Case Study: Computational Modelling in Structural Biology

Peter Minary
Department of Computer Science
Oxford University
Wolfson Building, Room 367
Oxford, OX1 3QD
Email: peter.minary@cs.ox.ac.uk

1 Background & Overview

1.1 Computational Modelling in Structural Biology

The rapidly increasing availability of experimental data for key macromolecules provides broad opportunities for the computational modelling and functional interpretation of their structures and one might envision routine and cheap *in silico* experiments replacing or aiding the interpretation of costly laboratory work. To achieve this goal, we might expect simulations to reveal conformational preferences that could explain a given functional working model of a nanomolecular machine or the mechanism how molecular switches trigger a biological process.

1.2 Conformational Sampling: Objectives and Limitations

The typical objective of conformational sampling is to obtain the expectation value and distribution of system specific observables that could range from simple distances such as the end to end distance in a protein to the 'volume' of a cavity that may accommodate a small molecule inhibitor. The expectation values are obtained by averaging over a phase space (Ω) in which the probability of visited states are described by a density distribution function, $f : \Omega \to \mathbb{R}$ that maps elements from phase space Ω to real numbers $\mathbb{R}$. Then, for any observable, $\alpha : \Omega \to \mathbb{R}$, the expectation value is defined as

$$\langle \alpha \rangle = \int_{\Omega} d\mathbf{L} \alpha(\mathbf{L}) f(\mathbf{L})$$

where $\mathbf{L} \in \Omega$. In the case of the canonical ensemble, the probability distribution function f is given by the Boltzmann distribution, $f(\mathbf{L}) = \exp(-\beta E(\mathbf{L}))/Q$, where the function $E : \Omega \to \mathbb{R}$ represents the energy of the system, $\beta = 1/kT$ (k is the Boltzmann constant and T the temperature) and Q is the partition function defined as $Q = \int_{\Omega} d\mathbf{L} \exp(-\beta E(\mathbf{L}))$. Therefore,

$$\langle \alpha \rangle = 1/Q \int_{\Omega} d\mathbf{L} \alpha(\mathbf{L}) \exp(-\beta E(\mathbf{L}))$$

In order to obtain $\langle \alpha \rangle$, one may average over all states visited in a numerical simulation (e.g. MD, MC) and the computed values can be compared with analog experimental observations. While this hope has been put forth since the first atomic resolution calculation (Levitt and Lifson, 1969) and experimental refinement (Levitt, 1974) on an entire biomolecule, the desired impact has not been delivered. The failure to adequately meet these expectations can be attributed to inherent limitations of basic molecular modelling methods such as molecular dynamics (MD) or standard Monte Carlo (MC) techniques: Firstly, their limited ability to explore globally the conformational space due to the large number of degrees of freedom that have to be explicitly treated (referred to as the *dimensionality problem*), and secondly their inefficiency in escaping from low energy conformational basins, hence conformational transitions occur with low probability (referred to as the *energy surface problem*).

2 Theory & Methods

In this section we provide the basic mathematical framework for some stochastic sampling algorithms and introduce a new methodology to reduce dimensionality in structure based computations.

2.1 Advanced Markov Chain Monte Carlo Algorithms

Regular Monte Carlo methods are exposed to the *energy surface problem* and have limited applicability to deliver expectation values, which can be better modelled by multi-canonical Monte Carlo schemes (Geyer, 1991, Minary and Levitt, 2006). Here, we review the theory behind both conventional and these more advanced Markov Chain Monte Carlo Methods.

Let us assume that the conformational state of a system, $X = \mathbf{L} \in \Omega$, could take different values $X_1, X_2, ...X_n$ during a simulation. We say that the simulation is guided by a Markov process if the future state is entirely determined by the present state. Formally, $P(X_{n+1}|X_0 = \mathbf{L_0}, ..., X_n = \mathbf{L_n}) = P(X_{n+1}|X_n = \mathbf{L_n})$, where X refers to the random variable that is assigned to different conformational states during the simulation. Next, one can define a Markov Chain, $X^{(.)}$ as a sequence of states $\{X_0, X_1, X_2, ..., X_n\}$ generated by the above Markov process. In Markov Chain Monte Carlo, we draw samples from a distribution, $f(X) = \phi(X)/N$, which is only known up to a normalizing constant, N and the consecutive elements are generated by the following rule: $1.X \rightarrow X'$, $q(X, X') = q(X', X)$; $2.P_{acc}(X \rightarrow X') = min\{1, \phi(X)/\phi(X')\}$. Here, q refers to a symmetric (often multivariate Gaussian) distribution.

In a conformational (Markov Chain) Monte Carlo trajectory we usually generate a Markov Chain, $X^{(T)}$ that samples from a Boltzmann distribution, $f(X, T)$, where $X(= \mathbf{L})$ denotes a conformational state (see 1.2) and T is the temperature. At low temperatures or even at physiological conditions ($T = 300K$), $X^{(T)}$ may include samples from only one particular conformational state often close to the initial starting structure, $X_0^{(T)}$ of the trajectory. This phenomena is very common as the small rate of escaping from low energy conformations hinders the global exploration of the conformational space. A widely accepted solution to this problem is parallel tempering (Geyer, 1991), in which one executes $K + 1$ Markov Chains, $\{X^{(T_0)}, X^{(T_1)}, ..., X^{(T_K)}\}$ in parallel each sampling from a Botzmann distribution, $f_i(X, T_i)$, at temperature, T_i, $\{i = 0, ..., K\}$. At step n, neighboring Markov Chains, $X^{(T_i)}$ and $X^{(T_{i+1})}$ may exchange states with a probability $P_{acc} = min\{1, f_{i-1}(y)/f_{i-1}(z) \times f_i(z)/f_i(y)\}$, where $y = X_n^{(T_i)}$ and $z = X_n^{(T_{i+1})}$. In this way, lower order (temperature) chains can visit more diverse conformational states while each chain still samples according to the proper canonical probability of occurrence at its own temperature.

2.2 Hierarchical Natural Move Monte Carlo Methods

Cartesian variables often represent the finest degrees of freedom (DOFs), $\mathbf{L}_f$, that span the full phase space, Ω_f. However this choice may become impractical with increasing system size due to the large number of independent variables. To reduce *dimensionality* one could use a proper subspace, $\Omega_s \subset \Omega_f$ spanned by a smaller number of generalized degrees of freedom, $\mathbf{L}_s$. For example, $\mathbf{L}_s$ may represent dihedrals (torsional angles) of a chain of connected particles/joints (Minary and Levitt, 2008, Minary and Levitt, 2006, Stein et al., 2006). Another solution is to change the conformation along torsional angles independently in many local segments while breaking chains connecting them, then use a closure algorithm to restore continuity (Minary and Levitt, 2010). Within this latter scheme, often referred to as Natural Move Monte Carlo one may use degrees of freedom (DOFs) that can be inferred from experimental observations, e.g. 'natural' moves.

Hierarchical Natural Move Monte Carlo (HNM-MC) (Sim et al., 2012) allows adaptive decomposition of the system into arbitrary structural segments of hierarchically increasing complexity. In this scheme, embedded subspaces of increasing sizes are defined, from Ω_s to Ω_f such that $\Omega_s = \Omega_n \subset \Omega_{n-1} ... \subset \Omega_1 = \Omega_f$, with each subspace representing a desired level of the hierarchy. In this way the energy surface along highest order collective DOFs, $\mathbf{L_n}$ are smoothed by introducing a hierarchical set of finer lower order DOFs, $\mathbf{L_{n-1}}, ..., \mathbf{L_1}$. Figure 1 presents such hierarchical DOFs for a large fractal-like RNA.

3 Numerical Experiments

3.1 Energy Functions

Sampling algorithms are generally assigned to a task of producing a distribution or evaluating an expectation value in a system described by the independent variable(s), $\mathbf{L}$ and a corresponding energy function, $E(\mathbf{L})$. For example, $E(L) = \epsilon_0[L^2 - a^2]^2/a^4$ describes a simple one-dimensional quadratic double well potential, where ϵ_0 is the height of the barrier and a is the separation of the wells. A more complex *energy surface* may be mimicked by using a Fourier series on the interval $[0, I)$,

$$E(L) = \sum_{i=1}^{20} c(i) \sin(i2\pi L/I)$$

where the coefficients $c(i)$ could be randomly chosen. While many of the all-atom empirical energy functions (MacKerrel, 1998, Cornell, 1995) represent a similarly *rough* energy surface, adequately capturing the conformational heterogeneity of complex biomolecules also requires a large number of Cartesian degrees of freedom (DOFs). While explicit treatment of these DOFs would hinder the global exploration of conformational space, the use of local torsional or essential (sometimes also called natural) DOFs often provide a computationally tractable alternative.

3.2 Structure of the Exercises

The numerical exercises are grouped into **Work Packages** and are available via the **Tutorial** link at the MOSAICS website[1]. There are 4 **Work Packages**, (**WPs**) and students are recommended to complete them in increasing order. In a two days case study, students may complete **WP1-2** in the first day and and **WP3-4** may left for the second day. More advanced and interested students may consider research tutorials [2,3,4].

The covered areas are as follows:

1. **WP1** *Gaining familiarity with one of the popular algorithms that has superior performance over conventional Markov Chain Monte Carlo (Metropolis et al., 1953) (MCMC) in overcoming energy barriers and facilitates the global exploration of a state space (Ω). By comparing performance to the standard MCMC scheme one could demonstrate the benefits of using methods that are designed to tackle the energy surface problem.*

2. **WP2** *Learning the practical use of algorithms that provide solution to the dimensionality problem. This will be done through running molecular simulations on small biological structures so that benchmark calculations can be readily preformed during the exercise. The benefits of using these techniques will be demonstrated through performing analog simulations using conventional Cartesian degrees of freedom.*

3. **WP3** *Study the same systems used in 2 and demonstrate the combine effects of algorithmic schemes including the ones learned in 1 and 2. These studies will demonstrate that modern computational simulations harvest the synergetic effect of advanced algorithmic schemes when custom designing an algorithmic protocol for a particular application in hand.*

4. **WP4** *This WP presents a modelling task for realistic application: to design and optimize algorithms to explore the conformational space of one of the first RNA riboswitch structures. The exercise require all the knowledge and experience students gain during completing the previous Work Packages.*

[1] www.cs.ox.ac.uk/mosaics
[2] http://www.cs.ox.ac.uk/mosaics/protein.html
[3] http://www.cs.ox.ac.uk/mosaics/rna.html
[4] http://www.cs.ox.ac.uk/mosaics/dna.html

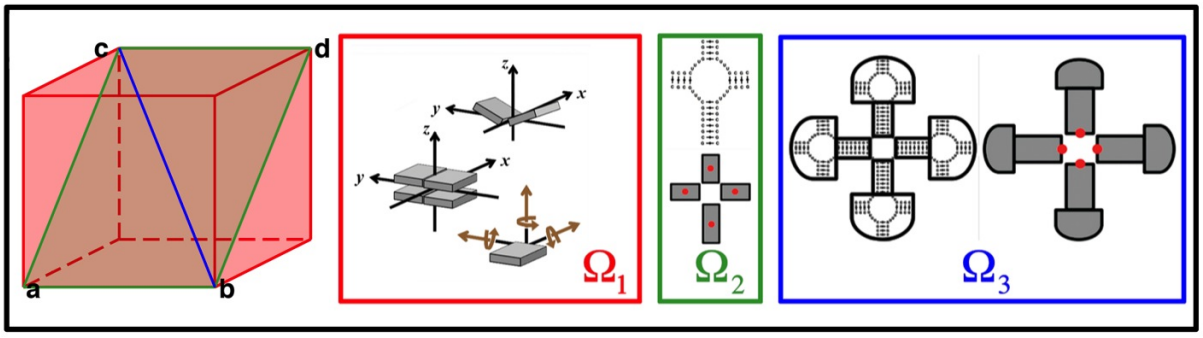

Figure 1: (A) Example of three embedded subspaces, $\Omega_1 \subset \Omega_2 \subset \Omega_3$ spanned by hierarchical degrees of freedom $\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3$ of a large fractal-like RNA modeled by Hierarchical Natural Move Monte Carlo (HNMMC). Ω_1, Ω_2 and Ω_3 spaces represent the arrangement of individual and groups of nucleotides in helices and four helix arms, respectively. The relative size of each subspace is illustrated by a geometric analogy: Ω_1 spans the volume of cube (red), Ω_2 spans the plane abcd (green) and Ω_3, spans the line cb (blue).

References

W. Cornell. 1995. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197.

C. J. Geyer. 1991. Computing science and statistics. In *Proc. 23rd Symp. Interface 156-163*, pages 156–163.

Michael Levitt and Shneior Lifson. 1969. Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.*, 46:269–279.

Michael Levitt. 1974. Energy refinement of hen egg-white lysozyme. *J. Mol. Biol.*, 82:393–420.

A. MacKerrel. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A.H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092.

Peter Minary and Michael Levitt. 2006. Discussion of the equi-energy sampler. *The Annals of Statistics*, 34:1636–1641.

Peter Minary and Michael Levitt. 2008. Probing protein fold space with a simplified model. *J. Mol. Biol.*, 375:920–933.

Peter Minary and Michael Levitt. 2010. Conformational optimization with natural degrees of freedom: A novel chain breakage/closure algorithm. *J. Comp. Biol.*, 17:993–1010.

Peter Minary. 2007. Methodologies for Optimization and SAmpling in Computational Studies (MOSAICS). Software release, Stanford University (2007) and Oxford University (2012), http://www.cs.ox.ac.uk/mosaics.

Adelene Y. Sim, Michael Levitt, and Peter Minary. 2012. Modeling and design by hierarchical natural moves. *Proc. Nat. Acad. Sci.*, 109(8):2890–2895.

Evan G. Stein, Luke M. Rics, and Axel T. Brunger. 2006. Torsion-angle molecular dynamics as a new efficient tool for nmr structure calculation. *Journal of Magnetic Resonance*, 34:1636–1641.