baker and Z. Dauter. Secondary-structure matching, a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica*, 6:2256–2268, 2004.
6. O. Camoglu, T. Kahveci, and A. K. Singh. Psi: Indexing protein structures for fast similarity search. In *Proceedings of Elventh International Conference on Intelligent Systems for Molecular Biology*, pages 81–83, 2003.
7. L. P. Chew, K. Kedem, D. P. Huttenlocher, and J. Kleinberg. Fast detection of geometric substructure in proteins. *Journal of Computational Biology*, 6:313–325, 1999.
8. P.-H. Chi, G. Scott, and C.-R. Shyu. A fast protein structure retrieval system using image-based distance matrices and multidimensional index. In *Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering*, pages 522–532, 2004.
9. U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Science*, 3:522–524, 1994.
10. L. Holm, S. Kaariainen, P. Rosenstrom, and A. Schenkel. Protein structure database searching by dalilite v. 3. *Bioinformatics*, 2008.
11. L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233:123–138, 1993.
12. L. Holm and C. Sander. The fssp database of structurally aligned protein fold families. *Nucleic Acids Res.*, 22:3600–3609, 1994.
13. V. A. Ilyin, A. Abyzov, and C. M.Leslin. Structural alignment of proteins by a novel topofit method, as a superimposition of common volumes at a topomax point. *Protein Science*, 13:1865–1874, 2004.
14. I. Koch, T. Lengauer, and E. Wanke. An algorithm for finding maximal common subtopolgies in a set of protein structures. *Journal of Computational Biology*, 3-2:289–306, 1996.
15. U. Lessel and D. Schomburg. Similarities between protein 3D structures. *Protein Engineering*, 7(10):1175–1187, 1994.
16. Martin, A. C. R. The Ups and Downs of Protein Topology; Rapid Comparison of Protein Structure. *Protein Engineering*, 13:829-837, 2000.
17. W. Liu, F. Mao, L. Lai, and Y. Han. Protein fold recognition based on structural classification. *Acta Biophysica Sinica*, 15:126–136, 1999.
18. W.-C. Lo, P.-J. Huang, C.-H. Chang, and P.-C. Lyu. Protein structrual similarity search by ramachandran codes. *BMC Bioinformatics*, 2007.
19. Z. Lu, Z. Zhao, S. Garcia, and B. Fu. New algorithm and web server for finding proteins with similar 3d structures. In *the 2008 International Conference on Bioinformatics & Computational Biology (BIOCOMP'08)*, pages 674–680, 2008.

20. T. Madej, J. F. Gibrat, and S. H. Bryant. Threading a database of protein cores. *Proteins*, 23:356–369, 1995.
21. K. Mizguchi and N. Go. Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Engineering*, 8:353–362, 1995.
22. K. Mizuguchi, C. M. Deane, T. L. Blundell, and J. P. Overington. Homstrad: A database of protein structure alignments for homologous families. *Protein Science*, 7:2469–2471, 1998.
23. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.
24. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath-a hierarchic classification of protein domain structures. *Structure*, 5:1093–1108, 1997.
25. S. D. Rufino and T. L. Blundell. Structure-based identification and clustering of protein families and superfamilies. *Journal of Comput. Aided. Mol. Dec.*, 233:123–138, 1994.
26. C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. In *Proceedings Struct. Funct. Genet.*, pages 56–68, 1991.
27. I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Engineering*, 11:739–747, 1998.
28. A. P. Singh and D. L. Brutlag. Hierarchical protein superposition using both secondary structure and atomic representation. In *Proc. Intelligent Systems for Molecular Biology*, pages 284–293, 1997.
29. J.-M. Yang and C.-H. Tung. Protein structure database search and evolutionary classification. *Nucleic Acids Research*, 1:3646–3659, 2006.
30. J. Ye, R. Janardan, and S. Liu. Pairwise protein structure alignment based on an orientation-independent backbone representation. *Journal of Bioinformatics and Computational Biology*, 4(2):699–717, 2005.
31. Y. Ye and A. Godzik. Database searching by flexible protein structure alignment. *Protein Science*, 13(7):1841–1850, 2004.
32. Z. Zhao and B. Fu. A flexible algorithm for pairwise protein structure alignment. In *Proceedings International Conference on Bioinformatics and Computational Biology 2007*, pages 16–22, 2007.
33. Z. Zhao, B. Fu, F. J. Alanis, and C. M. Summa. Feedback algorithm and web-server for protein structure alignment. *Journal of Computational Biology*, 15(5):505 – 524, June 2008.
34. M. Jiang, Y. Xu, and B. Zhu. Protein structure-structure alignment with discrete frechet distance. *Journal of Bioinformatics and Computational Biology*, 6(1):51 – 64, February 2008.
35. W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12): 2577 – 2637, 1983.
36. H. Berman, J. Westbrook, and Z. Feng. The Protein Data Bank. *Nucleic Acids Res*, 28(1): 235 – 242, January 2000.

Seebeck coefficient and thermal conductivity in doped C_{60}

Wendong Wang,[1] Zhenjun Wang,[1] Jinke Tang,[1,a)] Shizhong Yang,[2] Hua Jin,[2] Guang-Lin Zhao,[2] and Qiang Li[3]
[1]*Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA*
[2]*Department of Physics, Southern University and A & M College, Baton Rouge, Louisiana 70813, USA*
[3]*Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, New York 11973, USA*

(Received 7 January 2009; accepted 4 March 2009; published online 27 March 2009)

Pressed bulk samples of C_{60} doped with P, Co, Al, and Bi have been investigated for their thermoelectric properties. These samples show extremely low thermal conductivity, typically in the range of 0.1–0.3 W/Km at room temperature. The Seebeck coefficients of Co, Al, and Bi doped C_{60} solids are in the tens of μV/K; however, for P doped C_{60} samples, a very large Seebeck coefficient in the order of 10^3 μV/K was observed. The value of the Seebeck coefficient seems to depend sensitively on the P concentration and changes sign upon annealing at 100 °C. *Ab initio* density functional theory calculations show that the calculated electronic structures and the activation energies strongly depend on the dopants in C_{60} solids. The high Seebeck coefficient in studied P doped C_{60} is due to the system's unique dopant and concentration. © *2009 American Institute of Physics.*
[DOI: 10.1063/1.3106303]

I. INTRODUCTION

Advanced thermoelectric materials will play an important role in energy harvesting from waste heat, geothermal energy, as well as solar energy. Solar energy can be converted into electricity in a standalone thermoelectric generator or in devices that combine with solar cells. High Seebeck coefficient and electrical conductivity and low thermal conductivity are required for advanced thermoelectric materials for power generation.[1–3] C_{60} semiconductors possess one of the lowest known thermal conductivities of all materials,[4] and yet the electronic properties can be tuned by appropriate doping.[5] The super-low thermal conductivity of C_{60} semiconductors is mainly attributed to its low Debye temperature (70 K). The formation of C_{60} is due to the weak *van der Waals* interaction. All the atomic bonds in C_{60} fullerenes are fully saturated and exist in all three dimensions. This novel feature of C_{60} fullerenes leads to a low Debye temperature in solids. The unique properties of C_{60} fullerenes provide opportunities to develop novel thermoelectric composite materials. We have explored the methods in the preparation of doped C_{60} bulk samples and investigated their thermoelectric properties both experimentally and via *ab initio* density functional theory (DFT) calculations.[6] In this paper, we report a greatly enhanced Seebeck coefficient in phosphorus (P) doped C_{60}, yet its thermal conductivity remains extremely low. We also discuss the results obtained on C_{60} samples doped with Co, Al, and Bi.

II. SAMPLE PREPARATION

Bulk C_{60} samples doped with P, Co, Al, and Bi were prepared by mixing powders of P, Co, Al, Bi, and C_{60} at appropriate atomic ratios and pressing them into pellets. The pellets were sealed in quartz tubes and heated just below 600 °C for several days. X-ray diffraction patterns of the

a)Electronic mail: jtang2@uwyo.edu.

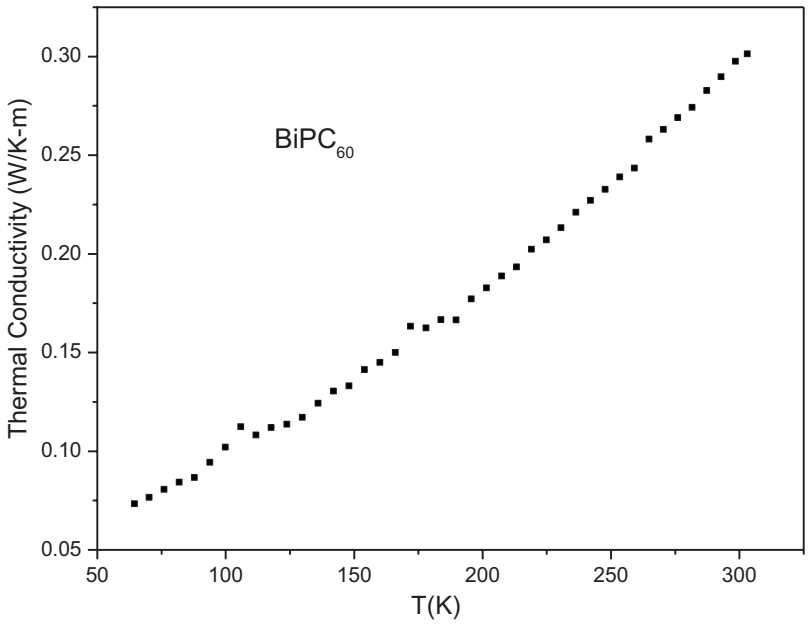


FIG. 1. Thermal conductivity of Bi and P co-doped C_{60}, $BiPC_{60}$.

samples show single phase C_{60} in face-centered cubic (fcc) structure. The thermoelectric properties, including Seebeck coefficient, thermal conductivity, and electrical resistivity of the samples, were measured with a Quantum Design PPMS system from low temperature up to 400 K.

III. EXPERIMENTAL RESULTS ON THE THERMAL TRANSPORT PROPERTIES OF Bi, Al, Co, AND BiP CO-DOPED C_{60}

Figure 1 shows the thermal conductivity of bismuth and phosphorus co-doped C_{60} bulk semiconductor samples ($BiPC_{60}$). The thermal conductivity of the bulk samples was less than 0.075 W/Km at 70 K and about 0.3 W/Km at 300 K. Such a room temperature thermal conductivity is an order of magnitude lower than those of the typical thermoelectric materials in current commercial applications and is similar to that of insulating bricks. Measurements on Co, Al, Bi, and P doped C_{60} semiconducting samples all showed extremely low thermal conductivity even when their electrical resistivity can be significantly reduced from that of the undoped and nearly insulating C_{60}. Figure 2 shows the resistivity ρ of Bi and P co-doped $BiPC_{60}$ sample over the temperature range from 50 to 300 K on a logarithmic scale. The resistivity decreases with increasing temperature and shows a tendency of decreasing further above the room temperature. At 300 K, the resistivity is only 0.06 Ω m, which demonstrates that the resistivity of the doped C_{60} semiconductors is highly tunable.

The Seebeck coefficient S varies with the selection of dopants (see Fig. 3). At room temperature, S is about 0.6, 14, 14, and 22 μV/K for Bi/P (co-doping), Al, Co, and Bi doping, respectively. The atomic ratio of the dopant to carbon atoms of C_{60} is fixed at 1:60 for the samples shown. Bi doped C_{60} appears to exhibit higher Seebeck coefficient than other cases. S is positive at room temperature and decreases with decreasing temperature. At 100 K, the Seebeck coefficient for $BiPC_{60}$ becomes negative as shown in Fig. 4, indicating a dominant role played by electrons at low temperatures for this sample.

IV. THERMAL TRANSPORT PROPERTIES OF P DOPED C_{60}

For P doped C_{60} samples, very interesting behaviors were observed. Figure 5 shows the thermal conductivity, Seebeck coefficient, electrical resistivity, and the figure of merit of a P doped sample (PC_{60}). Data are shown over the temperature range between 300 and 400 K. Below 300 K,

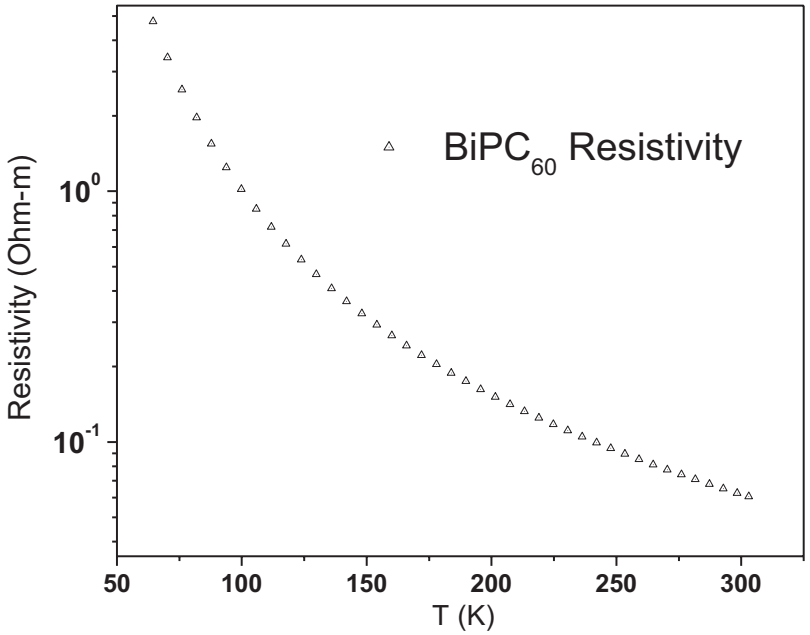


FIG. 2. Electrical resistivity of $BiPC_{60}$.

the resistance of the sample is too large to obtain reliable data on it and on the Seebeck coefficient. A very large Seebeck coefficient was observed. S reaches 1740 μV/K at 390 K and shows a tendency to increase further with increasing temperature. In order to confirm the observation, we have separately prepared a second sample. Similar increment in the Seebeck coefficient over the same temperature range was observed although the magnitude of the increase was not as large (S $\sim$70 μV/K). The electrical resistivity of this second sample was lower than the first. The differ-

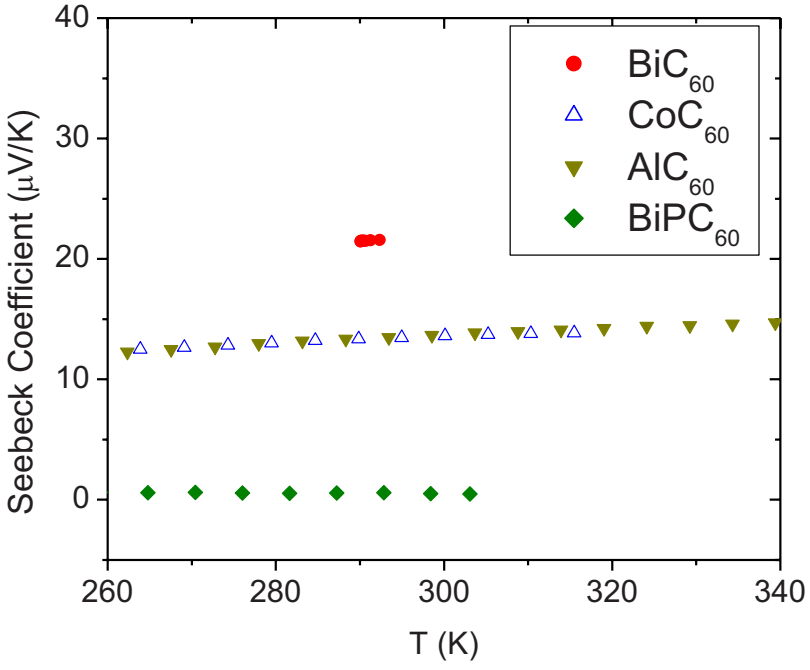


FIG. 3. Seebeck coefficient of C_{60} with various dopants.

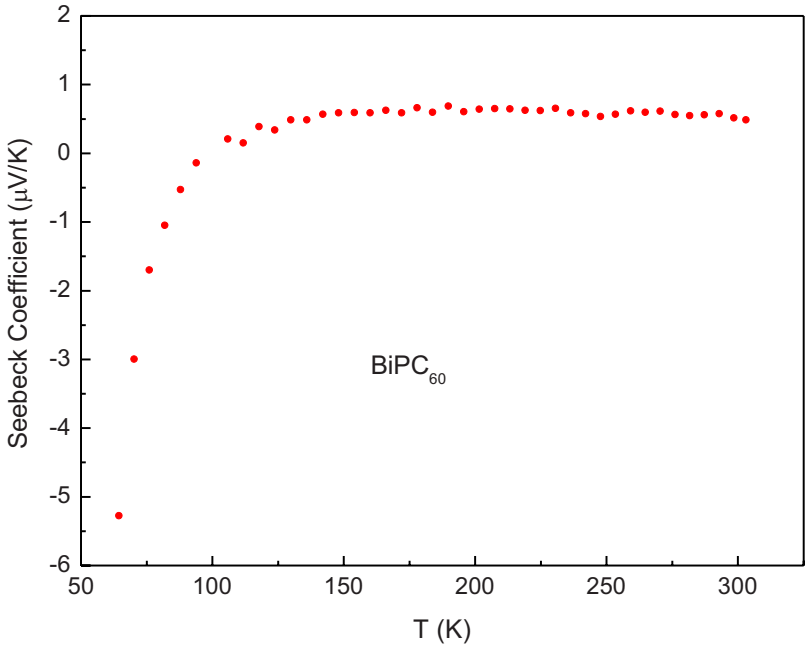


FIG. 4. Seebeck coefficient of $BiPC_{60}$ that shows a change from positive to negative values as the temperature decreases.

ences between the two samples are most likely due to the difference in the actual P concentration in the samples. Such a drastically increased Seebeck coefficient should have great implications on the potential application of this material. The thermal conductivity of the PC_{60} samples is about 0.06 W/Km at 50 K and about 0.1 W/Km at 300 K. Because the electrical resistivity of the P doped samples is quite high ($\rho = 10^1 - 10^3$ Ω m at room temperature), its ZT is rather small (10^{-5}). The greatest potential impact of the large Seebeck coefficient found in P doped C_{60}, we believe, lies in using it in C_{60} containing nanocomposites.

The concentration of phosphorous in C_{60} appears to be unstable at moderately high temperature. It is observed that the Seebeck coefficient changes with time when the sample is kept at about 100 °C in vacuum. It gradually decreases and finally becomes negative. The negative value of the Seebeck coefficient is consistent with that reported for undoped C_{60} films,[5] suggesting that the P is being pushed out of the C_{60} lattice under such conditions. However, any effect due to the oxygen absorbed in C_{60}, if any, cannot be ruled out.[5,7]

When the doping concentration is increased, sample P_4C_{60} (nominal composition) shows a much reduced Seebeck coefficient of 3 μV/K at 300 K (data not shown). The Seebeck coefficient remains positive down to 50 K for the P_4C_{60} sample, although it decreases with decreasing temperature.

A property important to application is that all doped C_{60} semiconductor samples studied here are chemically stable in air at room temperature, unlike alkali-metal doped C_{60}.

V. ELECTRONIC STRUCTURE OF DOPED C_{60}

In order to facilitate the search of high performance thermoelectric materials in experiments, we performed first-principles density functional theory calculations for the electronic structure of doped C_{60}. In an earlier report,[6] we presented the results for the electronic structure of C_{60} semiconductors under controlled doping with B, N, and Co atoms. We found that boron and cobalt doped, face-centered cubic C_{60} have the electronic structures of *n*-type semiconductors. Nitrogen doped fcc C_{60} solid has an electronic structure similar to those of a *p*-type semiconductor, with shallow impurity energy levels near the top of the valence bands of the host material.

We further calculated the electronic structure of Bi and P doped C_{60} that we studied experi-

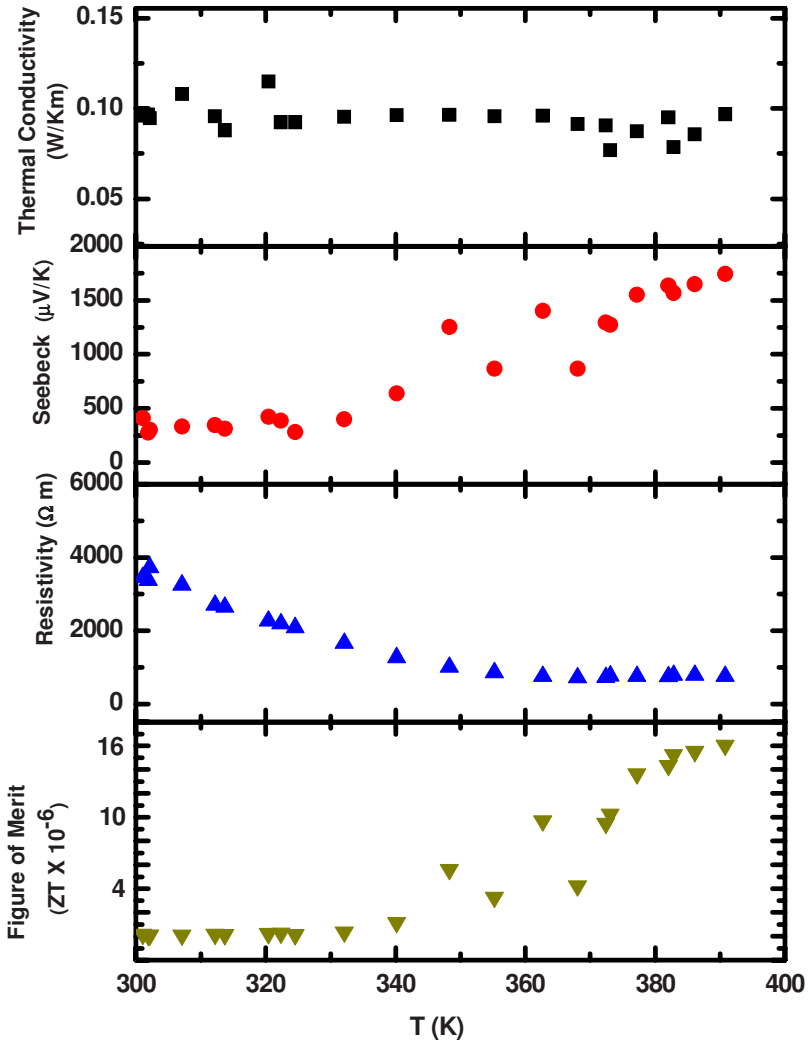


FIG. 5. Thermal conductivity, Seebeck coefficient, electrical resistivity, and figure of merit of a P doped sample (PC_{60}).

mentally. We used Vienna *ab initio* simulation package (VASP)[8] to perform the calculations. Our first-principle density functional calculations used the projector augmented wave (PAW) method and included the relativistic effects in the calculations.[9] We used the exchange-correlation potential in the generalized gradient approximation (GGA) and compared with those of the local density approximation (LDA) calculations. The 1*s*, 2*s*, and 2*p* electron states of P atom were treated as the core states as those of free atom in a frozen core approximation. The 3*s* and 3*p* electron states of P atom were included as the valence states. For Bi atom, 5*d*, 6*s*, and 6*p* electron states were included as the valence states and other deeper energy states were treated as the core states. We used a super-cell approach that included 60 carbon atoms and 1 doping atom (1:60 doping concentration) as well as 240 carbon atoms and 1 doping atom (1:240 doping concentration) in the comparative calculations. We implemented spin polarized electron density calculations. All the atomic coordinates and unit-cell volumes were relaxed in the *ab initio* DFT calculations. With the plane-wave energy cutoff at 450 eV, the calculated total energies converged to the order of about 0.01 meV. The residue forces on atoms were set at a value of 10 meV/Å. In the super-cell method, we used a $4\times4\times4$ and $1\times1\times1$ Monkhost grids in the *k* space sampling for the 1:60 and 1:240 doping concentrations, respectively. The Bader charges[10] were calculated for both the dopant atoms and the host C_{60} atoms.

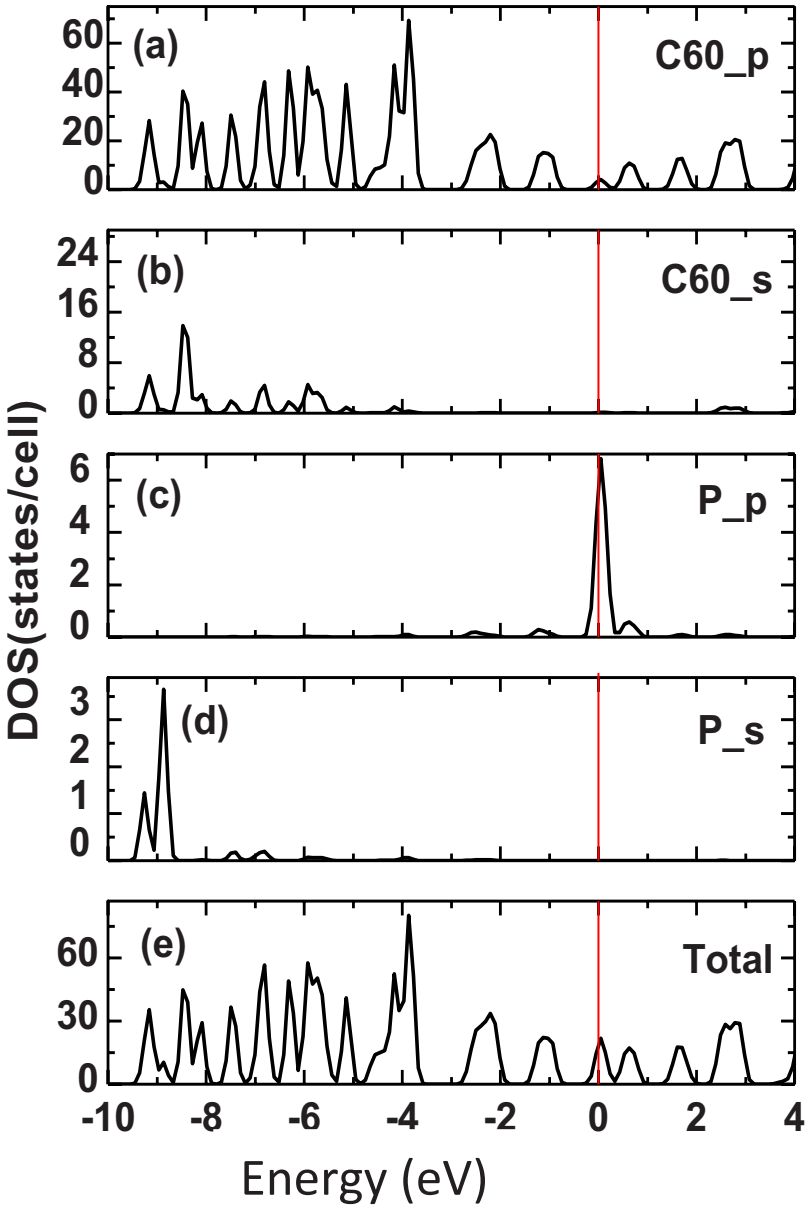


FIG. 6. The partial and total density of states of P doped C_{60}(1:60). The Fermi level is at 0.0 eV.

The partial and total electron density of states (DOS) of P and Bi doped C_{60} semiconductors were presented in Figs. 6 and 7, respectively. In all of the doped C_{60} semiconductors considered, the total energy is lower for the dopant atoms at the tetrahedral site than for other sites such as the octahedral sites. Consequently, in this article, we report the results of the dopant Bi and P atoms at the tetrahedral sites of C_{60} host material. We carefully tested both GGA and LDA calculations and found that the results are consistent, so we report the GGA results in this article, unless explicitly stated. One of the significant issues in the semiconductor research is the selection of suitable shallow impurity energy levels and activation energies for a desired working temperature. The activation energy also depends on the doping concentration, since the doping impurity energy level is no longer discrete in high doping concentration region but forms into an impurity band. Following the general theory in semiconductor physics, the carrier concentrations of an *n*-type semiconductor can be described by the following equation,[11]

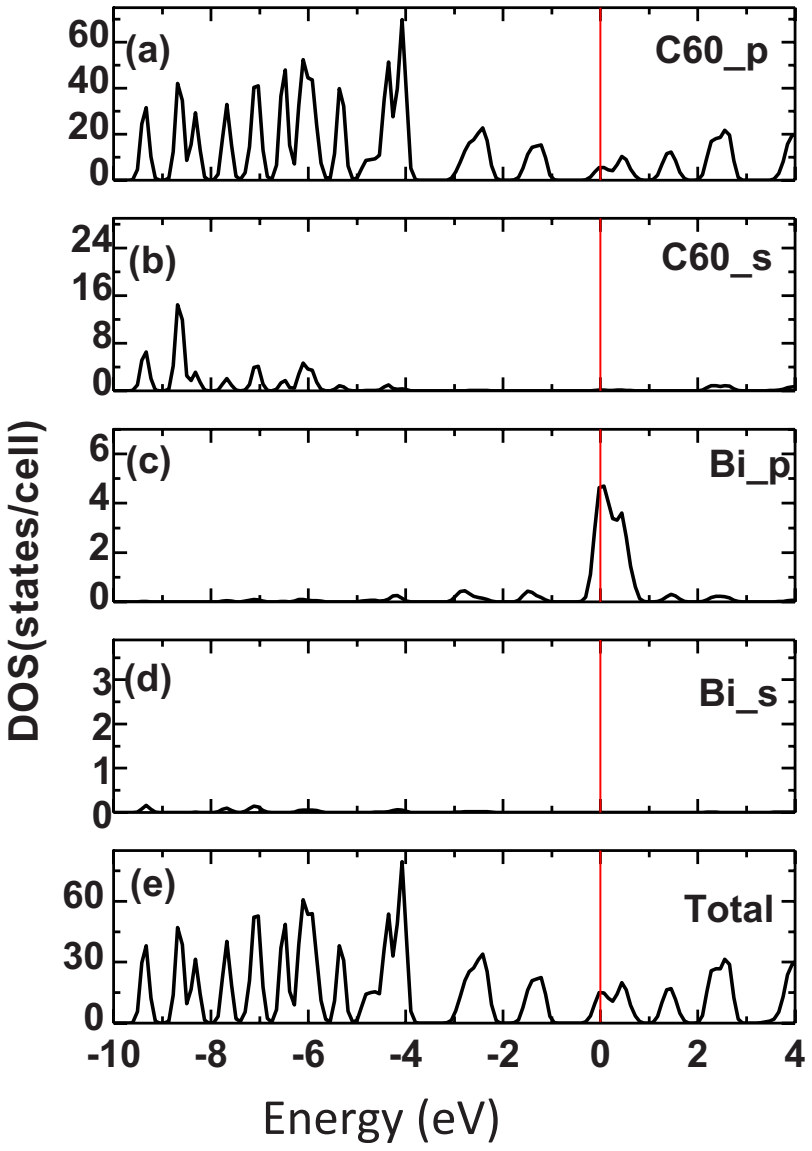


FIG. 7. The partial and total density of states of Bi-doped C_{60}(1:240). The Fermi level is at 0.0 eV.

$$\frac{n^2}{N_d - n} = \frac{N_c}{2} \exp\left(-\frac{\varepsilon_c - \varepsilon_d}{kT}\right),$$

where n is the n-type carrier concentration; N_d and ε_d are the density of donors and the donor impurity energy level, respectively; N_c and ε_c are the effective density of conduction-band states and conduction-band edge energy, respectively; $\varepsilon_c - \varepsilon_d$ is the activation energy; and kT is the thermal energy. From the calculated results of the electronic structure of the doped C_{60} semiconductors, we found that the phosphorus p-states form the shallow impurity electron states that are located below the conduction band edge of host C_{60} solid around the Fermi level (see Fig. 6). The impurity electron states in Bi doped C_{60} semiconductor (Fig. 7) merged with the conduction band edge of the host C_{60} solid. Consequently, the carrier activation energy in Bi doped C_{60} solid are very low (nearly zero) and would not have a large thermoelectric effect. Phosphorus doped FCC C_{60} semiconductor may have the carrier activation energies that are comparable to the room temperature thermal energy and may present a high thermoelectric effect around room tempera-

TABLE I. The calculated results of P and Bi doped (1:60 concentration) C_{60} semiconductors using DFT-GGA VASP program. The positive and negative signs of ΔQ mean losing and gaining electrons in $|e|$. The positive and negative signs of $\Delta V/V$ denote expansion and contraction of unit-cell volume, respectively. M_B is the magnetic moment of the system in μ_B.

Dopant	Dopant site	Type	$\Delta V/V$	$M_B(\mu_B)$	ΔQ (e)
P	Tetra	*n*	+0.27%	0.0	+0.033
Bi	Tetra	*n*	+2.19%	0.0	+0.30

ture, as demonstrated in our experiments. We also used a second computation package (a LCAO package adapted from the Ames Laboratory of Department of Energy) to perform the electronic structure calculations and confirmed the results from the VASP package. Some of the calculated results of P and Bi doped C_{60} semiconductors with 1:60 concentrations are also summarized in Table I. The calculated electronic structure of P and Bi doped C_{60} solids is similar to *n*-type semiconductors, consistent with the observed negative Seebeck coefficient at low temperatures in $BiPC_{60}$. Unfortunately, low temperature data on the Seebeck coefficient of PC_{60} are unreliable and were not taken for BiC_{60} to make a comparison. There is a small expansion of +0.27% and a relatively large expansion +2.19% of the unit cell volumes in the P and Bi doped FCC C_{60} solids, respectively. The calculated charge transfers for these two cases indicate that P and Bi atoms lose some electrons to C_{60} at about +0.033 and +0.30, respectively. There is no net magnetic moment in both P and Bi doped C_{60} solids.

VI. CONCLUSIONS

By carefully tuning the dopants, concentrations, and sample processing parameters of doped C_{60}, we can efficiently improve the thermoelectric performance of this new type of thermoelectric material. The thermal conductivity of the doped C_{60} is extremely small, about 0.1–0.3 W/Km at room temperature. For P doped C_{60}, we have observed a very large Seebeck coefficient in the order of 10^3 μV/K.

ACKNOWLEDGMENTS

This work is funded in part by the National Science Foundation Award No. CBET-0754821 and UW/SER MGF grant. Q.L. was supported by the U. S. Dept. of Energy, Office of Basic Energy Science, under Contract No. DE-AC-02-98CH10886.

[1] D. M. Rowe and C. M. Bhandari, *Modern Thermoelectrics* (Holt, Renehart and Winston, Eastbourne, 1983).
[2] H. J. Goldsmid, *Electronic Refrigeration* (Pion Limited, London, 1986).
[3] D. M. Rowe, *Thermoelectrics Hankbook—Macro to Nano* (CRC-Taylor & Francis, Boca Raton, FL, 2006).
[4] R. C. Yu, R. Heid, C. Meingast, P. Adelmann, D. Lorens, and R. Malhotra, Phys. Rev. Lett. **68**, 2050 (1992).
[5] N. Hayashi, K. Kanai, Y. Ouchi, and K. Seki, Proceedings Materials Research Society, Vol. 965E, Fall 2006, 0965 S13 03.
[6] G. L. Zhao, S. Yang, D. Bagayoko, J. Tang, and Z. Wang, Diamond Relat. Mater. **17**, 749 (2008).
[7] H. Tada, H. Touga, M. Takada, and K. Matsushige, Appl. Phys. Lett. **76**, 873 (2000).
[8] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993); G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996); G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996); VASP 2003 manual, see website: http://cms.mpi.univie.ac.at/vasp/.
[9] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999); P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).
[10] R. F. W. Bader, *Atoms in Molecules—A Quantum Theory* (Oxford University Press, Oxford, 1990).
[11] S. Wang, *Fundamentals of Semiconductor Theory and Device Physics* (Prentice Hall, Englewood Cliffs, NJ, 1989), Chap. 6.