Redeveloping and Optimizing the Interactive Chromatin Modeling [ICM] Web Server

Inderbir Sondh[1] Gyanadeep Kankawala[2] Tom Bishop[3]

1University of Pittsburgh Swanson School of Engineering, Department of Bioengineering

2Louisiana Tech University College of Engineering and Science, Department of Electrical Engineering

3Louisiana Tech University College of Engineering and Science, Departments of Physics and Chemistry

LA-SiGMA Louisiana Alliance for Simulation-Guided Materials Applications

Abstract

The Interactive Chromatin Modeling Web Server [ICM] is a tool for modeling DNA and chromatin interactively. The outputs include three-dimensional models of both free and folded forms of DNA as well as a nucleosome energy level diagram. The ICM kernel includes routines that assign Known DNA parameters to each base step of the sequence. That parameter (.par) file is Subsequently translated into a .XYZ file that can be rendered graphically by molecular modeling software such as VMD. The algorithm used to convert parameter files is that presented by El Hassan(1). The aim of the project is to redevelop the current FORTRAN ICM code using C++ with the expectation that the new program will efficiently process DNA sequences up to millions of base pairs long and introduce increased functionality as compared with the FORTRAN code.

Background: Parameter and XYZ Files + DNA

A DNA strand is made up of a long chain of four different bases: Adenine, Thymine, Cytosine, and Guanine (A, T, C, and G). Familiar B-form DNA exists in a double helix configuration, each base being paired with its corresponding partner (A to G and C to T). Moving along a strand of DNA will give its sequence (e.g. GATCCG). Each step such as G-A, A-T, and so on (referring to the example sequence just mentioned) constitutes a "base-step".
Each of the 16 possible base-steps has a unique set of six helical parameters. Shift, Slide, and Rise dictate movement (Angstroms) in the X, Y, and Z directions, respectively, while Tilt, Roll, and Twist indicate rotation (degrees) about the respective axes. When a DNA sequence is entered into ICM, a parameter file is the first file produced.

6197 base_pairs
*** local base-pair & step parameters ***

	Shear	Stretch	Stagger	Buckle	Prop-Tw	Opening	Shift	Slide	Rise	Tilt	Roll	Twist
G-C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-T	-0.06	-0.02	-0.03	0.13	-6.91	0.42	-0.05	0.22	3.23	-0.30	3.72	32.99
T-A	-0.06	-0.02	-0.03	0.13	-6.91	0.43	-0.00	-0.08	3.12	0.00	2.01	30.18
C-G	0.03	-0.02	-0.02	-1.43	-7.78	0.20	0.05	0.22	3.23	0.30	3.71	32.99
C-G	0.03	-0.02	-0.02	-1.44	-7.79	0.20	0.15	-0.28	3.34	0.15	5.68	29.57
G-C	0.03	-0.02	-0.02	-1.42	-7.77	0.20	0.00	0.30	3.07	0.00	8.07	27.24
T-A	-0.06	-0.02	-0.03	0.13	-6.92	0.43	0.05	0.04	3.19	-0.27	2.13	32.00
C-G	0.03	-0.02	-0.02	-1.43	-7.78	0.20	0.05	0.22	3.23	0.30	3.71	32.99
G-C	0.03	-0.02	-0.02	-1.42	-7.77	0.20	0.00	0.30	3.07	0.00	8.07	27.24

Part of an example parameter file. Only the last six parameters are used in our calculations. Note that Twist and rise are always much larger in magnitude than the other parameters, giving DNA its signature double helix shape.

The .par file is converted to an XYZ (.xyz) file using El Hassan's algorithm. A file in .xyz form can be directly fed into VMD, Jmol or similar software to produce a three dimensional rendering. The graphics can be modified to provide clarity and ease of viewing.

COMMENT	TcB par2xyz		
CA	0.00000	0.00000	0.00000
H1	1.00000	0.00000	0.00000
H2	0.00000	1.00000	0.00000
H3	0.00000	0.00000	1.00000
CA	-0.01227	0.23463	3.22933
H1	0.82456	0.77837	3.16564
H2	-0.55635	1.07355	3.24274
H3	0.04846	0.25807	4.22721

In our model, DNA is shown as a yellow bead, with each bead representing 5 base pairs.
To accurately study flexibility of DNA, one must consider the helical parameters as well as the effects of temperature. Our C++ program takes inputted temperature (K) and constructs a Gaussian distribution (centered around the standard parameter value) for each parameter, from which a random number is generated. The standard deviation of the distribution is specified in the original ICM paper by Bishop(2).

Method: Application of the El Hassan Algorithm

The process to convert a .par file to a .xyz file involves rotating and translating the coordinate axes of a base pair by adding and multiplying its coordinate rotation matrices (denoted by R_x,R_y, and R_z) after substituting parameter values for θ.

$$R_x(\theta) = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & \cos(\theta) & -\sin(\theta) \\ 0.00 & \sin(\theta) & \cos(\theta) \end{pmatrix}$$

$$R_y(\theta) = \begin{pmatrix} \cos(\theta) & 0.00 & \sin(\theta) \\ 0.00 & 1.00 & 0.00 \\ -\sin(\theta) & 0.00 & \cos(\theta) \end{pmatrix}$$

$$R_z(\theta) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0.00 \\ \sin(\theta) & \cos(\theta) & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$$

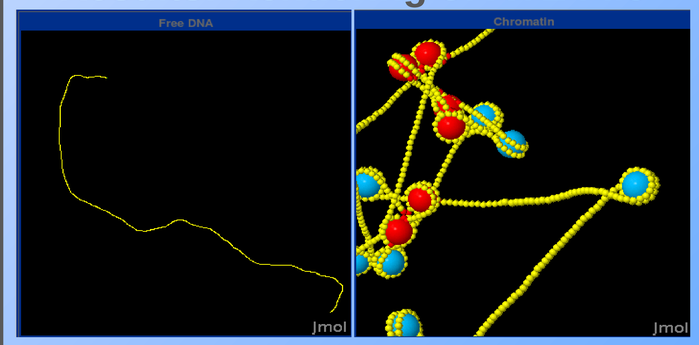
Left: The beginning rotation matrices are defined using standard direction sines and cosines.
Bottom: (10) gives a 3x3 matrix containing the mid-step triad located between the two base pairs. The columns of this matrix are multiplied by Shift, Slide, and Rise (Dx, Dy, and Dz respectively) in (11) to obtain the coordinates of CA. Adding the values of CA to the corresponding columns of T_{i+1} from (9) gives H1, H2, and H3.

$$\mathbf{T}_{i+1} = \left[\mathbf{R}_z\left(\frac{\Omega}{2} - \phi\right) \mathbf{R}_y(\Gamma) \mathbf{R}_z\left(\frac{\Omega}{2} + \phi\right) \right] \mathbf{T}_i \qquad (9)$$

$$\mathbf{T}_{mst} = \left[\mathbf{R}_z\left(\frac{\Omega}{2} - \phi\right) \mathbf{R}_y\left(\frac{\Gamma}{2}\right) \mathbf{R}_z(\phi) \right] \mathbf{T}_i \qquad (10)$$

$$\mathbf{r}^o_{i+1} = \mathbf{r}^o_i + D_x \mathbf{x}_{mst} + D_y \mathbf{y}_{mst} + D_z \mathbf{z}_{mst} \qquad (11)$$

Results: Rendering of XYZ File



Left: Jmol was used to produce these graphic models. This specific sequence was ~6200 base pairs long. ICM must be efficient enough to process sequences millions of base pairs long, interactively.

Right: Histones are shown in blue. A red histone indicates close contact and thus steric hindrance and/or an unstable energy conformation.

Results: Running Time Data

Using UNIX's "time" feature, we collected running times of our C++ parameter-to-xyz conversion when processing sequences up to 1,000,000 base pairs long.

Running Time Comparison of New and Old Code				
	C++		FORTRAN	
# of Base Pairs	Running Time (s)	CPU Usage (%)	Running Time (s)	CPU Usage (%)
1000000	68.3	74.5	25.69	100
100000	7.3	74.8	2.54	99.6
10000	0.83	66.2	0.25	100
1000	0.09	55.5	0.03	66.6
100	0.02	50	0.003	0
10	0	0	0	0

Discussion: Work In Progress

Currently there are some file I/O inefficiencies that cause an unnecessarily long running time, longer than the current ICM code in FORTRAN. This is not acceptable, as the goal is to make a more efficient program. Fixing these inefficiencies along with other program organization issues is predicted to reduce the running time to be at least as fast as the FORTRAN code. Also, the current C++ ICM code only supports 'Free DNA' modeling, the chromatin and nucleosome code is still under development and testing.

Discussion/Conclusion

There is strong evidence that once the C++ code is better organized and I/O problems are fixed, the program running time will be cut-down significantly. One current plan is to introduce a new data structure that encapsulates all the data normally in a .par file. This would involve a multidimensional array containing base-pairs along with their twelve corresponding helical parameters. Having such a structure will prevent having to access the .par file so many times. It was also discovered that the running time of the program is linked linearly with the size of the sequence file. e.g. a 10000 base pair file takes 3 seconds, a 100000 one would take 30 seconds. Further code analysis will be done to pursue a more favorable mathematical correlation.

Acknowledgments

This material is based upon work supported by the National Science Foundation under the NSF EPSCoR Cooperative Agreement No. EPS-1003897 with additional support from the Louisiana Board of Regents.

LOUISIANA TECH UNIVERSITY

References

1) El Hassan, M.A. and Calladine C.R., 1995, The Assessment of the Geometry of Dinucleotide Steps in Double-Helical DNA; a New Local Calculation Scheme, *J. Mol. Biol.*, Vol. 251, p. 648-664.
2) Bishop, T.C. and Stolz, R.C., 2010, ICM Web: the interactive chromatin modeling web server, *Nucleic Acids Research*, Vol. 38, Web Server Issue. DOI: 10.1093/nar/gkq496