Energy Landscapes and Accelerated Molecular-Dynamical Techniques for the Study of Protein Folding

John K. Prentice
Wild Rose Biophysics
Boulder, CO

BioMed Seminar
University of New Mexico Physics and Astronomy Department
November 3, 2004
What is a Protein?

See http://www.accessexcellence.org/RC/IL/GG/prot_Struct.html

Wild Rose Biophysics
Peptide Units Combine to Form Proteins

Space Filling Model  Stick Model  Ribbon Model

See http://www.dl.ac.uk/SRS/PX/special/special.html
The Native Structure of Proteins is Unique and Complex

Bovine F1 ATP Synthase

See http://www.dl.ac.uk/SRS/PX/special/special.html
So How Do Proteins Fold?

Denatured state  ▸  Folded (native) state

Thanks to John Chodera, Dill Biophysics Group, UCSF
The Protein Folding Problem

The “holy grail”: Compute structure from sequence \textit{ab initio}

This involves exploring the conformational space of the protein using “first-principle physics” methods in order to identify the native structure.
A Realistic Energy Funnel Landscape

Denatured state

Native state

Folding Funnel

Wild Rose Biophysics
How to Theoretically Explore Folding? The Molecular Dynamics Approach

The AMBER forcefield:

\[ V(q) = \sum_{\text{bonds}} K_r (r-r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta-\theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi-\gamma)] + \sum_{i<j} \left[ \frac{A_{ij}}{R_{ij}^2} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \]

\[ H(q,p) = \frac{1}{2} p^T M^{-1} p + V(q) \]
Molecular Dynamics

Numerically solve Newton’s 2nd law over small time increments to compute the position and momentum of the atoms in the protein over time.

\[ m_i \frac{d\vec{v}_i}{dt} = \sum_{\substack{j=1 \atop j \neq i}}^{N} \vec{F}_{i,j} \]

\[ \dot{q} = \nabla_p \mathcal{H}; \quad \dot{p} = -\nabla_q \mathcal{H} \]
All-Atom MD Calculations of Folding

• These calculations include all atoms in the protein, implicit or explicit solvent, and account for “all” forces between the various atoms.

• Each MD timestep is no more than 2 fs \((10^{-15}\text{ s})\).

• A 4 \(\mu\text{s}\) calculation would require \(2 \times 10^9\) timesteps to simulate observed “real life” folding times.

• At 10 ms per timestep, such a simulation would require 231 single CPU days to compute a single trajectory on a modern high-end workstation.

• Brute force, single trajectory full-folding calculations are thus in general not practical at this time.
Statistical Molecular Dynamics

• Consider a protein consisting of $N$ atoms whose instantaneous positions and momenta at time $t$ are given by the 3N dimensional vectors $\vec{r}(t)$ and $\vec{p}(t)$.

• The time average of a property $A$ is defined as:

$$\langle A \rangle_t = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau dt \ A[\vec{r}(t), \vec{p}(t)]$$

• Numerically, we calculate this as:

$$\langle A \rangle_t \approx \frac{1}{M} \sum_{m=1}^M A[\vec{r}(t_m), \vec{p}(t_m)]$$
Statistical Molecular Dynamics (2)

Note that the ergodic hypothesis says that

\[
\langle A \rangle_t = \langle A \rangle_e
\]

where the ensemble average of property A is given by

\[
\langle A \rangle_e = \iiint d\vec{p} \ d\vec{r} \ A(\vec{r}, \vec{p}) \rho(\vec{r}, \vec{p})
\]

and the probability density in the canonical (PVT) ensemble is given by

\[
\rho(\vec{r}, \vec{p}) = \frac{\exp\left(-\frac{E(\vec{r}, \vec{p})}{k_B T}\right)}{Q}
\]
Folding Times are Often Dominated by the Time to Escape Kinetic Traps

- Low energy trajectories tend to get trapped in metastable local minima.
- Higher energy (temperature) trajectories avoid kinetic traps.
- By varying temperatures to kick a trajectory out of a kinetic trap, we can accelerate MD folding simulations.
Simulations are conducted in parallel at temperatures $T_1, T_2, \ldots, T_{\Lambda-1}, T_{\Lambda}$. Because each replica can undergo cycles of heating and cooling, conformation space is broadly sampled.

**The Replica Exchange Method**

$P = \min\{1, \exp(-\Delta)\}$

*Thanks to John Chodera, Dill Biophysics Group, UCSF*
Details of the REM (1)

Let $X = (x_1^{i(1)}, ..., x^{i(\Lambda)}) = (x_1^{[i]}, ..., x^{[\Lambda]})$ stand for a “state” in this generalized ensemble.

- The superscript and subscript label the replica and temperature respectively.
- Each state $X$ is specified by $\Lambda$ sets of coordinates and momenta of $N$ atoms in replica $i(\lambda)$ at temperature $T_\lambda$. We assume each replica is in the canonical (PVT) ensemble.

$$x_\lambda^{[i(\lambda)]} \equiv \left( \vec{r}^{[i(\lambda)]}, \vec{p}^{[i(\lambda)]} \right)_\lambda$$
Details of the REM (2)

The replicas are non-interacting. Therefore, the weight factor for the state $X$ in this generalized ensemble is given by the product of the Boltzmann factors for each replica (each at a single temperature):

$$ W_{REM} (X) = \exp \left\{ - \sum_{i=1}^{\Lambda} \beta_{\lambda(i)} H \left( \bar{r}^{[i]}, \bar{p}^{[i]} \right) \right\} $$

$$ = \exp \left\{ - \sum_{\lambda=1}^{\Lambda} \beta_{\lambda} H \left( \bar{r}^{[i(\lambda)]}, \bar{p}^{[i(\lambda)]} \right) \right\} $$

where

$$ \beta = \frac{1}{k_B T} $$
Details of the REM (3)

Now consider exchanging a pair of replicas in the generalized ensemble. Specifically, exchange the temperatures between replicas \( i \) and \( j \). Thus:

\[
X = \left( \ldots, x^{[i]}_{\lambda}, \ldots, x^{[j]}_{\gamma}, \ldots \right) \rightarrow X' = \left( \ldots, x^{[j]'}_{\lambda}, \ldots, x^{[i]'}_{\gamma}, \ldots \right)
\]

where

\[
\begin{align*}
    x^{[i]}_{\lambda} &\equiv (\vec{r}^{[i]}, \vec{p}^{[i]})_{\lambda} \rightarrow x^{[i]'}_{\lambda} &\equiv (\vec{r}^{[i]}, \vec{p}^{[i]'})_{\lambda} \\
    x^{[j]}_{\gamma} &\equiv (\vec{r}^{[j]}, \vec{p}^{[j]})_{\gamma} \rightarrow x^{[j]'}_{\gamma} &\equiv (\vec{r}^{[j]}, \vec{p}^{[j]'})_{\lambda}
\end{align*}
\]
Details of the REM (4)

The post-exchange momenta are defined by:

\[
\begin{aligned}
    \vec{p}'[i] & \equiv \sqrt{\frac{T_y}{T_\lambda}} \vec{p}[i] \\
    \vec{p}'[j] & \equiv \sqrt{\frac{T_\lambda}{T_y}} \vec{p}[j]
\end{aligned}
\]

This specific scaling of momenta correctly rescales the kinetic energy to be consistent with the equipartition of energy theorem

\[
\langle K(\vec{p}) \rangle_T = \left\langle \sum_{k=1}^{N} \frac{\vec{p}_k^2}{2m_k} \right\rangle_T = \frac{3}{2} Nk_B T
\]

where \( K \) is the kinetic energy.
Details of the REM (5)

In order for this exchange process to converge toward an equilibrium distribution in the canonical ensemble (for each temperature), it is sufficient to impose the detailed balance condition on the transition probability $w(X \rightarrow X')$:

$$W_{REM}(X)w(X \rightarrow X') = W_{REM}(X')w(X' \rightarrow X)$$

This can be satisfied by the usual Metropolis criterion:

$$w(X \rightarrow X') \equiv w(x_i^\gamma | x_j^\lambda) = \begin{cases} 
1 & \text{for } \Delta \leq 0 \\
\exp(-\Delta) & \text{otherwise}
\end{cases}$$

where $U$ is the potential energy and

$$\Delta = [\beta^\gamma - \beta^\lambda](U(\tilde{r}[i]) - U(\tilde{r}[j]))$$
Details of the REM (6)

The canonical expectation value of a quantity $A$ at temperature $T_\lambda (\lambda = 1, \ldots, \Lambda)$ can then be calculated by

$$\langle A \rangle_{T_\lambda} = \frac{1}{M} \sum_{m=1}^{M} A \left( x_{\lambda}^{[f(\lambda, m)]} (t_m) \right)$$

Expectations values for intermediate temperatures can be obtained using the multiple-histogram reweighting technique (WHAM).
Simulations are conducted in parallel at temperatures $T_1, T_2, \ldots, T_{\Lambda-1}, T_{\Lambda}$. Because each replica can undergo cycles of heating and cooling, conformation space is broadly sampled.

$P = \min\{1, \exp(-\Delta)\}$

Thanks to John Chodera, Dill Biophysics Group, UCSF
Advantages of REM

• Exchanging temperatures kicks replicas out of kinetic traps and accelerates the simulated folding process.

• Each replica calculation can be done on a different processor on a parallel computer. There is little interprocessor communication required, so the problem is virtually “embarrassingly parallel”.

• Thus, on a parallel computer with sufficient processors, there is no “wall clock” penalty for the additional work required in REM while the simulation time to reach the native state is reduced because the sampling of phase space is improved.
Results for Met-enkephallin

Time series of temperature exchange and total potential energy for one replica in an 8 replica REM calculation of Met-enkephalin (Tyr-Gly-Gly-Phe-Met).

Results for Met-enkephallin

Fig. 4. Distributions of a pair of dihedral angles ($\phi, \psi$) of Gly-2 for: (a) $T = 200$ K from a regular canonical MD simulation, (b) $T = 200$ K from the replica-exchange MD simulation, (c) $T = 700$ K from a regular canonical MD simulation, and (d) $T = 700$ K from the replica-exchange MD simulation.

MREM Method

• Multiplexed REM (MREM) is closely related to REM, but uses multiple replicas at the same temperature in addition to replicas at different temperatures. This improves sampling and further reduces folding times.

• Rhee and Pande performed MREM calculations for the 23 amino acid model protein BBA5.

• The calculations used 20 temperatures and 200 replicas per temperature.

• The MREM calculations reached the native state approximately 10 times faster than a constant temperature molecular dynamics calculation.
FIGURE 3  Stereo representations of (a) a folded conformation example and (b) the native structure. For simplicity, only Cα backbone and selected side chains (Val3, Phe8, Leu14, Leu17, and Leu18) are shown. The β-hairpin (residues 2-7) and the α-helix (residues 12-20) regions are represented in blue and red, respectively.
Results for BBA5 Using MREM

FIGURE 4  Snapshots of three selected folding trajectories obtained at every 10-ns simulation time. The same coloring scheme in Fig. 3 is adopted.
Results for BBA5 Using MREM

FIGURE 6  Evolutions of populations with native-like characters: (a) α-helix, (b) β-turn; (c) evolution of the average RMSD₆. Solid lines represent the results at 279 K of temperature obtained with MREMD method. Dotted lines represent the results from CTMD method at the same temperature.
Another Way to Accelerate Folding

Note that transition rates merely reflect the average time required for a transition. Some trajectories will transition faster than others and some slower. The parallel replica method exploits this as another method for accelerating simulations.
The Parallel Replica Method

• Consider $N$ identical conformation replicas at the same temperature, but with randomized velocities.

• The probability that a particular simulation has crossed the barrier is:

$$P_1(t) = k \exp(-kt)$$

• For $M$ simulations, the probability of the first stimulation crossing the barrier is

$$P_M(t) = MP_1(t) \left[ 1 - \int_{0}^{t} d\tau P_1(\tau) \right]^{M-1}$$

$$= Mk \exp(-Mkt)$$

_Pande, et. al., Biopolymers, vol 68, 2003._
Folding@Home

- Parallel replica requires infrequent exchange of information between replicas. It is therefore nearly embarrassingly parallel.

- Folding@Home distributes the computation of each replica to individual PCs across the World Wide Web.

- To date, over 1 million computers have participated in folding@home.

- This has made possible the calculation of realistic complete folds for moderate sized proteins out to the equivalent of milliseconds.

See http://folding.stanford.edu
Conclusions

• Folding rates are dominated by time spent in kinetic traps. In silico brute-force folding of proteins of biological interest via a single molecular dynamics trajectory is not in general feasible with current computers.

• The replica exchange and parallel replica techniques avoid this problem by artificially minimizing the time spent in kinetic traps (at the expense of destroying information about the kinetics).

• Speedups of 10x in calculation of native structures have been realized for MREM compared to constant temperature MD.

• Speedups of several orders of magnitude for calculation of native structures have been realized using parallel replica.
Acknowledgements

I would like to acknowledge the invaluable assistance in preparing this talk of John Chodera and Ken Dill of the Dill Biophysics Group, UCSF.