A BAYESIAN APPROACH TO CALCULATING FREE ENERGIES OF CHEMICAL AND BIOLOGICAL SYSTEMS

Andrew Pohorille
NASA-Ames Research Center, USA
(e-mail: pohorill@max.arc.nasa.gov, fax: 650-604-1088)

Abstract

A common objective of molecular simulations in chemistry and biology is to calculate the free energy difference, ΔA, between states of a system of interest. Important examples are protein-drug interactions, protein folding and ionization states of chemical groups. However, accurate determination of ΔA from simulations is not simple. This can be seen by representing ΔA in terms of a one-dimensional integral of $exp(-\Delta E/k_BT) \times P(\Delta E)$ over ΔE. In this expression, ΔE is the energy difference between two states of the system, $P(\Delta E)$ is the probability distribution of ΔE, k_B is the Boltzmann constant and T is temperature. For finite systems, $P(\Delta E)$ is a distorted Gaussian. Note that the exponential factor weights heavily the low ΔE tail of $P(\Delta E)$, which is usually known with low statistical precision.

One way to improve estimates of ΔA is to model $P(\Delta E)$. Generally, this approach is rarely successful. Here, however, we take advantage of the "Gaussian-like" shape of $P(\Delta E)$. As is known in physics, such a function can be conveniently represented by the square of a "wave function" which is a linear combination of Gram-Charlier polynomials. The number of terms, N, in this expansion supported by the data must be determined separately. This is done by calculating the posterior probability, $P(N/\mathbf{\Delta E})$, where $\mathbf{\Delta E}$ stands for all sampled values of ΔE. In brief, the dependence of the likelihood function on the coefficients of the expansion, $\mathbf{C_N}$ is marginalized by determining their optimal values using Lagrange multipliers, and then expanding $P(\mathbf{\Delta E})/\mathbf{C_N}, \mathbf{N})$ around the optimal solution. Special care needs to be taken to ensure convergence of this expansion. As expected, the maximum likelihood solution consists of two terms. One is related to the optimal values of $\mathbf{C_N}$ and always increases with N. The second term is an "Ockham's Razor" penalty. It involves a multivariate Gaussian integral on the N-dimensional hypersphere, which arises due to mormalization. This integral cannot be calculated analytically, but accurate approximations, which properly account for problem symmetries, can be obtained.

The method offers the largest improvements over conventional approaches when $P(\Delta E)$ is broad and sample size is relatively small. This makes is particularly suitable for computer aided drug design, in which the goal is to screen rapidly a large number of potential drugs for binding with the protein target.