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

Andrew Pohorille [*] and Eric Darve [†]

[*]NASA Ames Research Center,
Exobiology Branch, MS 239–4
Moffett Field, California 94035–1000, USA
[†]Department of Mechanical Engineering,
Stanford University
Stanford, California 94025, 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, which makes it applicable to challenging problems, such as protein-drug interactions.

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

INTRODUCTION

To understand and control 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.

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. Thus, calculating A is equivalent to estimating Q, which is usually very difficult. In practice, however, we are interested in free energy *differences*, ΔA, between two systems, say 0 and 1, which can be expressed as [1]:

$$\Delta A = -\beta^{-1} \ln \frac{\int \exp\left[-\beta U_1\left(\mathbf{x}\right)\right] d\mathbf{x}}{\int \exp\left[-\beta U_0\left(\mathbf{x}\right)\right] d\mathbf{x}} = -\beta^{-1} \ln \langle \exp\left\{-\beta\left(\Delta U\right)\right\} \rangle_0 \tag{2}$$

Here, $U_0(\mathbf{x})$, and $U_1(\mathbf{x})$, are potential energies for systems 0 and 1, respectively, $\Delta U = U_1(\mathbf{x}) - U_0(\mathbf{x})$ 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:

$$\Delta A = -\beta^{-1} \ln \int \exp\left(-\beta \Delta U\right)\ P_0(\Delta U)\ d\Delta U \tag{3}$$

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 $\exp\left(-\beta \Delta U\right)$. 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.

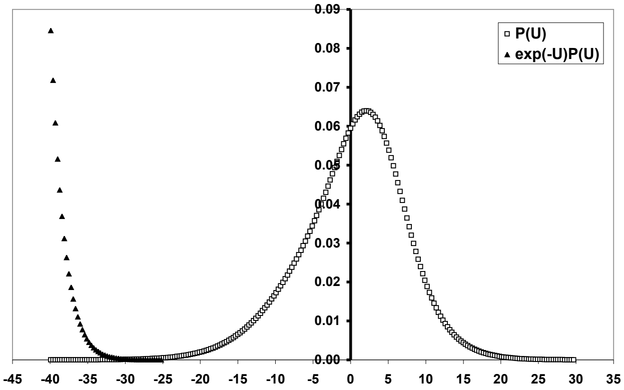

FIGURE 1. $P_0(\Delta U)$ (open squares) and the integrand in equation (3), $\exp\left(-\beta \Delta U\right) P_0(\Delta U)$ (triangles). Only the right side of the integrand is sampled, which precludes accurate estimation of the integral.

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 this region. 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 suitable for a broad class of 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

$$P_0(\Delta U) = \left(\sum_{n=0}^{\infty} c_n \varphi_n \left(\Delta U \right) \right)^2 \tag{4}$$

where c_n is the n–th coefficients of the expansion and ϕ_n is the n–th normalized Gram–Charlier polynomial and related to the n–th Hermite polynomial by:

$$\varphi_n \left(x \right) = \frac{1}{\sqrt{2^n \pi^{1/2} n!}} H_n \left(x \right) \exp \left(-x^2/2 \right) \tag{5}$$

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

$$\sum_n c_n^2 = 1 \tag{6}$$

In practice only the first few coefficients can be determined with sufficient accuracy. This means that (4), or any other expansion, is useful only if it converges quickly. This raises a question: how to determine the optimal order of the expansion, N, and the coefficients $\{c_n\}$, $n \leq N$ in (4)? 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.

We 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 (4) 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 (4) truncated after the first $N+1$ terms

$$P(N \mid X) = \frac{P(X \mid N) P(N)}{P(X)}. \tag{7}$$

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

$$P(X \mid N) = \int P(X, C_N \mid N) dC_N = \int P(X \mid C_N, N) P(C_N \mid N) dC_N \tag{8}$$

where dC_N stands for $dc_0 \dots dc_N$ 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 (6) using Lagrange multipliers. We first note that for statistically independent samples

$$P\left(X \mid C_N, N\right) = \prod_{\mu=1}^{M} P\left(x_\mu \mid C_N, N\right) \tag{9}$$

where $P\left(x_\mu \mid C_N, N\right)$ 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 (4), the function to be minimized is:

$$f(C,N) = 2\sum_\mu [\ln \sum_n c_n \varphi_n(x_\mu)] + \lambda \sum_n c_n^2. \tag{10}$$

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 $\{c_n\}$

$$\sum_\mu \frac{\varphi_m\left(x_\mu\right)}{\sum_n c_n \varphi_n\left(x_\mu\right)} + \lambda c_m = 0 \tag{11}$$

which are solved simultaneously with (6).

The value of λ can be found to be:

$$\lambda = \lambda \sum_m c_m^2 = -\lambda \sum_\mu \frac{\sum_m c_m \varphi_m\left(x_\mu\right)}{\sum_n c_n \varphi_n\left(x_\mu\right)} = -M \tag{12}$$

Equation (11) has a simple interpretation. If we apply the relation

$$\frac{1}{M}\sum_\mu f(x_\mu) \approx \int f(x)P(x)dx. \tag{13}$$

for a discrete sample of a function $f(x)$ to the sum on the left hand side of (11) and take advantage of orthonormality of φ_n we obtain

$$\sum_n \left[\frac{1}{M}\sum_\mu \frac{\varphi_m(x_\mu)\varphi_n(x_\mu)}{(\sum_p c_p \varphi_p(x_\mu))^2}\right] c_n = \sum_n c_n \int \varphi_m(x)\varphi_n(x)dx = c_m. \tag{14}$$

This means that (11) are $N+1$ equations that enforce orthonormality conditions on φ_n sampled at $\{x_\mu\}$.

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:

$$P\left(X \mid C_N, N\right) = \exp\left[\ln P\left(X \mid C_N, N\right)\right] \tag{15}$$

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

$$\ln P(X \mid C_N, N) = \ln P(X \mid C_N^0, N) + 2\sum_{k=1}^{\infty}(-1)^{k+1}\frac{1}{k}\sum_{\mu}(S_\mu)^k \tag{16}$$

where

$$S_\mu = \frac{\sum_m \Delta c_m \varphi_m(x_\mu)}{\sum_n c_n^0 \varphi_n(x_\mu)} \tag{17}$$

and $\Delta c_n = c_n - c_n^0$. If we truncate the expansion in (16) after second–order

$$P(X \mid C_N, N) = P\left(X \mid C_N^0, N\right)\exp\left(2\sum_\mu S_\mu - \sum_\mu S_\mu^2\right). \tag{18}$$

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:

$$2\sum_\mu S_\mu = 2\sum_m \Delta c_m \sum_\mu \frac{\varphi_m(x_\mu)}{\sum_n c_n^0 \varphi_n(x_\mu)} = 2M\sum_m \Delta c_m c_m^0 = -M\sum_m \Delta c_m^2. \tag{19}$$

In the second equality we used (11), 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:

$$2\sum_\mu S_\mu = -\Delta C^T \mathbf{M} \Delta C \tag{20}$$

where ΔC is a $N+1$ dimensional vector with the coefficients Δc_n, ΔC^T is its transpose and $\mathbf{M}$ is an $(N+1)\times(N+1)$ matrix, whose entries are $M\delta_{mn}$.

We can proceed similarly with the second-order term. Using (17) we obtain:

$$\sum_\mu S_\mu^2 = \Delta C^T \mathbf{A} \Delta C \tag{21}$$

where $\mathbf{A}$ is a $(N+1)\times(N+1)$ matrix, whose entries are:

$$A_{nm} = \sum_\mu \frac{\varphi_n(x_\mu)\varphi_m(x_\mu)}{\left[\sum_n c_n^0 \varphi_n(x_\mu)\right]^2}. \tag{22}$$

After substituting (20) and (21) to (18) and defining $\mathbf{\Lambda} = \mathbf{A} + \mathbf{M}$, we obtain:

$$P(X \mid C_N, N) = P\left(X \mid C_N^0, N\right)\exp\left(-\Delta C^T \mathbf{\Lambda} \Delta C\right) \tag{23}$$

which we substitute to (8) to obtain

$$P(X \mid N) = 2P\left(X \mid C_N^0, N\right)\int \exp\left(-\Delta C^T \mathbf{\Lambda} \Delta C\right) P(C_N \mid N) dC_N \tag{24}$$

where the extra factor of 2 comes from the fact that our definition of c_n (see Eq.(4)) admits solutions C_N^0 and $-C_N^0$.

We take the prior, $P(C_N \mid N)$, to be uniform, subject to the constraint (6). This means that it is uniform on a N-dimensional unit hypersphere and is zero otherwise.

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

To calculate this integral, let's first observe that the matrix $\mathbf{A}$ is such that:

$$A_{nm} = \sum_\mu \frac{\varphi_n(x_\mu)\,\varphi_m(x_\mu)}{\left[\sum_q c_q^0 \varphi_q(x_\mu)\right]^2} \approx M \int \varphi_n(x)\varphi_m(x)dx = M\delta_{nm} \tag{26}$$

Therefore, $\mathbf{\Lambda} = \mathbf{M} + \mathbf{A}$ is close to $2M\mathbf{I}$. In practice M is very large and $\exp\left(-\Delta C^T \mathbf{\Lambda} \Delta C\right)$ is sharply peaked around C_N^0. To a good approximation, we can replace the integral over the sphere by an integral over the plane tangent to the sphere at C_N^0. This allows calculating the integral analytically. This can be done by defining the rotation matrix $\mathbf{R}$ such that $\mathbf{R}C_N^0 = (1, 0, \cdots, 0)^T$. Then, the plane tangent at C_N^0 is mapped onto the plane tangent at $(1, 0, \cdots, 0)$. Define $N \times N$ matrix $\mathbf{\Lambda^r}$ as:

$$[\mathbf{\Lambda}^r]_{ij} = [\mathbf{R}\mathbf{\Lambda}\mathbf{R}^T]_{i+1,j+1} \tag{27}$$

We now change basis using the rotation $\mathbf{R}$:

$$\Delta C^T \mathbf{\Lambda} \Delta C = (\mathbf{R}\Delta C)^T \mathbf{R}\mathbf{\Lambda}\mathbf{R}^T (\mathbf{R}\Delta C) \tag{28}$$

Using standard techniques to calculate multivariate Gaussian integrals, we obtain:

$$\int \exp\left(-\Delta C^T \mathbf{\Lambda} \Delta C\right) dC_N = \int \exp(-Z^T \mathbf{\Lambda}^r Z)\, dz_1 \cdots dz_N = \sqrt{\frac{\pi^N}{|\det \mathbf{\Lambda}^r|}} \tag{29}$$

The determinant of $\mathbf{\Lambda}^r$ can be simply related to $\mathbf{\Lambda}$. Observe first that C_N^0 is an eigenvector of $\mathbf{\Lambda}$ with eigenvalue $2M$:

$$[\mathbf{A}C_N^0]_n = \sum_\mu \frac{\varphi_n(x_\mu) \sum_m c_n^0 \varphi_m(x_\mu)}{\left[\sum_q c_q^0 \varphi_q(x_\mu)\right]^2} = \sum_\mu \frac{\varphi_n(x_\mu)}{\sum_q c_q^0 \varphi_q(x_\mu)} = M c_n^0 \tag{30}$$

This implies that:

$$\mathbf{R}\mathbf{\Lambda}\mathbf{R}^T (1, 0, \cdots, 0)^T = (2M, 0, \cdots, 0)^T \tag{31}$$

The matrix $\mathbf{R}\mathbf{\Lambda}\mathbf{R}^T$ is therefore of the form:

$$\mathbf{R}\mathbf{\Lambda}\mathbf{R}^T = \begin{bmatrix} 2M & * \\ 0 & \mathbf{\Lambda}^r \end{bmatrix} \tag{32}$$

Consequently:

$$\det(\mathbf{\Lambda}) = 2M \det(\mathbf{\Lambda}^r) \tag{33}$$

We finally obtain a very simple approximation for our Gaussian integral over a sphere:

$$\int \exp\left(-\Delta C^T \mathbf{\Lambda} \Delta C\right) dC_N = \sqrt{\frac{2M\pi^N}{|\det \mathbf{\Lambda}|}} \tag{34}$$

This yields the final expression for $P(X \mid N)$:

$$\ln P(X \mid N) = \ln P\left(X \mid C_N^0, N\right) - \frac{1}{2}\left(\ln|\det \mathbf{\Lambda}| - N\ln\pi - \ln 8M\right) \tag{35}$$

Note that $\ln|\det \mathbf{\Lambda}| \approx (N+1)\ln 2M$ so that the term in parenthesis is positive. This is as expected; the solution consists of two terms which change oppositely with N. The first term, which is the optimal (ML) solution, 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 (35) 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, $p_i\,(\Delta U)$. The mean values, $\langle \Delta U \rangle_i$, variances, σ_i, and weights w_i of each Gaussian were: (3.0, 4.0, 0.3), (0.0, 7,0, 0.5) and (-3.0, 9.0, 0.2) The resulting $P_0(\Delta U)$ is shown in Fig. 1. The main advantages of a multi-Gaussian $P_0(\Delta U)$ are that it can be easily sampled and the free energy, ΔA, can be calculated exactly.

For this system we generated 20 datasets of 100,000 statistically independent values of x. We also generated a dataset of 2,000,000 values of x by combining the previous datasets. For each dataset, we calculated the free energy from (3) and from the expansion (4) for $0 \leq N \leq 20$, with the ML coefficients C_N^0 determined from (11). 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 $N = 0$ is the Gaussian approximation for $P_0(\Delta U)$, equivalent to the second-order free energy perturbation theory.

Next, we calculated $\ln P(X \mid N)$ from (35) 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, which is between 9 and 11 for different datasets, and determined the corresponding free energies. Averaged over all 20 datasets, the free energy is -40.4 ± 0.4, which is close to the exact value of -41.9. In contrast, the free energies obtained directly from (3) and from the second-order (Gaussian) approximation reproduce the correct values of ΔA rather poorly . The average free energies in these two approaches are -28.3 ± 0.6 and -24.6 ± 0.5, respectively. For the sample of 2,000,000, the value of N increases to 13, because this dataset contains more information. The free energy in this case is -44.9. Numerical tests indicate that the second order approximation in (18) and the approximation to $\det \mathbf{\Lambda}$ from the end of the previous section are both accurate.

In addition, we generated datasets of 100,000 values of x sampled from a Gaussian with the mean zero and $\sigma = 8$. The ML solution for N was always zero (pure Gaussian).

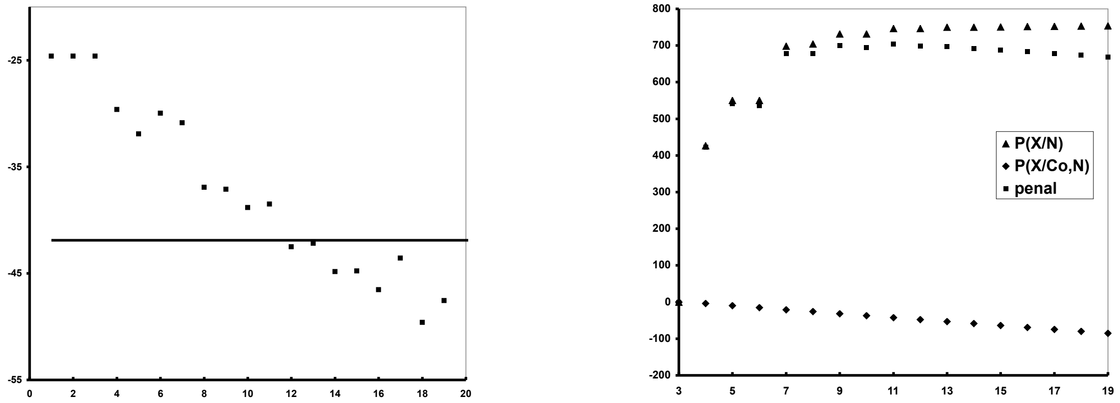

FIGURE 2. Left panel: the ML free energies calculated from (4) and averaged over 20 datasets as functions of the number of terms, N in the expansion. The solid horizontal line represents the exact free energy. Right panel: a typical result for $\ln P(X \mid N)$ (triangles) $\ln P\left(X \mid C_N^0, N\right)$ (squares) and the “Ockham penalty" (diamonds), calculated from (35), as functions of N.

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 difficult cases when $P_0(\Delta U)$ is broad and skewed, which means that the two states are fairly dissimilar. In such cases, the method is a promising alternative to more expensive strategies of stratification and importance sampling.

In the future, we will systematically investigate how the quality of the approximation depends on the shape of the distribution and sample size, and apply our method to data generated in molecular dynamics simulations of chemical and biological systems.

REFERENCES

1. C. Chipot and A. Pohorille (Eds.), Free energy calculations. Theory and application in chemistry and biology. Springer, 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.