Locally enhanced sampling in free energy calculations: Application of mean field approximation to accurate calculation of free energy differences

Gennady Verkhivker, Ron Elber, and Wieslaw Nowak

University of Illinois at Chicago, Department of Chemistry, M/C 111, P.O. Box 4348, Chicago, Illinois 60680

(Received 10 July 1992; accepted 10 September 1992)

Computational methods such as the free energy perturbation and the thermodynamic integration\(^1\) attracted considerable attention during the last few years. These simulation techniques provided a link between structural information, microscopic potentials and macroscopic thermodynamic measurements. Especially popular are the applications to biological molecules, such as proteins and DNA. Nevertheless, the routine use of these methodologies must be always weighted against the considerable computer time associated with their applications. Hundreds to thousands of Cray YMP hours are required to obtain converged results for a mutation of a single residue in a protein. Convergence is assumed when backward and forward calculations yield similar results and the relevant torsional space was sampled effectively. This means that in pursuing this type of calculation the user time is at least several months, leaving little room for checking the sensitivity of the results to the simulation parameters. A method to enhance the efficiency of these calculations, while maintaining their accuracy is very much to be desired.

In this Communication we restrict the discussion to heterogeneous systems and to free energy differences of small and local changes. An example is a mutation of a single residue in a protein. Obviously a protein is not homogeneous and the change described is local. The effects of single point mutations on protein activity and stability are of wide biochemical interest and therefore the methodology to be discussed is worth pursuing. Under these restrictions a more accurate simulation technique is proposed that provides considerably better sampling of configuration space and better convergence properties. This is at a computational cost comparable to the usual applications.

In the regular application of thermodynamic integration (TI) we consider two thermodynamic states that are described by different Hamiltonians. In the computational example, one thermodynamic state is of the protein lupine leghemoglobin with phenylalanine (phe) at the 29th position (native leghemoglobin) and the second thermodynamic state is of lupine leghemoglobin with tryptophan (trp) at the same position. Leghemoglobin is a small plant protein that binds oxygen very strongly. The phenylalanine at the 29th position was suggested as the “door” for the ligand escape from the binding pocket.\(^3\) The above (hypothetical) mutation is proposed as means to increase the barrier for diffusion. Here we consider the effect of the mutation on the stability of the folded state of the protein.

We do not consider the effect of the mutation on the unfolded state. Estimate for the later and the effect on the barrier will be discussed in future work.\(^4\)

The Hamiltonians for the native protein and the mutant are called \(H_{\text{phe}}\) and \(H_{\text{trp}}\), respectively. In the TI we define a Hamiltonian that interpolates between the two states—\(H_\lambda = (1 - \lambda)H_{\text{phe}} + \lambda H_{\text{trp}}\), where \(\lambda\) is varied from 0 to 1. The free energy difference—\(\Delta F_{\lambda} = \Delta F_{\lambda,\lambda + \Delta \lambda}\)—between the states described by the Hamiltonians \(H_\lambda\) and \(H_{\lambda + \Delta \lambda}\) is given by \(\Delta F_{\lambda} = \langle H_\lambda - H_{\lambda + \Delta \lambda} \rangle_{H_\lambda}\), where the \(\langle \cdots \rangle_{H_\lambda}\) denotes thermal average with respect to the Hamiltonian \(H_\lambda\). The total free energy difference is obtained by the summation of all the \(\Delta F_{\lambda,\lambda + \Delta \lambda}\).

The thermal averaging in heterogeneous systems is usually performed with a time sequence obtained from a molecular dynamics trajectory. To compute a single molecular dynamics step (and a single energy difference to be used in the average) the forces between all the particles in the system are calculated. The resources required for this calculation are at least linear with the number of particles. In our model, leghemoglobin includes 1558 atoms.

The application of the locally enhanced sampling (LES) in TI is outlined schematically in Fig. 1. We replace the usual thermodynamic path (A–D) by the alternative route A–B–C–D. As is explained below in the B–C part we “win” and the overhead is the two (short) paths: A–B and C–D. Let us discuss these two paths first. Using the LES methodology we split a single “real” phe to L phe fragments and a single “ghost” trp to L trp ghost pieces. The fragments do not see each other and the protein feels the average force of all the pieces (see below for mathematical details). Similar procedure is applied in the last step (path segment C–D) in which L trp fragments are shrunk to a single full sized trp while L phe ghosts are merged to one.

The path segment B–C modifies linearly the L phe to L trp. The path segment B–C is comparable to the original path (path A–D) except that now it is done for L copies of trp and L copies of phe. The existence of the multiple copies is a major advantage of the proposed protocol since considerably more sampling for possible side chain conformations is obtained.

Let us now switch to a more formal discussion. The multiple copies of side chains originate from a mean field approximation. Note however that the present formalism is exact. Let \(\rho_\lambda(T, \Gamma)\) be the probability density of velocities and coordinates in the system (\(\Gamma\) denotes both coordinates

and velocities). In the proposed mean field approximation $\rho$ is set to a Hartree product. Going back to the leghemoglobin example, we set $\rho \approx \rho_{\text{prot}}\rho_{\text{phe}}\rho_{\text{trp}}$ where $\rho_{\text{prot}}$ is the density of the rest of the protein. The density is translated to equations of motion by a specific choice of the different densities,

$$\rho_{\text{prot}} = \delta[\Gamma_{\text{prot}} - \Gamma^0_{\text{prot}}(t)]$$

and

$$\rho_{\text{phe/trp}} = \sum_{i=1}^{L} w_i \delta[\Gamma_{\text{phe/trp}} - \Gamma^0_{\text{phe/trp}}(t)],$$

where $\rho_{\text{phe/trp}}$ denotes either the phenylalanine (phe) density or the tryptophan (trp) density. The $w_i$ are the weights of the different fragments. They are equal to 1/L in the B-C path and are changed linearly to zero or to one in the A-B and the C-D paths. The Hartree tree describes a mean field system in which a single protein is moving in the average field of $L$ and $L'$ (enhanced) residues. The mean field energy associated with this density is

$$E^\text{LES}_A = E_{\text{prot}}(\Gamma^0_{\text{prot}}) + (1 - \lambda) \sum_{i=1}^{L} w_i E^\text{pest}_{\text{phe}}(\Gamma^0_{\text{phe}}, \Gamma^0_{\text{prot}}) + \lambda \sum_{j=1}^{L'} w_j E^\text{pest}_{\text{trp}}(\Gamma^0_{\text{trp}}, \Gamma^0_{\text{prot}}).$$

This means field energy is used in the equations of motion, as well as in the calculations of data points $\Delta E^\text{LES}_{A, A' + b}$

$$\langle E^\text{LES}_A - E^\text{LES}_{A + b} \rangle_{b, A}.$$ The existence of a distribution of side chains (in the form of $L$ copies) provides self averaging in a single dynamics step and therefore the convergence is faster. Furthermore, another important point in free energy calculation is to explore as extensively as possible the conformational space. As was shown in Ref. 5, the barriers for conformational transitions are reduced in the mean field energy. It is therefore possible to explore better the configuration space in LES as compared to the original system. In Fig. 2 we compare the sampling of $\chi_2$ of trp during a trajectory employed in a single window. Clearly LES is much more effective in exploring the $\chi_2$ space (similar results were found for the phe). In fact, the low error bars in a free energy difference calculated using a single copy procedure are artificial, since the distribution of $\chi_2$ that was actually explored in 20ps is very limited. The more extensive exploration of the configurational space in LES comes at the expense of moderately larger error bars. The quality of the results for comparable computational effort is considerably better for the LES than for the direct calculations.

Hence, LES is advantageous with respect to the usual applications in two aspects: First it provides more statistics and second it reduces the barriers between alternate conformations and therefore reduces the risk of being trapped in a small fraction of the available configuration space. The same protocol is also used in the annihilation and
the creation of the extra fragments, except that the $\lambda$ parameter now corresponds to the relative weight of the different fragments. For the shrinking calculation (converting $L$ fragments to only one residue) one of the copies (it can be any of the $L$) increases in size and the rest of the copies decrease in size. A point of concern is that A–B (and C–D) are associated with the creation and the annihilation of particles, a type of calculation that is known to be problematic. Note first that the process considered does not require the formation of a new cavity, since a fully grown side chain was already there. We consider the splitting of this side chain to $L$ fragments that occupy approximately the same space. Furthermore, the different copies are allowed to overlap with each other which minimized their effective volume. The calculation is more similar to cavity growth than to cavity formation. In the first step $w_i$ was varied from $1/L$ to 0.96 and the other $w_i$ to 0.04/$L - 1$. This change can be calculated in the backward and the forward directions and therefore can be tested in a reasonable way. In the second step we modified $w_i$ to 1.0 and the other $w_i$ to zero. Of course in the second step we calculate the free energy difference only in the backward direction (eliminating the excess particles). The creation of particles in reverse is a common practice in the insertion of new particles. The contribution of the last segment was small ($\approx 0.1$ kcal/mol) and was further tested by repeating the last calculation using two $\lambda$ steps. The low error bars in the A–B and C–D paths are tabulated in Fig. 1. Another comment concerning the annihilation of particles is that their interactions are scaled to zero linearly. However, we maintained bond constraints on the "ghost" residues. The contribution of these constraints to the free energy was subsequently calculated using the method of van Gunsteren and was found to be very small.

In the leghemoglobin study we set $L = L' = 5$. The thermodynamic averages were calculated using the program MOIL. The empirical force field in MOIL is mostly the combination of AMBER and OPLS. The only difference from Ref. 9 was that the improper torsion energy function and parameters were taken from the CHARMM paper. The dielectric constant was 1 and the scaling factors for the van der Waals and the electrostatic 1–4 interactions were 8 and 2, respectively. The cutoff distance was 9 Å. The long range forces were smoothly truncated using a shift function. The bonds for the phe and trp parts were kept fixed using the SHAKE algorithm.

To compare the LES to the regular simulation we pursued first a straightforward TI calculation employing 20 equally spaced windows interpolating between a single phe and a single trp. For each window we used 10 ps for equilibration and 20 ps for data collection at 300 K. In addition to TI we also performed thermodynamic perturbation calculation with similar quality of results. The errors were estimated as the difference between the backward and the forward calculations (estimates based on the variance were found to be absurdly small). We calculated independent trajectories in the backward and the forward directions to estimate the same free energy—$\Delta F^f = \Delta F_{\lambda \rightarrow \lambda + 1}$ and $\Delta F^b = \Delta F_{\lambda + \Delta \lambda \rightarrow \lambda}$. The error was estimated as $(\Delta F^b + \Delta F^f)/2$.

Our first trial of calculating directly path A–D with 20 windows did not converge. The total free energy difference in the forward direction (from phe to trp) was 1.9 kcal/mol and in the backward direction was 5.8 kcal/mol, giving an average value (forward direction) of $-1.95$ kcal/mol and error bars of $\pm 3.8$ kcal/mol. Only with significantly more effort that included human identification of problematic windows (that were primarily near $\lambda \approx 0$ and near $\lambda \approx 1$) and the addition of more intermediate $\lambda$ states it was possible to obtain converged results. The results obtained with 35 windows are reported in Fig. 1. In the converged calculation we obtain a free energy difference from phe to trp of 4.7 kcal/mol, from trp to phe of $-3.1$ kcal/mol and error bars of $\pm 0.83$ kcal/mol. These error bars should not be taken too seriously since according to Fig. 2 the conformational space sampled in the single copy calculation is too restricted.

In contrast to the usual calculation (path A–D) the LES calculation did converge in the first trial using only 24 windows to 5.1 ± 1.5 kcal/mol. The errors of the different paths: A–B B–C C–D, are assumed independent. Therefore the total error is estimated as the square root of the sum of the individual squares of the errors. The LES included 20 windows for the B–C path and two windows for each of the paths A–B and C–D. Since calculation of a single window in the LES and in the single copy protocol requires essentially the same computational resources, LES was significantly less expensive. We then refined the LES calculation using 24 windows for the B–C part (total of 28 windows) to obtain the value of 5.25 ± 1.07 kcal/mol. The values and the error bars for the refined calculations are reported in Fig. 1. The somewhat larger error bars in the LES calculations are more realistic than the single copy calculation since as is demonstrated in Fig. 2 the conformational space was explored much better in LES and specifically the sampling of the $\chi^2$ space in LES was essentially complete.

To summarize: With significantly less computational effort the LES protocol gave us results of comparable accuracy and significantly enhanced sampling as compared to straightforward single copy calculations.

This research was supported by NIH Grant GM41905.

---

*Present address: Institute of Physics, N. Copernicus University, ul. Grudziadzka 5, PL-87-100 Torun, Poland.


5. W. F. van Gunsteren, in Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications, edited by W. F. van-