Simulating Biomolecules with Variable Protonation State: Constant-pH Molecular Dynamics Simulations with NAMD

Brian Radak

Leadership Computing Facility, Argonne National Laboratory

Computational Biophysics Workshop – Enhanced Sampling and Free Energy Calculations
University of Illinois, Urbana-Champaign
September 26, 2017
Acknowledgements (other people to blame)

**Univ of Chicago**
- Benoît Roux
- Donghyuk Suh
- Huan Rui
- Yunjie Chen

**ALCF**
- Sunhwan Jo (SilcsBio)
- Wei Jiang

**UIUC**
- Jim Phillips
- Chris Chipot
- Abhi Singharoy
pH Effects in Biochemistry

variability of pH by region


enzyme rate vs. pH


pH gradients at cell surfaces
Constant pH and the semi-grand canonical ensemble

- Conventional MD samples a canonical ensemble:
  \[ Q = \int dx \, e^{-\beta U(x)} \]

- Constant-pH MD samples a semi-grand canonical ensemble:
  \[ \Xi(\text{pH}) = \sum_{\lambda \in S} Q_\lambda 10^{-n_\lambda \text{pH}} \]

The added interaction is between the number of protons, \( n_\lambda \), and a pH bath. \( \lambda \) is a new variable designating the protonation state.
Networks of protonation states

conventional MD

constant pH MD
Networks of protonation states

$E-R_1:H:R_2H \xleftrightarrow{-H^+} E-R^-_1:R_2H$

$E-R_1:H:R_2^- \xleftarrow{-H^+} E-R^-_1:R_2^-$

conventional MD

$E-R_1:H:R_2H \xleftrightarrow{-H^+} E-R^-_1:R_2H$

$E-R_1:H:R_2^- \xleftarrow{-H^+} E-R^-_1:R_2^-$

constant pH MD

$N = 1 \quad 2 \quad 3 \quad 4$
pH as a *thermodynamic force*

- Classical MD utilizes *mechanical* forces

\[
F = -\nabla U[x(t)] = m \frac{\partial v}{\partial t}; \quad v = \frac{\partial x}{\partial t}
\]

- pH may be regarded as a *thermodynamic force*

\[
\ln 10 \text{pH} = -\frac{\partial \ln \Xi}{\partial n_\lambda}
\]

Mechanical forces – deterministic/stochastic dynamics
Thermodynamic forces – probabilistic “dynamics”

\[
P_\lambda(\text{pH}) \propto Q_\lambda 10^{-n_\lambda \text{pH}}
\]
How do we define nodes in the network?

Consider a system with $m$ sites:

$$\lambda = \{ \lambda_1, \lambda_2; \ldots \lambda_s, \lambda_{s+1}; \ldots \lambda_m \}$$
Protonation state probabilities/populations

\[ \langle A(x, \lambda) \rangle_{pH} = \sum_{\lambda \in S} \int d x \ A(x, \lambda) e^{-\beta U(x; \lambda)} 10^{-n x_pH} / \Xi(pH) \]

\[ P_{\lambda_s} = \langle \lambda_s \rangle_{pH} \quad \text{– the probability that site } s \text{ is occupied} \]

There are two kinds of terms in the summation, \( \lambda_s = 0/1 \)

\[ \Xi(pH) = \Xi_0(pH) + \Xi_1(pH)10^{-pH} \]

thus,

\[ \langle \lambda_s \rangle_{pH} = \frac{\Xi_1(pH)10^{-pH}}{\Xi_0(pH) + \Xi_1(pH)10^{-pH}} = \frac{1}{1 + \frac{\Xi_0(pH)}{\Xi_1(pH)} 10^{pH}} \]
Connection to thermodynamics

\[
\langle \lambda_s \rangle_{\text{pH}} = \frac{1}{1 + \frac{\Xi_0(\text{pH})}{\Xi_1(\text{pH})}10^{\text{pH}}}
\]

compares to the Henderson-Hasselbalch equation such that

\[
pK_a(\text{pH}) = -\log \frac{\Xi_0(\text{pH})}{\Xi_1(\text{pH})},
\]

except that now \( pK_a(\text{pH}) \) is \( \text{pH} \) dependent. One often uses the approximation:

\[
pK_a(\text{pH}) \approx pK_a^{(a)} + (1 - n) \left( \text{pH} - pK_a^{(a)} \right),
\]

where \( n \) is the Hill coefficient and \( pK_a^{(a)} \) is the “apparent” \( pK_a \).
We can now see that the fraction of simulation time spent in a given protonation state is directly impacted by the difference of the $pK_a$ of a residue/site and the pH.
That’s great – how do we sample the states?

1. Sample the configuration space of a given state (i.e., sample $x$ for a given $Q_{\lambda}$)

2. Change between protonation states according to the number of protons and the given pH (i.e., sample $\lambda$ and choose a new $Q_{\lambda}$)

This may be regarded as a Gibbs sampling, whereby the configuration and state are sampled in an alternating fashion.
A problem! Environmental response

- (De)Protonation is a significant electrostatic event.
- Non-trivial reorganization of solvent, possibly solute.
- Naive sudden changes in protonation are likely to cause high energy configurations and/or steric clashes.
Possible solutions to the solvent clash problem

- **Auxillary Implicit Solvent**

- **Continuous Fractional Proton**

- **Discrete Copy Fractional Proton**
“Fast” alchemical growth

- Swap the protonation state by using time-dependent interactions.
- Gradually stronger interactions will induce solvent response.
- Clashes are avoided by using the natural dynamics of the model.
The neMD/MC constant pH paradigm

- Drive alchemical growth with *nonequilibrium* work
- Accept/reject with a generalized Metropolis criterion

The neMD/MC constant pH paradigm

- Drive alchemical growth with *nonequilibrium* work
- Accept/reject with a generalized Metropolis criterion

Beyond Gibbs sampling: Hybrid MD and neMD/MC

We now alternate conventional sampling with MD ($\mathbf{x}$) and Metropolis Monte Carlo sampling ($\mathbf{x}$ and $\lambda$):

$$\rho(\mathbf{x}, \lambda) T(\mathbf{x}, \lambda \rightarrow \mathbf{x}', \lambda') = \rho(\mathbf{x}', \lambda') T(\mathbf{x}', \lambda' \rightarrow \mathbf{x}, \lambda)$$

such that the neMD/MC transition probability is:

$$T(\mathbf{x}, \lambda \rightarrow \mathbf{x}', \lambda') = \min \left[ 1, \frac{\rho(\mathbf{x}', \lambda')}{\rho(\mathbf{x}, \lambda)} \right] = \min \left[ 1, e^{-\beta W 10^{-\Delta n pH}} \right]$$

(If you’d like, MD uses the probability $T(\mathbf{x} \rightarrow \mathbf{x}') = 1$.)
Important considerations

- How long should I sample the equilibrium stage?

- How long should I sample the nonequilibrium stage (the “switch time,” $\tau_{\text{switch}}$)

- Rejecting a nonequilibrium trajectory is expensive, how can we avoid doing that so much?
The two-step “inherent” $pK_a$ algorithm

\[
T(x, \lambda \rightarrow x', \lambda') = T^{(i)}(\lambda \rightarrow \lambda') T^{(s)}(x \rightarrow x'|\lambda \rightarrow \lambda')
\]

\[
T^{(i)}(\lambda \rightarrow \lambda') = \min \left[ 1, 10^{pK_a^{(i)}(\lambda,\lambda') - \Delta npH} \right]
\]

- neMD/MC can be split into two parts
  1. $T^{(i)}$ – only depends on $\lambda$ and the pH – CHEAP
  2. $T^{(s)}$ – depends on the switch ($W$) – COSTLY

- Effort is shifted by estimating a parameter, $pK_a^{(i)}$
- Optimal efficiency achieved for exact $pK_a$
- Dramatically improved performance on wide pH ranges!
- Can do even better for systems with more than two states.

Let’s look at this graphically

- It’s silly to try to add/remove protons to/from acidic/basic residues at high/low pH
- Transitions are proposed in proportion to the estimated population.
What about after we’ve proposed a switch?

- A short switch will not change much and likely be rejected.
- A long switch is expensive (limit of a single switch – BAD).
- Since the switch success depends on the work, let’s analyze that.
Work and force fluctuations – a typical neMD/MC cycle

Theoretical and Empirical Performance Analysis

- Well-defined criteria for optimization.
- Cost is quite tractable.

NAMD Constant pH features

- Flexible Tcl interface source
  ```
  ...lib/namdcph/namdcph.tcl
  ```
- PSF build procedure is unchanged (automated psfgen)
- Implemented with PME and full electrostatics
- Normal CPU scaling (no GPU yet) - depends on alchemy
- Companion analysis script cphanalyze

```plaintext
pH 7.0
  cphNumstepsPerSwitch 7500 ;# run 7500 steps per switch
  cphRun 5000 10 ;# run 10 cycles of 5000 MD steps
```
Output and Analysis

- Normal usage requires multiple pH values ("titration curves")
- New output cphlog and cphrst
- New checkpoint files psf/pdb
- Can boost performance with WHAM (cphanalyze)
- Can also analyze residue correlations
Reference amino acids are well-reproduced

- Adjustments to force field enforce empirical reference values
- Implicitly model solvated proton and bond energy effects
- Bonus - accurate reproduction of tautomeric ratios!
What about challenging environments?

- Single titratable peptide (AADAA)
- Lipid relaxation is slow (slower than water)
- Low dielectric region should perturb pKa in obvious way
- Significant shifts due to low dielectric region.
- Switch time of 10 ps is sufficient.

Staph nuclease (SNase) - A constant pH benchmark
Benchmarking of SNase pK$_a$ values

- Good correlation with measured values for carboxylates
- Bonus - estimates for HIS and LYS

<table>
<thead>
<tr>
<th>residue</th>
<th>this work</th>
<th>his work</th>
</tr>
</thead>
<tbody>
<tr>
<td>LYS</td>
<td>24</td>
<td>8.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.45)</td>
</tr>
<tr>
<td>HIS</td>
<td>8</td>
<td>6.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.56)</td>
</tr>
<tr>
<td></td>
<td>121</td>
<td>5.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.50)</td>
</tr>
</tbody>
</table>
Concluding Remarks/Future Directions

- Things we are working on:
  - Performance improvements in alchemy – CUDA support
  - Titratable lipids and phosphates

- Things we would like to work on:
  - psfgen improvements – support for Drude
  - Better visualization support in VMD
  - More powerful interface for analysis (PyNAMD)
Acknowledgements (other people to blame)

Univ of Chicago
- Benoît Roux
- Donghyuk Suh
- Huan Rui
- Yunjie Chen

ALCF
- Sunhwan Jo (SilcsBio)
- Wei Jiang

UIUC
- Jim Phillips
- Chris Chipot
- Abhi Singharoy