All-Atom Free Energy Calculations of Biological Molecules

Sahand Jamal Rahi

May 16, 2006

1 Introduction

To understand biological and chemical phenomena at the molecular level it is important to measure the free energy of the system, which may be a protein in solution. When temperature and resulting solvation effects are important the energy of the system alone does not suffice to describe the properties of the system.

The following discussion is largely based on references [1, 2].

Unfortunately, computing the free energy is difficult. Classically, it involves computing an integral over all possible states of the system convering the entire phase space:

$$Z = \frac{1}{h^{3N} N!} \int d\Gamma \exp(-\beta H(\Gamma)) \tag{1}$$

and integrating out momentum

$$Z = Z_{\text{ideal gas}} V^{-N} \int dr \exp(-\beta U(r)). \tag{2}$$

The final integral captures the configurational contribution to the partition function so we shall call it Z_{conf}.

For a system with many positional degrees of freedom even this integral is too difficult to compute. But since the probability density for a configuration denoted by a $3N$-dimensional position vector r' is:

$$p(r') = \frac{\exp(-\beta U(r'))}{Z_{\text{conf}}} \tag{3}$$

we may hope that by simulating a system and computing some average we may tease out Z_{conf}. If a system behaves ergodically the time average of any quantity equals its ensemble average. So, for example, we may study:

$$\langle \exp(\beta U(r)) \rangle = \int dr \frac{\exp(\beta U(r)) \exp(-\beta U(r))}{Z_{\text{conf}}} = \frac{V^N}{Z_{\text{conf}}} \tag{4}$$

We would find the average quantity on the left hand side by simulating the system at room temperature and find Z_{conf}. However, as noted in reference [2], large negative energies predominate in the simulation path due to their favorable Boltzmann weight whereas these states contribute exponentially small numbers to the average. One cannot expect acceptable convergence of the average within the usual time lengths of computer simulations.

2 Summary of Popular Free Energy Computation Technique(s)

Usually, free energy *differences* between different systems are of greater interest than their absolute values:

$$\begin{aligned} F_B - F_A &= -\beta^{-1}\log(Z_B) + \beta^{-1}\log(Z_A) = -\beta^{-1}\log\left(\frac{Z_B}{Z_A}\right) \\ &= -\beta^{-1}\log\left[\frac{\int d\Gamma e^{-\beta H_B(\Gamma)}}{\int d\Gamma e^{-\beta H_A(\Gamma)}}\right] \\ &= -\beta^{-1}\log\left[\frac{\int d\Gamma e^{-\beta \Delta H_{BA}(\Gamma)} e^{-\beta H_A(\Gamma)}}{\int d\Gamma e^{-\beta H_A(\Gamma)}}\right] \\ &= -\beta^{-1}\log\langle e^{-\beta \Delta H_{BA}(\Gamma)}\rangle_A \end{aligned} \tag{5}$$

In order to be able to express $H_B(\Gamma) = \Delta H_{BA}(\Gamma) + H_A(\Gamma)$ the phase spaces of system A and B must be compatible or must be made compatible. For example, if system A has more atoms than system B one inserts atoms with no mass and no interactions into B to make the phase spaces formally identical. This ought to always be possible if the comparison between systems A and B makes physical sense.

The final ensemble average of $e^{-\beta\Delta H_{BA}}$ can be obtained through simulation: System A is simulated for sufficiently long times and $e^{-\beta\Delta H_{BA}}$ is averaged over the simulation steps. A simulation of system A will tend to sample the low-energy states of system A. If these do not correspond to low-energy states of system B, which means that the simulation does not sample the system close enough to the favorable configurations of system B, ΔH_{BA} will predominantly be large. In this case, again, convergence of the average will be slow at best.

By the way, note that, of course, A and B can be switched, the answer ought to be the reciprocal of $\langle e^{-\beta\Delta H_{BA}(\Gamma)}\rangle_A$. However, usually it is not, due to the sampling errors.

In order to minimize the sampling error effects, the free energy difference is usually broken up into free energy differences with intermediate states 'between' A and B with Hamiltonians $H(\lambda) = (1-\lambda)H_A + \lambda H_B$. Such Hamiltonians may not be physical but that does not matter.

$$\begin{aligned} F_B - F_A = F(\lambda=1) - F(\lambda=0) &= \sum_{i=0}^{N-1}\left[F\left(\frac{i+1}{N}\right) - F\left(\frac{i}{N}\right)\right] \\ &= -\beta^{-1}\sum_{i=0}^{N-1}\log\langle e^{-\beta\Delta H_{\lambda_{i+1}\lambda_i}(\Gamma)}\rangle_{\lambda_i} \end{aligned} \tag{6}$$

and in the limit as $N \to \infty$:

$$F_B - F_A = \int_0^1 d\lambda \frac{\partial F(\lambda)}{\partial \lambda} = \int_0^1 d\lambda \left\langle \frac{\partial H(\lambda)}{\partial \lambda}\right\rangle_\lambda . \tag{7}$$

Both equations 6 and 7 lead to different simulation treatments. The derivative-integral approach, called the thermodynamic integration technique, appears to be more popular. Here, the derivative of the energy with respect to λ is averaged for a system governed by $H(\lambda)$ for many values of λ and the integral is approximated by a sum. The method involving averages over exponentials (6) is refered to as Free Energy Perturbation (FEP), largely for historical reasons since perturbation theory has little to do it.

Most other techniques that I have encountered in my literature search seek to improve on these techniques outlined here, some are more interesting than others, I have chosen to try these out for this final project before moving to more complicated techniques.

3 Shift in pKa (acidity) of the Amino Acid Aspartic Acid due to Presence of Protein

I followed an online tutorial (reference [5]). The goal is to repeat the computations published in reference [4]. The amino acid aspartic acid has a different acidity depending on its environment: as part of a protein or in free form. The $pK_a$is defined as follows and the shift we are looking for is given in the second line:

$$pK_a = -\log_{10} K_a = -\log_{10} \frac{[H+][A-]}{[HA]} = -\log_{10} e^{-\beta\Delta G} = \frac{\beta\Delta G}{\log 10} \tag{8}$$

$$pK_a(\text{in protein}) - pK_a(\text{free}) = \frac{\beta\Delta\Delta G}{\log 10} \tag{9}$$

Throughout we shall assume that $\Delta F \approx \Delta G$ since volume and pressure changes ought to be negligible.

3.1 Methods

The simulation package Amber 9.0 was used with the amber 94 force field and the Generalized Born solvent model of Onufriev, Bashford, and Case ([3]). Four structures were generated: protonated and unprotonated aspartic acid with methyl groups protecting the amino and the carboxylate groups as well as thioredoxin protein with protonated and unprotonated ASP26. As mentioned above, degrees of freedom cannot simply be changed if one wants to interpolate between Hamiltonians, hence, the Hydrogen atom in the unprotonated structures was not removed but its interactions were simply turned off, which is equivalent to it disappearing. (Although the mass of the dummy Hydrogen cannot be turned off, it does not affect the answer since Amber does not output $\frac{\partial H}{\partial \lambda}$ but $\frac{\partial U}{\partial \lambda}$). Notice that the free energy of the free proton is not accounted for at all because it cancels when computing the difference in the free energies. Its electrostatic interaction with the two anions may be different, which could be ignored if some natural salt concentration was taken into account, however, this was not done.

One could evaluate $\frac{\partial H}{\partial \lambda}$ for λ evenly spaced between 0 and 1 and then add the values multiplied by the spacing to obtain the integral. In the literature, however, the method of Gaussian Quadrature is widely used and recommended. The Gaussian Quadrature formula requires n prespecified data points and gives the exact formula for the integral for polynomials fitting the data up to degree $2n + 1$.

3.2 Results

Following the Gaussian quadrature prescription the following simulations were run:

Interpolations between Structures	Total Length of Simulation	Length of Simulation used for Average	λ	$\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda$ [kcal/mol]
protonated → unprotonated free aspartic acid	3 ns	2 ns	0.0000	-57.9283
			0.1127	-58.3121
			0.5000	-59.6893
			0.8873	-61.8471
			1.0000	-62.1924
protonated → unprotonated ASP26 in thioredoxin	1 ns	800 ps	0.0000	N/A
			0.1127	-49.8394
			0.5000	-53.6001
			0.8873	-62.1712
			1.0000	N/A

The first five simulations were run on my computer, for the latter three I used the results given in the tutorial since the computational time for them would exceed 1 month. The Gaussian Quadrature method turns out not to require the value of the integrand at the end points, so they have not been determined for the second set of simulations. In the following I have used the tabulated weights for the Gaussian Quadrature integral approximation:

$$\begin{aligned}\Delta F(\text{free aa}) &= \int_0^1 d\lambda \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_\lambda \\ &\approx 0.27777 * (-58.3121\text{kcal/mol}) + 0.44444 * (-59.6893\text{kcal/mol}) \\ &+ 0.27777 * (-61.8471\text{kcal/mol}) \\ &= -59.9049\text{kcal/mol}\end{aligned} \tag{10}$$

$$\begin{aligned}\Delta F(\text{aa in protein}) &\approx 0.27777 * (-49.8394\text{kcal/mol}) + 0.44444 * (-53.6001\text{kcal/mol}) \\ &+ 0.27777 * (-62.1712\text{kcal/mol}) \\ &= -55.0886\text{kcal/mol}\end{aligned} \tag{11}$$

$$\begin{aligned}\text{pK}_\text{a}(\text{in protein}) - \text{pK}_\text{a}(\text{free}) &= \frac{\beta \Delta\Delta G}{\log 10} \\ &\approx \frac{(-55.0886\text{kcal/mol} + 59.9049\text{kcal/mol})}{2.303 * 0.001987 * 300\text{kcal/mol}} \\ &= \frac{4.8163\text{kcal/mol}}{2.303 * 0.001987 * 300\text{kcal/mol}} \\ &= 3.5\end{aligned} \tag{12}$$

The value of $\Delta\Delta G$ obtained agrees surprisingly well with the experimental value of 4.8kcal/mol. Simonson et al. [4], whose work has been imitated here, however, found $\Delta\Delta G$ to be further from the experimental value although their simulations were run for longer periods. This most likely means that I have just been lucky.

The result, in any case, shows that the protein strongly suppresses the acidity of aspartic acid. This may be highly relevant biologically, however, it is beyond the scope of this project to investigate this aspect further.

References

[1] Christophe Chipot and Andrew Pohorille. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*. Springer, 2007.

[2] P. M. King. Free energy via molecular simulation: A primer. In Wilfred F. van Gunsteren, Paul K. Weiner, and Anthony J. Wilkinson, editors, *Computer Simulation of Biomolecular Systems*, pages 267–314. ESCOM, 1993.

[3] A. Onufriev, D. Bashford, and D. A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*, 55:383–394, 2004.

[4] Thomas Simonson, Jens Carlsson, and David A. Case. Proton binding to proteins: pKa calculations with explicit and implicit solvent models. *J. Am. Chem. Soc.*, 126:4167–4180, 2004.

[5] Ross Walter and Mike Crowley. http://amber.scripps.edu/tutorials/advanced/tutorial6/index.htm.