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

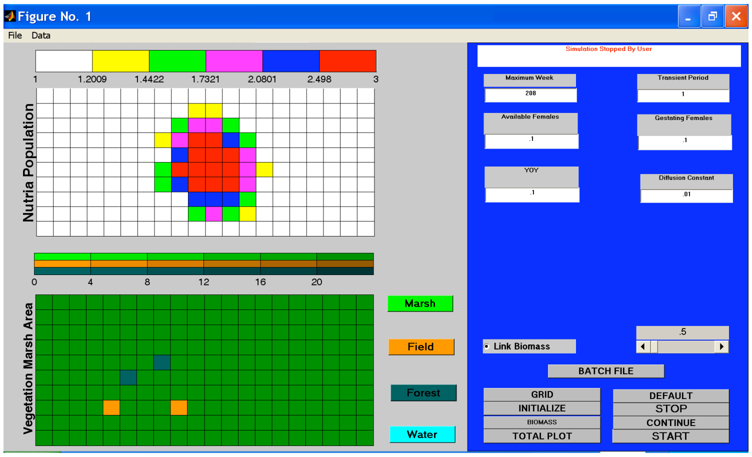
Project 2. Spatial Modeling of the Dynamics of Invasive Nutria

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.



Simulation Stopped By User

Maximum Week: 208
Transient Period: 1
Available Females: .1
Gestating Females: .1
YOY: .1
Diffusion Constant: .01
● Link Biomass: .5

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.

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

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.

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.

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.