A Bayesian Approach to Calculating Free Energies in Chemical and Biological Systems

Andrew Pohorille

NASA Ames Research Center,
Exobiology Branch, MS 239–4
Moffett Field, California 94035–1000, USA

Abstract. A common objective of molecular simulations in chemistry and biology is to calculate the free energy difference between systems of interest. We propose to improve estimates of these free energies by modeling the underlying probability distribution as a the square of a "wave function", which is a linear combination of Gram-Charlier polynomials. The number of terms, N, in this expansion is determined by calculating the posterior probability, $P(N \mid X)$, where X stands for all energy differences sampled in a simulation. The method offers significantly improved free energy estimates when the probability distribution is broad and non–Gaussian, even if sample size is small. This makes it applicable to challenging problems, such as protein–drug interactions.

Key Words: Free energy, Gram–Charlier polynomials, Maximum Likelihood.

INTRODUCTION

To understand chemical and biological processes at a molecular level, it is often necessary to examine their underlying free energy behavior. This is the case, for instance, in protein folding, protein–ligand, protein–protein and protein–DNA interactions, and in drug partitioning across the cell membrane. These processes, which are of paramount importance in the fields of biotechnology and computer-aided, rational drug design, cannot be predicted reliably without the knowledge of the associated free energy changes.

The Helmholtz free energy, A in the canonical ensemble can be expressed in terms of the partition function, Q

$$A = -\beta^{-1} \ln Q = -\beta^{-1} \ln \frac{1}{N! h^{3N}} \int \exp\left[-\beta H\left(\mathbf{x}, \mathbf{p}_x\right)\right] d\mathbf{x} d\mathbf{p}_x \tag{1}$$

where N is the number of particles, h is the Planck constant, $\beta = 1/kT$, k is the Boltzmann constant and T is temperature. From this equation it follows that calculating A is equivalent to estimating Q, which is a very difficult undertaking. In both experiments and calculations, however, we are interested in free energy *differences*, ΔA, between two systems, say 0 and 1, described by the partition functions Q_0 and Q_1, respectively.

$$\Delta A = -\beta^{-1} \ln Q_1/Q_0 \tag{2}$$

This equation indicates calculating ΔA requires determining the ratio of Q_1/Q_0 rather than individual partition functions. On the basis of computer simulations this can be done in various ways [1]. One approach is to transform Eq. (2) as follows:

$$\Delta A = -\beta^{-1} \ln \frac{\int \exp\left[-\beta U_1(\mathbf{x})\right] d\mathbf{x}}{\int \exp\left[-\beta U_0(\mathbf{x})\right] d\mathbf{x}} = -\beta^{-1} \ln \langle \exp\{-\beta(\Delta U)\}\rangle_0 \tag{3}$$

Here, potential energy functions for systems 0 and 1 are $U_0(\mathbf{x})$, and $U_1(\mathbf{x})$, respectively,

$$P_0(\mathbf{x}) = \frac{\exp\left[-\beta_0 U_0(\mathbf{x})\right]}{Z_0} \tag{4}$$

is the probability density function of finding system 0 in the microstate defined by particle positions $\mathbf{x}$, $\Delta U = U_1(\mathbf{x}) - U_0(\mathbf{x})$ and $\langle \ldots \rangle_0$ denotes an average over the ensemble 0. This indicates that ΔA can be calculated by sampling system 0 only. Since ΔA is evaluated as the average of a quantity that depends only on ΔU, it can be expressed as a one–dimensional integral over energy difference:

$$\Delta A = -\beta^{-1}\ \ln \int \exp\left(-\beta \Delta U\right)\ P_0(\Delta U)\ d\Delta U \tag{5}$$

where $P_0(\Delta U)$ is the probability distribution of ΔU sampled for system 0. If energies were the functions of a sufficient number of identically distributed random variables, then $P_0(\Delta U)$ would be a Gaussian, as a consequence of the central limit theorem. In practice, it deviates from a Gaussian, but is still "Gaussian–like". To yield free energy, $P_0(\Delta U)$ is integrated with the Boltzmann weighting factor $\exp(-\beta \Delta U)$. This means that the poorly sampled, negative ΔU tail of the distribution provides the dominant contribution to the integral, whereas the contribution from the well sampled region around the peak of $P_0(\Delta U)$ is small. This is illustrated in Figure 1.

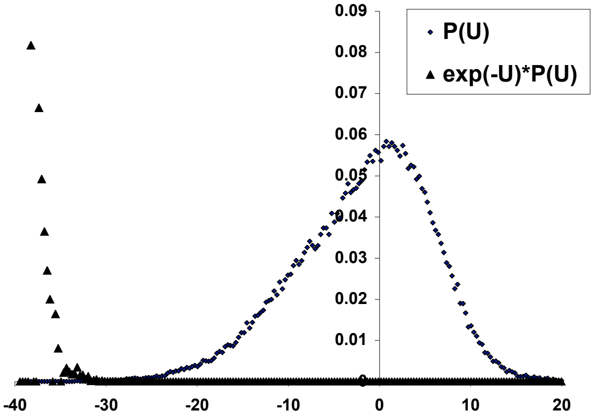

FIGURE 1. $P_0(\Delta U)$ (circles) and the integrand in equation (5), $\exp(-\beta \Delta U)\, P_0(\Delta U)$ (triangles). Only the right side of the integrand is sampled, which precludes accurate estimation of the integral.

It would be natural to exploit our knowledge of the whole $P_0(\Delta U)$, rather than its low–ΔU tail only. The simplest strategy is to model $P_0(\Delta U)$ as an analytical function or a series expansion whose adjustable parameters are determined primarily from the well–sampled region of the function. In general, such approach fails, because its reliability deteriorates away from the region, in which the function is known with a good accuracy. Here, however, we might be successful, because $P_0(\Delta U)$ is smooth and Gaussian–like. So far, there have been only a few attempts at modeling $P_0(\Delta U)$. One is to represent it as a linear combination of Gaussian functions [2]. Another model is sometimes called the "universal" probability distribution function [3], because it has been suggested that it suitably represents global quantities in a broad class of finite–size, equilibrium or non–equilibrium systems characterized by strong correlations and self–similarity. Below we propose a different model and a more systematic approach to the problem.

A BAYESIAN APPROACH TO MODELING THE PROBABILITY DISTRIBUTION

We expand $P_0(\Delta U)$ using Gram–Charlier polynomials, which are the products of Hermite polynomials and a Gaussian function [4] and are particularly suitable for describing near–Gaussian functions. To ensure that $P_0(\Delta U)$ is always positive, we take

$$P_0(\Delta U) = \left(\sum_{n=0}^{\infty} c_n \varphi_n \left(\Delta U \right) \right)^2 \tag{6}$$

where c_n is the n–th coefficients of the expansion and ϕ_n is the n–th normalized Gram–Charlier polynomial, identical to the wave functions for the n–th excitation levels of the quantum harmonic oscillator and related to the n–th Hermite polynomial by:

$$\varphi_n \left(x \right) = \frac{1}{\sqrt{2^n \pi^{1/2} n!}} H_n \left(x \right) \exp \left(-x^2/2 \right) \tag{7}$$

The coefficients $\{c_n\}$ are constrained by the normalization condition for $P_0(\Delta U)$

$$\sum_n c_n^2 = 1 \tag{8}$$

The expansion in (6) is complete and convergent. This nice, formal property is, however, not particularly helpful in practice because only the first few coefficients in the expansion can be determined from simulations with sufficient accuracy. This means that (6), or any other expansion, is useful only if it converges quickly.

The above considerations raise a question: how to determine the optimal N and the coefficients $\{c_n\}$, $n \leq N$ in (6)? If the expansion is truncated too early, some terms that contribute importantly to $P_0(\Delta U)$ are lost. On the other hand, terms above some threshold carry no information, and only add noise to the probability distribution.

Our follow a standard Bayesian approach to find the optimal N. The data consist of M statistically independent samples of ΔU collected in computer simulations. For convenience, the energies are taken in units of β, rescaled to $x = U/\sqrt{2}\sigma$, where σ is the variance of $P_0(\Delta U)$, and shifted such that zero of energy is equal to the average ΔU. The M-dimensional vector with the values of x and the N-dimensional vector with the coefficients in the expansion (6) are denoted X and C_N, respectively. The goal is to calculate the posterior probability, $P(N \mid X)$, that the data were generated from the expansion (6) truncated after the first $N+1$ terms

$$P(N \mid X) = \frac{P(X \mid N)P(N)}{P(X)}. \tag{9}$$

If the prior, $P(N)$, is uniform for all N between 0 and N_{max} the posterior becomes proportional to the likelihood function, $P(X \mid N)$

$$P(N \mid X) \propto P(X \mid N). \tag{10}$$

The probability, $P(X \mid N)$ of generating data X given N depends on C_N. Since we are not interested in this dependence here, we marginalize C_N

$$P(X \mid N) = \int P(X, C_N \mid N) dC_N = \int P(X \mid C_N, N) P(C_N \mid N) dC_N \tag{11}$$

where dC_N stands for $dc_0 \ldots dc_N$ and the second equality follows from the product rule.

Next, we expand $P(X \mid C_N, N)$ around $P(X \mid C_N^0, N)$, where C_N^0 stands for the N-dimensional vector with the maximum likelihood (ML) coefficients, c_n^0. To obtain C_N^0 we find the extremum of $\ln P(X \mid C_N, N)$, subject to the normalization constraint (8). The problem can be readily solved using Lagrange multipliers. We first note that for statistically independent samples

$$P(X \mid C_N, N) = \prod_{\mu=1}^{M} P(x_\mu \mid C_N, N) \tag{12}$$

where $P(x_\mu \mid C_N, N)$ is the probability of generating a sample point x_μ from an expansion of $P_0(\Delta U)$ to order N. After substituting the explicit form of $P(x_\mu \mid C_N, N)$ from (6), the function to be minimized is:

$$f(C, N) = 2\sum_\mu [\ln \sum_n c_n \varphi_n(x_\mu)] + \lambda \sum_n c_n^2. \tag{13}$$

where λ is the Lagrange multiplier. For $f(C, N)$ to be an extremum, its first derivatives with respect to $\{c_n\}$ must vanish. This leads to a set of $N+1$ equations for $\{c_n\}$

$$\sum_\mu \frac{\varphi_m(x_\mu)}{\sum_n c_n \varphi_n(x_\mu)} + \lambda c_m = 0 \tag{14}$$

which are solved simultaneously with (8).

Equations (14) have a simple interpretation. If we apply the relation

$$\frac{1}{M} \sum_\mu f(x_\mu) \approx \int f(x) P(x) dx. \tag{15}$$

for a discrete sample of a function $f(x)$ to the sum on the left hand side of (14) and take advantage of orthonormality of φ_n we obtain

$$\sum_{\mu} \frac{\varphi_m(x_\mu)}{\sum_n c_n \varphi_n(x_\mu)} = M \sum_n c_n \int \varphi_m(x) \varphi_n(x) dx = M c_m. \tag{16}$$

This means that (14) are $N+1$ equations that enforce orthonormality of φ_n sampled at $\{x\}$. From these equations it also follows that $\lambda = -M$.

Returning to $P(X \mid C_N, N)$, we first note that the direct expansion of this probability density around $P(X \mid C_N^0, N)$ diverges. Instead, we represent $P(X \mid C_N, N)$ as:

$$P(X \mid C_N, N) = \exp\left[\ln P(X \mid C_N, N)\right] \tag{17}$$

and expand $\ln P(X \mid C_N, N)$ in the Taylor series. This yields:

$$\ln P(X \mid C_N, N) = \ln P(X \mid C_N^0, N) + 2 \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k} \sum_{\mu} (S_\mu)^k \tag{18}$$

where

$$S_\mu = \frac{\sum_m \Delta c_m \varphi_m(x_\mu)}{\sum_n c_n^0 \varphi_n(x_\mu)} \tag{19}$$

and $\Delta c_n = c_n - c_n^0$. If we truncate the expansion in (18) after second–order

$$P(X \mid C_N, N) = P\left(X \mid C_N^0, N\right) \exp\left(2 \sum_{\mu} S_\mu - \sum_{\mu} S_\mu^2\right). \tag{20}$$

In the absence of the normalization constraint the linear term would vanish. In this case, however, it does not, but it can be easily evaluated:

$$2 \sum_{\mu} S_\mu = 2 \sum_m \Delta c_m \sum_{\mu} \frac{\varphi_m(x_\mu)}{\sum_n c_n^0 \varphi_n(x_\mu)} = 2M \sum_m \Delta c_m c_m^0 = -M \sum_m \Delta c_m^2. \tag{21}$$

In the second equality we used (14), and in the third we took advantage of the relation $2 \sum_n \Delta c_n c_n^0 = -\sum_n \Delta c_n^2$. The linear term can be represented in a matrix notation:

$$2 \sum_{\mu} S_\mu = -\Delta C^T \mathbf{M} \Delta C \tag{22}$$

where ΔC is a N–dimensional vector with the coefficients Δc_n, ΔC^T is its transpose and $\mathbf{M}$ is a $N \times N$ matrix, whose entries are $M \delta_{mn}$.

We can proceed similarly with the second–order term. Using (19) we obtain:

$$\sum_{\mu} S_\mu^2 = \Delta C^T \mathbf{A} \Delta C \tag{23}$$

where $\mathbf{A}$ is a $N \times N$ matrix, whose entries are:

$$A_{nm} = \sum_{\mu} \frac{\varphi_n(x_\mu) \varphi_m(x_\mu)}{\left[\sum_n c_n^0 \varphi_n(x_\mu)\right]^2}. \tag{24}$$

After substituting (22) and (23) to (20) and defining $\mathbf{\Lambda} = \mathbf{A} - \mathbf{M}$, we obtain an equation for $P(X \mid C_N, N)$ in a χ^2 form

$$P(X \mid C_N, N) = P\left(X \mid C_N^0, N\right) \exp\left(-\Delta C^T \mathbf{\Lambda} \Delta C\right) \tag{25}$$

which we substitute to (11) to obtain

$$P(X \mid N) = P\left(X \mid C_N^0, N\right) \int \exp\left(-\Delta C^T \mathbf{\Lambda} \Delta C\right) P(C_N \mid N) dC_N \tag{26}$$

We take the prior, $P(C_N \mid N)$, to be uniform, subject to the constraint (8). This means that it is uniform on a N-dimensional unit hypersphere and is zero otherwise. Since the constraint has already been included in the equation through $\mathbf{M}$ we get:

$$P(X \mid N) = P\left(X \mid C_N^0, N\right) \int \exp\left(-\Delta C^T \mathbf{\Lambda} \Delta C\right) dC_N. \tag{27}$$

This is a standard multivariate Gaussian integral that can be evaluated by calculating the determinant or through diagonalization of $\mathbf{\Lambda}$. For sample sizes that we deal with in real simulations, the Gaussians are always quite sharp. Then, after integration we have:

$$P(X \mid N) = P\left(X \mid C_N^0, N\right) \prod_{n=0}^{N} \frac{2\sqrt{\pi}}{\sqrt{\lambda_0 \ldots \lambda_N}} \tag{28}$$

where $\lambda_0 \ldots \lambda_N$ are the eigenvalues of $\mathbf{\Lambda}$ and the extra factor of 2 in front of $\sqrt{\pi}$ follows from the quadratic form of $P_0(\Delta U)$, which always yields two symmetric solutions for C. More conveniently

$$\ln P(X \mid N) = \ln P\left(X \mid C_N^0, N\right) - \left[\frac{1}{2}\sum_{n=0}^{N} \ln \lambda_n + N \ln 2\sqrt{\pi}\right] \tag{29}$$

As expected, the solution for the logarithm of the posterior consists of two terms which change oppositely with N. The first term, which represents the optimal (ML) solutions, always increases with N towards its asymptotic value. The second term, which represents an “Ockham razor" penalty for increasing the number of terms in the expansion, decreases with N.

SIMULATION RESULTS

For a numerical test of (29) we chose a challenging case, in which $P_0(\Delta U)$ is broad and clearly non–Gaussian. Instead of considering a real chemical system, we constructed a synthetic $P_0(\Delta U)$, which resembled those of systems with ionic interactions, but was a linear combination of 3 Gaussians, $p_i(\Delta U)$, with different mean values and variances:

$$P_0(\Delta U) = \sum_{i=1}^{3} w_i p_i(\Delta U) \tag{30}$$

where w_i was the weight of the i–th Gaussian, subject to the constraints $w_i \geq 0, \sum w_i = 1$. The mean values, $\langle \Delta U \rangle_i$, variances, σ_i, and weights of each Gaussian were: (3.0, 4.0, 0.3), (-3.0, 7,0, 0.5) and (-6.0, 9.0, 0.2) The resulting $P_0(\Delta U)$ is shown in Fig. 1. The main advantages of using a multi–Gaussian $P_0(\Delta U)$ are that it can be easily sampled and that the free energy, ΔA, can be calculated exactly as:

$$\Delta A = -\ln \sum_{i=1}^{3} w_i \exp\left(-\langle \Delta U \rangle_i + \sigma_i^2/2\right) \tag{31}$$

For this system we generated 20 datasets of 100,000 and 20 datasets of 1,000 statistically independent values of x. For comparison, we also generated 20 datasets of 100,000 values of x sampled from a Gaussian with the mean value of zero and $\sigma = 8$. For each dataset, we calculated the free energy from (5) and from the expansion (6) for $0 \leq N \leq 15$, with the ML coefficients C_N^0 determined from (14). The results averaged over all 20 datasets, are displayed in Fig 2. As can be seen, the free energy decreases nearly monotonically with N. Note that $N = 0$ corresponds to the Gaussian approximation for $P_0(\Delta U)$, and is equivalent to the second–order free energy perturbation theory.

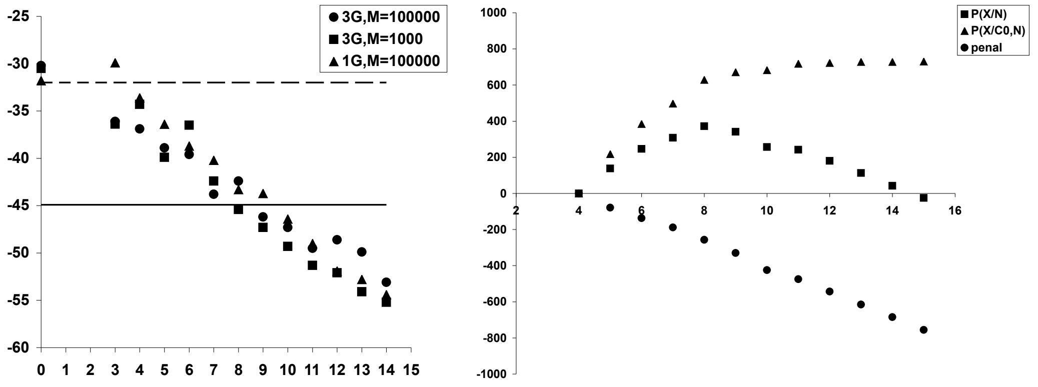

FIGURE 2. Left panel: the ML free energies calculated from (6) and averaged over 20 datasets as functions of the number of terms, N in the expansion. The results for 100,000 and 1,000 values of x for $P_0(\Delta U)$ described by (31) and for a single Gaussian are denoted by diamonds, squares and triangles, respectively. The solid and dashed horizontal lines represent the exact free energies for the two functions used. Right panel: a typical result for $\ln P(X \mid N)$ (triangles) $\ln P\left(X \mid C_N^0, N\right)$ (squares) and the "Ockham penalty" (diamonds), calculated from (29), as functions of N.

Next, we calculated $\ln P(X \mid N)$ from (29) for each dataset. Its typical behavior is shown in the right panel of Fig 2. It increases for small N, passes through a maximum and then slowly decreases with N. From this dependence we identified the ML values of N and determined the corresponding free energies. These energies were averaged over 20 datasets and the root mean square deviation (RMSD) was calculated. The results are collected in Table 1. The free energies obtained directly from (5) poorly reproduce the correct values of ΔA, as might be expected from Fig. 1. Also as expected, the second–order (Gaussian) approximation is very good for the purely Gaussian $P_0(\Delta U)$, but not for the asymmetric $P_0(\Delta U)$. In contrast, the ML estimate of ΔA for this $P_0(\Delta U)$ approximates the exact free energy very well. For the large sample, the average ML value of N is 7.2 with a small variance. For the small sample, ΔA is overestimated because N is consistently slightly smaller than that for the large sample, presumably because the small dataset contains less information. For the purely Gaussian case, the ML solution for N fluctuates markedly around 4.2 and the estimated ΔA matches the exact value poorer that the second–order formula.

Table 1. Free energies calculated using different approaches.

system/ sample size	Exact	Eq.(5)	Gaussian approximation	ML	RMSD (ML)
3G/100,000	-44.9	-20.5	-30.2	-43.7	1.2
3G/1,000	-44.9	-19.8	-30.5	-41.8	2.6
1G/100,000	-32.0	-23.1	-31.8	-34.3	2.1

Instead of using the ML value of N, one can terminate the series in (6) when the statistical error on $\varphi_n(x)$ becomes larger than its average over the sample. Interestingly, this heuristic criterion yields similar results as a better justified ML criterion.

CONCLUSIONS

We have shown that modeling probability densities of ΔU as a series well suited to describe Gaussian–like distributions, combined with a ML approach to determining the number of terms and the coefficients of the expansion, yields markedly improved estimates of free energy differences between two states of a system. The improvement is particularly evident in the most difficult cases when $P_0(\Delta U)$ is broad and skewed, which means that the two states are fairly dissimilar. In such cases, the proposed method is a highly promising alternative to more expensive strategies of stratification and importance sampling. The reduced cost makes the method particularly suitable, for example, for computer aided drug design, in which the goal is to screen rapidly a large number of potential drugs for binding with their protein target.

The modeling approach also represents a conceptual departure from the traditional view that free energy differences can be reliably estimated only if configurations from the low–ΔU tail of $P_0(\Delta U)$ are adequately sampled. Instead, it is proposed that information contained in the well sampled part of $P_0(\Delta U)$ might be sufficient to calculate free energies, at least in the absence of persistent quasi non–ergodicities.

REFERENCES

1. C. Chipot and A. Pohorille (Eds.), Free energy calculations. Theory and application in chemistry and biology. Springer, to be published (2006).
2. G. Hummer, L. R. Pratt and A. E. Garcia, Multistate Gaussian model for electrostatic solvation free energies. J. Am. Chem. Soc. **119**, 8523Ð8527 (1997).
3. S. T. Bramwell, K. Christensen, J. Y. Fortin, P. C. W. Holdsworth, H. J. Jensen, S. Lise, J. M. Lopez, M. Nicodemi, J. F. Pinton, and M. Sellitto. Universal fluctuations in correlated systems. Phys. Rev. Lett. **84**, 3744Ð3747 (2000).
4. G. Szego, Orthogonal polynomials. 4th Edition, American Mathematical Society: Providence, 1975.