Advanced Topics in Molecular Dynamics: Sampling & Solvation

Vijay Pande
OpenMM Workshop, February 13, 2009
Challenge: timescales

-Kinetics: reaching experimental timescales
- Thermodynamics: convergence
  - Are the results independent of initial conditions?
New Application: OpenMM Zephyr

• Goals
  • make MD easy to run
  • easy but correct setup
    (not just PDB -> MD, but think about
    protonation, missing residues, etc)
  • easy to run on GPU’s
  • visual feedback

• Under the hood
  • Wrap GPU enabled MD code
  • use MMtools (Pande group, SimTk.org) or new Gromacs set up tools
  • Use VMD IMD interface for visualization (leverage a standard in molecular
    visualization)

• Use of real time visualization
  • immediate feedback is not just fun, but can be useful
  • key to correct setup, etc
Large speed increases seen using GPU

<table>
<thead>
<tr>
<th>Molecule</th>
<th># atoms</th>
<th>ns/day</th>
<th>speedup*</th>
<th>GFLOPS (GPU)</th>
<th>GFLOPS (x86)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fip35</td>
<td>544</td>
<td>576</td>
<td>128</td>
<td>311</td>
<td>657</td>
</tr>
<tr>
<td>villin</td>
<td>582</td>
<td>529</td>
<td>136</td>
<td>328</td>
<td>692</td>
</tr>
<tr>
<td>lambda</td>
<td>1254</td>
<td>202</td>
<td>255</td>
<td>547</td>
<td>1153</td>
</tr>
<tr>
<td>a-spectrin</td>
<td>5078</td>
<td>17</td>
<td>735</td>
<td>805</td>
<td>1702</td>
</tr>
</tbody>
</table>

(*comparing a GTX280 to a single core of a 3GHz core 2 duo using the AMBER code)
How accurate are atomistic physical models?

**ELECTROSTATICS**


**SOLVATION FREE ENERGY**


**THERMODYNAMICS**

Journal of Chemical Physics, 123 084108 (2005)

**KINETICS**

Annual Reviews of Biophysics 34 43-69 (2005)
Case study: implicit solvent

http://pande.stanford.edu
How does Generalized Born (GB) work?

- Break down water into dielectric and hydrophobicity

\[
\Delta G_{solv} = \Delta G_{el} + \Delta G_{nonel},
\]

- Make an ansatz for the form of the dielectric

\[
\Delta G_{el} \approx \Delta G_{gb} = -\frac{1}{2} \sum_{ij} \frac{q_i q_j}{f_{GB}(r_{ij}, R_i, R_j)}
\]

\[
f_{GB} = \left[ r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_i R_j) \right]^{\frac{1}{2}}
\]

- Must calculate the Born Radii (R_i)

\[
R_i^{-1} = \rho_i^{-1} - \frac{1}{4\pi} \int_{solute} \theta(|r| - \rho_i) \frac{1}{r^4} d^3r.
\]

- Can include salt effects

\[
\Delta G_{el} \approx \Delta G_{gb} = -\frac{1}{2} \sum_{ij} \frac{q_i q_j}{f_{GB}(r_{ij}, R_i, R_j)} \left( 1 - \frac{e^{-\kappa f_{gb}}}{\epsilon_w} \right)
\]
Can this be put on a more formal ground?

- Limiting case: single ion of radius $r$ yields the Born eq

$$\Delta G_{el} \approx -\frac{1}{2} \left(1 - \frac{1}{\varepsilon_w}\right) \frac{q^2}{r}$$

- For the linearized PB equation, one can derive the exact result in a spherical geometry (Kirkwood equation), which yields

$$\Delta G_{el} \approx \Delta G_{gb} = -\frac{1}{2} \sum_{ij} \frac{q_i q_j}{f_{GB}(r_{ij}, R_i, R_j)}$$

$$f_{Kirkwood} = \left[r_{ij}^2 + R_i R_j\right]^{1/2}$$

$$f_{GB} = \left[r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_i R_j)\right]^{1/2}$$

- The exp term can be considered an empirical fix for non-spherical geometries
Different Generalized Born models

• All have the same general form

\[ \Delta G_{el} \approx \Delta G_{gb} = -\frac{1}{2} \sum_{ij} \frac{q_i q_j}{f_{GB}(r_{ij}, R_i, R_j)} \]

\[ f_{GB} = \left[ r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_i R_j) \right]^{\frac{1}{2}} \]

• But differ in the calculation of the Born radii
  • Still (Original)
  • Hawkins, Cramer, Truhlar (“HCT”)
  • Onufriev, Bashford, Case (“OBC”)
  • Mongan, Simmerling, McCammon, Onufriev, Case (“GBneck”)

• Goal is to best model the nature of the dielectric region
Large number of force fields to choose

**AMBER**
- ff94: too helical (explicit solvent)
- ff96: too beta sheet (explicit solvent)
- ff99: not helical enough (explicit solvent)
- ff99sb: modifications to improve torsions
- ff03: latest, intended to be balanced

**OPLS**
- OPLS-ua (unified atom)
- OPLS-aa: classic all atom force field
- OPLS-aa/L: new torsions

**CHARMM**
- CHARMM19 (unified atom)
- CHARMM27 (latest)
- CMAP (new torsions for use with CHARMM27 or other CHARMM ff’s)

**Other**
- GROMOS (van G.)
- GROMACS
- Encad (Levitt)

**Polarizable force fields**
Test systems

protein G hairpin
16 residues

EK peptide
14 residues

multiple force fields
x multiple solvent models
x two test peptides
x three runs each
x 10 ns REMD runs
≈ 12 μs aggregate simulation time
≈ 60 CPU-years of compute time
Results for AMBER ff’s

<table>
<thead>
<tr>
<th>HCT</th>
<th>OBC</th>
<th>GBn</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="" alt="Images of structures for HCT, OBC, GBn with colors indicating hydrophobics, polar, positively charged, and negatively charged" /></td>
<td><img src="" alt="Images of structures for HCT, OBC, GBn with colors indicating hydrophobics, polar, positively charged, and negatively charged" /></td>
<td><img src="" alt="Images of structures for HCT, OBC, GBn with colors indicating hydrophobics, polar, positively charged, and negatively charged" /></td>
</tr>
</tbody>
</table>

color key: hydrophobics, polar, positively charged, negatively charged

(Scott Shell, UCSB; Ken Dill, UCSF)
Case study:
protein thermodynamics
Sampling methods

• **NVE MD: Constant energy**
  - often used to be most faithful to kinetics
  - important test of an MD code (no bugs or numerical issues)

• **NVT MD**
  - uses a thermostat
  - Vijay’s opinion (w/data): can be used for kinetics, if a thermostat is used carefully

• **Other thermodynamics methods**
  - **ST**: Serial Tempering
  - **REMD**: Replica Exchange Molecular Dynamics (aka parallel tempering)
  - **MSM**: Markov State Models
Next steps: Generalized Ensemble (GE) methods

- **Basic idea:** define new form of kinetics to overcome long timescale behavior
  - must be able to recover Boltzmann weighted configurations
  - but we can define what ever form of dynamics we want
  - and we can create new potential forms, as we can transform back
  - generalization of methods like “parallel tempering” or REMD

- **Game plan**
  - identify what are the factors limiting kinetics (high energy barriers? diffusion?)
  - pick states that drive against these factors
  - Define a new, reduced potential along these states
  - recover original Boltzmann weightings to calculated desired free energy

(this process will have similarities to the BAR step mentioned in the previous cases -- the question will be given a set of data, what’s the best prediction of free energies and Boltzmann weights)
The reduced potential

We define the reduced potential for a state \( k \) as a combination of terms

\[
   u_k(x) = \beta_k \left[ U_k(x) + p_k V(x) + \mu_k^T N(x) \right]
\]

with thermodynamic parameters for each state

- \( \beta_k \): inverse temperature
- \( U_k \): potential energy function
- \( p_k \): external pressure
- \( \mu_k \): chemical potential of exchangeable species

where

- \( x \): microstate or configuration
- \( V(x) \): volume of simulation box
- \( N(x) \): number of each chemical species in system

The distribution function is given by

\[
   p_k(x) = Z_k^{-1} \exp[-u_k(x)] \quad Z_k = \int dx \exp[-u_k(x)]
\]

Covers many common thermodynamic ensembles: NVT, NPT, \( \mu VT \), \( \mu PT \)

The method of expanded ensembles

Form an expanded ensemble by allowing transitions between thermodynamic states:

\[ p(x, k) = Z^{-1} \exp[-u_k(x) + g_k] \]

with partition function

\[ Z = \sum_{k=1}^{K} Z_k \exp[g_k] \]

where we have introduced log weights \( g_k \) to bias sampling of states.

Current configuration now consists of \((x, k)\) pair.

How do we conduct the simulations?

How do we conduct the simulations?

MD or MC moves can be used, or HMC if exact sampling is required.

Sampling could include grand-canonical moves for constant pH (and/or salt concentration).

Multiple ways to conduct MC state change move.
How do we choose states?

• High temperature barriers?
  • use high temperature replicas to overcome energy barriers: choose states as different temperatures (“simulated tempering”)
  • use umbrella sampling to drive the system throughout configuration space: choose states to be different anchor points for umbrella sampling calculations

• Alchemical transformations
  • Goal: calculate free energy difference between two Hamiltonians, via a scaling factor \( \mathcal{H}(\lambda) = \lambda \mathcal{H}_1 + (1-\lambda) \mathcal{H}_0 \)

• Convenient side effect
  • in many cases, we actually want the free energy as a function of the state, and this is obtained directly from the GE weights

Figure from D. Chandler
How do we choose the weights?

1. Start with a good initial guess

   Estimates of $<u_k>_k$ from short simulations of each state can provide an excellent guess.

   Even initial energies can provide a good initial guess.

2. Several options for automatic updating

   Wang-Landau method

   Bennett acceptance ratio (BAR)

   Weighted histogram analysis method (WHAM) or multistate BAR (MBAR)

   Adaptive Bayesian WHAM (ABWHAM)

   All are extremely simple to implement!
Adaptive Bayesian WHAM

• Setup
  • Consider a system that can be in K different states, and let $p_i$ be the probability for the i-th state
  • we want to estimate the parameters $p_i$ by means of weighted sampling
  • We seek an adaptive weighted sampling scheme as outlined on the right
  • Based on the estimates $p_i^{(n-1)}$ from the previous iteration step, new weights $w_i^{(n)}$ are determined in a way that leads to efficient sampling of states

• Adaptive scheme
  • Therefore, we attempt to develop a method in which only new data are needed for the update of estimates
  • We want to determine a new estimate $p_i^{(n)}$ from the knowledge of the new histogram $h^{(n)}$, the new weight $w_i^{(n)}$, and the previous estimate $p_i^{(n-1)}$

Estimate of $p_i$

Weight $w_i$

histogram $h$

Analysis (WHAM)

ABWHAM converges quickly

\[ f^{(n)}(\theta) \propto P(h^{(n)}|\theta,w^{(n)})f^{(n-1)}(\theta) \]

\[ P(h|\theta,w) = \frac{H!}{h_1! \cdots h_K!} \phi_1^{h_1} \cdots \phi_K^{h_K} \]

\[ P(h|\theta,w) = \frac{H!}{h_1! \cdots h_K!} (w_1 \theta_1)^{h_1} \cdots (w_K \theta_K)^{h_K} \]

Application: Fs-Peptide

C-α RMSD from ideal helix for a few representative replica walkers in SREM simulations.

A few transitions between folded and extended states for each replica are observed, indicating there is Reversible Folding.
Application: Fs-Peptide

Folded Initial Structure

Extended Initial Structure

Helical Content v.s. T

WHAM is used.


(Huang & Bowman)
Case study: small molecule drug design

http://pande.stanford.edu
Efficient free energy calculation: use forward and backward work distributions


\[
\left\langle \frac{1}{1 + \exp[\beta (M + W - \Delta G)]} \right\rangle_f = \left\langle \frac{1}{1 + \exp[\beta (M + W - \Delta G)]} \right\rangle_b
\]

\[M \equiv kT \ln \left( \frac{N_f}{N_b} \right)\]

• Plan
  • find \(P_f(W)\) and \(P_b(W)\)
  • average in a new way
  • Find \(\Delta G\) as the balancing point

• Benefit
  • two distributions are statistically linked
  • use one distribution to help flesh out the tails of the other

The tails of \(P_f(W)\) are constrained by the bulk of \(P_b(W)\) and vice versa
How does this work?

1. Start from Crooks’ fluctuation theorem

\[
\ln \left[ \frac{P_F(W)}{P_R(-W)} \right] = \beta(W - \Delta F),
\]

2. Use Bayesian method + normalization

\[
\frac{P(W|F)}{P(W|R)} = \frac{P(F|W)P(R)}{P(R|W)P(F)} = \frac{P(F|W)}{1 - P(F|W)}\frac{P(R)}{P(F)}.
\]

\[
P(F|W) + P(R|W) = 1
\]

\[
\ln \frac{P(F|W)}{1 - P(F|W)} = \beta(M + W - \Delta F)
\]

3. This leads to the probabilities

\[
P(F|W_i) = \frac{1}{1 + \exp[-\beta(M + W_i - \Delta F)]}
\]

\[
P(R|W_i) = \frac{1}{1 + \exp[\beta(M + W_i - \Delta F)]}
\]

These probabilities hold for all distributions: not parametric!

\[(Shirts, et al, PRL, 2004)\]
How does this work?

4. Define the likelihood

\[ L(\Delta F) = \prod_{i=1}^{n_F} P(F|W_i) \prod_{j=1}^{n_R} P(R|W_j). \]

5. Find maximum likelihood

\[
\frac{\partial \ln L(\Delta F)}{\partial \Delta F} = \sum_{i=1}^{n_F} \frac{1}{1 + \exp[\beta(M + W_i - \Delta F)]} - \sum_{j=1}^{n_R} \frac{1}{1 + \exp[-\beta(M + W_j - \Delta F)]} = 0.
\]

6. Result: new way to average

\[
\left\langle \frac{1}{1 + \exp[\beta (M + W - \Delta F)]} \right\rangle_f = \left\langle \frac{1}{1 + \exp[\beta (M + W - \Delta F)]} \right\rangle_b.
\]

7. Find the value of \( \Delta F \) which satisfies the above

(Shirts, et al, PRL, 2004)
Application: solvation free energies

• In our hands, BAR is most efficient
  ▪ FEP is a limiting case of BAR
  ▪ BAR appears to be more efficient than TI too

• Example test:
  ▪ 3-methlyindole (Trp sidechain analog)
  ▪ 1.0 ns at each intermediate
  ▪ We see BAR yields more precise answers for the same CPU time

<table>
<thead>
<tr>
<th>Method</th>
<th># Intermediates</th>
<th>Value (kcal/mol)</th>
<th>Precision (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TI</td>
<td>61</td>
<td>3.69</td>
<td>0.05</td>
</tr>
<tr>
<td>TI</td>
<td>8</td>
<td>4.41</td>
<td>0.21</td>
</tr>
<tr>
<td>BAR</td>
<td>8</td>
<td>3.68</td>
<td>0.05</td>
</tr>
<tr>
<td>FEP (0-&gt;A)</td>
<td>8</td>
<td>3.43</td>
<td>0.19</td>
</tr>
<tr>
<td>FEP (A-&gt;0)</td>
<td>8</td>
<td>6.01</td>
<td>0.43</td>
</tr>
<tr>
<td>FEP Average</td>
<td>8</td>
<td>4.72</td>
<td>0.24</td>
</tr>
</tbody>
</table>
Comparison with experiment

Hydration Free Energy of Amino Acid Sidechains

RMS deviations from experiment (kcal/mol):

- AMBER 1.35
- CHARMM 1.31
- OPLS-AA 0.85
Case study: protein folding kinetics

http://pande.stanford.edu
Progress of MD & experiment

Longest Simulation

- 1975: BPTI (Karplus)
- 1995: Protein A (Brooks)
- 1998: Villin (Kollman)
- 2002: BBA5 (Pande)

Fastest folding experiment

- BPTI
- SH3
- λ-repressor (Oas)
- BBA5 (Gruebele)

Time Scale:

- $10^{-15}$ femto
- $10^{-12}$ pico
- $10^{-9}$ nano
- $10^{-6}$ micro
- $10^{-3}$ milli
- $10^0$ seconds

Processes:

- Bond vibration
- Isomerization
- Water dynamics
- Helix forms
- Fastest folders
- Typical folders
- Slow folders
A very fast folding protein: $k_{\text{fold}} \sim 1/\mu\text{s}$

**villin headpiece**
mutant designed by the Eaton Lab
(Kubelka *et al*, *JMB* 2006)
Let’s look at a 1µs trajectory for villin: we see stochastic behavior.

Simulation details:
- villin headpiece (36 residues)
- Eaton mutant (0.7 µs folding time)
- explicit solvent
- 20,000 atoms total
- AMBER2003 force field

MD Engine:
- GROMACS 3.3.99 (CVS) code
- SMP on FAH

Visualization (VMD):
- spacefill: aromatic residues
- licorice: backbone
- rest: sticks
- color: N-C gradient

One trajectory of thousands, each on the >1 µs timescale

[http://simtk.org](http://simtk.org)
Looking at ensembles of simulations

• **Starting structures**
  - 9 different structures
  - generated by high temperature unfolding
  - different degrees of native like structure
  - some have helices, other contacts
  - some have no native structure at all

• **Ensemble of trajectories**
  - hundreds to thousands of trajectories per structure
  - each trajectory ~1-2 µs timescale (longer than experimental folding timescale of 0.7µs)

Ensemble data agrees with experiment

(Explicit solvent)

Fraction folded (via Trp-His distance) vs time

But is the experimental assay looking at folding?

_Fraction folded (via comparison to x-ray structure) vs time_

Comparison between explicit and implicit

(implicit solvent)

Fraction folded (via comparison to xray structure) vs time

(Ensign)
We find a heterogeneous set of folding pathways

- Do we see a single pathway or many different?

- Test this with a simple question: “Is the order of helix formation consistent between simulations?”
  - for 3 helices (villin), there are 3! = 6 possible orderings
  - histogram shows a very wide variation of pathways seen

- Other variations possible too
  - which key core contacts form first?

- A single trajectory (or even a few) would give a misleading picture of the folding dynamics

Histogram of folding kinetics: what is the order of formation of each helix A, B, C?
What have we learned about how proteins fold?

• What did we see in that trajectory?
  • starts with non-specific hydrophobic collapse
  • unfolds, breaks most contacts
  • refolds, with little native structure
  • some native persist over numerous folding/refolding cycles
  • eventually gets everything right

• What about other trajectories?
  • similar behavior in general, but different details
  • great heterogeneity in folding paths

• General lessons?
  • Folding is a stochastic process
    (if the folding time is 1ms, then it’s not ½ folded at 0.5 ms)
  • Dynamics of even small molecules can be complex & very heterogeneous
  • Even a few long trajectories aren’t enough to inform us about the true nature of the complex phase space -- we need a statistical picture
Case study: long timescale dynamics
How to overcome long timescales: stochastic kinetic sampling methods

Folding is a stochastic process with exponential kinetics

Fraction that fold:
\[ f(t) = 1 - \exp(-kt) \]

At short times, we get
\[ f(t) \approx kt \]

What if we run \( M \) Simulations in parallel each of time \( t \)?
\( Mkt \) will fold

Putting in real numbers: number that fold = \( Mkt = 10,000 \text{ simulations} \times 10,000\text{ns}^{-1} \times 100\text{ns} = 100 \text{ events!} \)
How about a new model based on states & rates

For example, consider villin with 3 helices A, B, & C
Run MD simulations from these initial states

Run ~100 trajectories from each state

(one can use many different means to generate initial conformations for MD; this is just an example)
Find new states and then repeat ...

Run a state decomposition algorithm to find new states, and then repeat

**Key concept**: timescales between states (small circles) are much faster than between $U \rightarrow F$
... until convergence

May take several (~5) rounds to converge

yields a complete description of long timescale kinetics & thermodynamics (predict rates, thermodynamics, & structure)
Automatic State Decomposition: An interactive algorithm

(Chodera & Singhal)

Collaboration between Swope, Dill, and Pande labs

Iterative refinement attempts to locate states for which there is a separation of timescales between fast intrastate dynamics and slow interstate dynamics.

Alanine dipeptide

Iterative refinement attempts to locate states for which there is a separation of timescales between fast intrastate dynamics and slow interstate dynamics.
Macrostates reveal a richer decomposition of configuration space than hypothesis-driven study (Chodera & Singhal).

In collaboration with Bill Swope & Jed Pitera (IBM Almaden) and Ken Dill (UCSF).
Adaptive sampling: a big step forward in efficiency

- **Molecular simulation as a statistical problem**
  - allows for on-the-fly adaptive methods
  - add simulations only where needed (to improve uncertainty)

- **Bayesian error analysis methods**
  - Impact
    - Optimize trajectory choice based on uncertainty
    - 100x to 1000x speed up -- calculate just what you need, not any more

Conclusions: a paradigm shift in simulation?

• Simulations are typically viewed as computational experiments
  • run, probe, then analyze
  • however, typically done anecdotally (<10 trajectories) due to computational expense

• New perspective
  • use simulations to build statistical models of the underlying phenomena
  • Bayesian inferential view of simulations -- simulations are used to parameterize our model

• Benefits
  • more powerful methods -- much longer timescales
  • a statistical view of the phenomena of interest (uncertainties, etc)
  • more much scalable than traditional MD
  • much more efficient (only simulate what you need to simulate)
A solution to the long timescale challenge?

- Use a series of complementary methods
  - default: single CPU does ~ 1ns/day ($10^9 \times$ gap)
  - Distributed computing ($10^4 \times$ to $10^5 \times$; cluster: $10^2 \times$)
  - GPU’s/streaming ($10^2 \times$ to $10^3 \times$)
  - MSMs/adaptive sampling ($10^2 \times$ to $10^3 \times$)
  - total: ($10^8 \times$ to $10^{11} \times = 0.1$ to 100 seconds per day)
Where to learn more

• Books:
  • Leach, *Molecular Modeling*: Great first resource
  • Gromacs manual ([http://gromacs.org](http://gromacs.org)): has full derivations and detailed explanations

• Wikipedia
  • believe it or not, it’s pretty well written and has lots of information

• Folding@Home: [http://folding.stanford.edu](http://folding.stanford.edu)