Improving the Efficiency and Reliability of Free Energy Perturbation Calculations Using Overlap Sampling Methods

NANDOU LU,[1] DAVID A. KOFKE,[2] THOMAS B. WOOLF[1,3]
[1]*Department of Physiology, School of Medicine, Johns Hopkins University, 725 North Wolfe Street, 206 Biophysics Building, Baltimore, Maryland 21205*
[2]*Department of Chemical Engineering, University at Buffalo, The State University of New York, Buffalo, New York 14260*
[3]*Department of Biophysics and Biophysical Chemistry, School of Medicine, John Hopkins University, Baltimore, Maryland 21205*

Received 26 June 2003; Accepted 4 August 2003

Abstract: A challenge in free energy calculation for complex molecular systems by computer simulation is to obtain a reliable estimate within feasible computational time. In this study, we suggest an answer to this challenge by exploring a simple method, *overlap sampling* (OS), for producing reliable free-energy results in an efficient way. The formalism of the OS method is based on ensuring sampling of important overlapping phase space during perturbation calculations. This technique samples both forward and reverse free energy perturbation (FEP) to improve the free-energy calculation. It considers the asymmetry of the FEP calculation and features an ability to optimize both the precision and the accuracy of the measurement without affecting the simulation process itself. The OS method is tested at two optimization levels: no optimization (simple OS), and full optimization (equivalent to Bennett's method), and compared to conventional FEP techniques, including the widely used direct FEP averaging method, on three alchemical mutation systems: (a) an anion transformation in water solution, (b) mutation between methanol and ethane, and (c) alchemical change of an adenosine molecule. It is consistently shown that the reliability of free-energy estimates can be greatly improved using the OS techniques at both optimization levels, while the performance of Bennett's method is particularly striking. In addition, the efficiency of a calculation can be significantly improved because the method is able to (a) converge to the right answer quickly, and (b) work for large perturbations. The basic two-stage OS method can be extended to admit additional stages, if needed. We suggest that the OS method can be used as a general perturbation technique for computing free energy differences in molecular simulations.

 J Comput Chem 25: 28–39, 2004

Key words: free energy simulation; perturbation calculation; overlap sampling; alchemical transformation

Introduction

The free energy is an important thermodynamic quantity for physical, chemical, and biological systems, and the computation of free energy has for a long time been one of the central tasks of molecular simulation.[1–4] As the calculation is a computationally intensive and time-demanding process, obtaining a reliable free energy estimate within feasible computational time usually imposes difficulties in a computer simulation. A reliable *and* efficient free energy calculation method is thus desired. Even with modern computers, an effective calculation algorithm is crucial for computer simulation as a practical tool for free energy studies of large, complex, and time-demanding systems, such as in all atom protein–ligand systems. In fact, free energy calculations are regarded as one of the greatest challenges for molecular simulation.[5]

Although many free energy calculation methods have been developed over decades, the free energy perturbation (FEP) technique and its extensions are popular choices for computing relative

Correspondence to: N. D. Lu; e-mail: nlu@groucho.med.jhmi.edu

Contract/grant sponsor: American Cancer Society; contract/grant number: ACS-RSG-01-048-01-GMC (to T.B.W.)

Contract/grant sponsor: U.S. Department of Energy, Office of Basic Energy Services; contract/grant number: DE-FG02-96ER14677 (to D.A.K.)

free energy in molecular simulations.[2,6] The basic FEP formula computes the free-energy difference between two systems according to[7]

$$\exp(-\beta\Delta A) = \langle \exp(-\beta u) \rangle_0, \quad (1)$$

where, $\Delta A = A_1 - A_0$ is the Helmholtz free-energy difference between systems 0 and 1, $\beta = 1/kT$ is the reciprocal temperature, with k being Boltzmann's constant, and T is the absolute temperature. $u = U_1 - U_0$ is the perturbation with U_0 and U_1 the Hamiltonians for systems 0 and 1, respectively (note the perturbation can be of quantities other than the Hamiltonian); the angle bracket represents a canonical ensemble average over the system specified by the subscript. For an isobaric and isothermal ensemble, eq. (1) gives the Gibbs free-energy difference. The system governing the simulation is often called the *reference* [i.e., system 0 for eq. (1)] and the other system is the *target.*

Understanding the source of error in a simulation is important for developing effective free-energy calculation techniques. The error in a free-energy calculation is due to the many numerical approximations in a computer simulation and can be seen as originating from two sources: the representation of the system and the sampling. Before further discussion on the source of error, we note that the reliability of a calculation has two distinct facets, representing both the *precision* and the *accuracy,* respectively. The precision is related to the random error of the calculation that describes the reproducibility of results; the accuracy is related to the systematic error, indicating the correctness of the measurement. The free-energy calculation shows both systematic and random errors caused by these different sources.

The approximations in the representation of a system (e.g., empirically derived potential model, small system size, etc.) on the computer will cause systematic errors in free-energy calculations. Such systematic errors will not be reduced by increasing the simulation length or by adopting a better free-energy calculation algorithm. Imperfect sampling contributes another major source of error in free-energy calculations.[8–14] Although statistical mechanics provides a formally exact equation for computing free energies—with the knowledge of the entire phase space of a system—a realistic computer simulation must be finite in time, and thus free energy has to be computed using a finite set of configurations. Therefore, finite sampling error is an intrinsic part of free-energy simulations, and can be exacerbated by the rough energy landscape of the system under study. For a perturbation calculation, it is important to realize that there are two separate kinds of finite sampling errors: (1) finite sampling of the phase space of the reference (e.g., system 0) during the simulation, and (2) finite sampling of the phase space of the target (system 1), or equivalently, finite sampling of the perturbation distribution, while simulating system 0. We emphasize that (1) and (2) are distinct error sources. For example, free-energy calculations would still suffer systematic and random errors due to the problem of sampling the target phase space even when conformations of the reference system are perfectly sampled.[9] Unlike the error due to the approximation in system representation, generally finite sampling errors decrease as the sample size (simulation length) is increased.[11] Algorithms handling phase-space sampling and/or free-energy perturbation methods can play an important role in reducing finite sampling errors and thus improving the reliability of results as well as simulation efficiency. Exploring a simple perturbation method to effectively overcome both the random and the systematic errors due to finite sampling of the perturbation is the focus of this study, where we assume configurational sampling of the reference system is perfect and thus is not a concern.

Until recently, study of finite sampling inaccuracy was rare compared to that for random error. Two different approaches have been developed to investigate the statistics of finite free-energy simulation results. One of them[11,12,15,16] analyzes the most likely behavior of the result of an individual finite-length simulation, and quantifies the most likely inaccuracy (as well as most likely imprecision) in terms of probability density functions of perturbations. The other approach[8,13,14] examines the average behavior of systematic free-energy error due to finite sampling where the free energy can be expressed as power-series expansion of simulation length, and an extrapolation approach can be used to estimate the free-energy difference with infinite sample size. These advances in the fundamental understanding of both systematic and random errors will provide solid grounds for developing better algorithms for free-energy applications.

The perturbations sampled in a FEP simulation contribute to the free-energy result nonuniformly, for example, the contribution is proportional to the Boltzmann factor in eq. (1). For the calculation described by eq. (1), those perturbations with small (negative) values of u have the most significant contribution to the free energy. Those perturbations result from sampling important configurations of the target system during perturbation trials. The ability to effectively perturb into important configurations in the course of a finite-length simulation is the key to achieving a reliable free-energy calculation. Besides the simulation length, this ability is mainly related to two factors: the magnitude of the perturbation, and the choice of reference (i.e., the direction of the perturbation).[11,12,16]

It is well known that the FEP technique works well only when the two perturbation systems are similar, i.e., the perturbation itself is small.[4,17–19] It was shown in a recent study on finite sampling error that the perturbation magnitude is appropriately characterized by the entropy difference ΔS between the two systems.[12] Sampling should proceed from the high-entropy system perturbing into the low-entropy one, and the chance of sampling the important configurations of the target is reduced with $\exp(\Delta S)$. Not only must the target system be of lower entropy, but also its configurations must form a subset of the important configurations of the reference system—there cannot be low-energy configurations of the target that are high-energy configurations of the reference. If this subset relation does not apply, a single stage FEP calculation will fail to provide an accurate result.[9,16]

A given pair of systems 0 and 1 may be such that their important configurations overlap partially, and neither has its important configurations forming a subset of the other. In this case a FEP calculation will produce incorrect results regardless of its calculation direction ("forward" or "reverse"). It is well known[6,8,18,20–23] that the calculation in one direction overestimates the free energy, and that in the other, it underestimates. It is less appreciated that there is an "asymmetry" characteristic of the FEP calculation: the magnitudes of free-energy errors in the op-

posite directions are generally not the same, because of the different degrees that the important configurations of one system might be a subset of the other. Nevertheless, it is a widely used practice to take the mean of the forward and reverse results as an estimate of free-energy difference, hoping that the systematic errors will cancel.[8,18,23–29] We refer to this as *direct averaging* (DA). The double-wide sampling method[22] is a multistage version based on the same assumptions as direct averaging. Unfortunately, the asymmetry in the magnitude of errors is not accounted for in the DA method. As a result, the error cancellation is usually not complete, and one does not have control over the inaccuracy.[9,15,16] Ironically, the average result may be not as reliable as that by a single-direction FEP calculation (in the better direction of the two)—one is better off choosing the right direction and putting all the computational effort there, but often it is not easy to ascertain which direction is better.

To compute the free-energy difference between two systems with a large perturbation, and/or for which the subset relation does not apply, one must use a multistage perturbation scheme where the overall perturbation is divided into a set of consecutive substages (or windows) each with an appropriately small perturbation magnitude for the FEP calculation, and each obeying the subset relation. The overall free-energy difference is obtained by summing over all those substage free energies computed using the standard FEP technique. One or more (perhaps nonphysical) intermediates need to be introduced between the initial and final states to construct these multiple stages. The impact of the intermediates to the calculation is twofold. On the one hand, additional simulations are required on states that otherwise are of no interest. As a result, the computational effort may be increased; on the other hand, these intermediates provide opportunity for optimization for a more reliable and/or efficient overall calculation. The optimization of a multistage FEP calculation usually involves the choice of the number of stages, the definition of each intermediate, the amount of sampling time for each subperturbation, and the perturbation direction for each stage. For a generalized insertion multistage FEP calculation (where perturbations always follow from one intermediate to the next in a decreasing entropy direction), the optimization parameters for achieving a reliable and efficient calculation have been quantified in terms of the entropy difference.[12,16]

The *overlap sampling* (OS) technique was proposed in a recent study[30] as a way to stage the FEP calculation for systems having overlapping but nonsubset configurations. The idea is that two FEP calculations are performed sampling the 0 and 1 systems, respectively, with each perturbing into a common intermediate M designed to have important configurations from the intersection of the configurations for 0 and 1, i.e., in their overlap configurations. Impressive results were obtained from OS FEP for chemical potential calculations of simple Lennard–Jones and water systems where the forward and reverse calculations demonstrate strong asymmetry. Interestingly, this previous study showed that with a particular choice of weighting function, OS becomes identical to Bennett's method.[31] In addition, it showed that the free energy result by Bennett's method is not only the most precise, but also the most accurate. It will be shown in this study that the OS technique has a wide working range and a fast convergence rate relative to direct FEP averaging and to single direction FEP calculations, and it ensures the improved efficiency for free-energy calculations.

In this study, OS techniques are examined with examples of free-energy evaluation for the alchemical transformations because this is a very common and important application for computer simulations. Three systems are selected: (1) an anion system, (2) methanol to ethane mutation, and (3) changes to the adenosine molecule; they are studied because they are simple enough for a rigorous examination, yet complex enough to cover many important features that should be encountered in free-energy calculations of much larger molecular systems, such as in a full-atom protein system. Common observations over all three systems will be valuable for understanding the methodology, and for providing solutions of efficient and reliable free-energy calculations for computationally demanding applications, such as rational drug design.

Although OS has been discussed elsewhere, we briefly repeat it in the next section for the sake of completeness. Then we describe simulation design, analysis protocols, and the details of the test systems. Results and Discussion are then given, followed by Concluding Remarks.

Overlap Sampling

The OS technique aims to reduce the free energy simulation error and to improve calculation efficiency by appropriately combining FEP sampling in both the forward and reverse directions. It starts with the formulation of a *conceptual* intermediate M, which has as its important configurations only those configurations important to both initial (0) and final (1) states—i.e., an overlap of 0 and 1. The purpose of doing so is to ensure accessibility of the phase space important to M in simulations that are conducted on both 0 and 1 states and perturbed to M. With this choice the configurations important to M are a subset of both the 0- and the 1-important configurations, and this is required for accurate evaluation of the FEP averages in each substage. A reasonable starting point for the Hamiltonian of the intermediate is

$$H_M = -kT \ln w(\Delta H) + (H_0 + H_1)/2 \tag{2}$$

where w is a weighting function allowing adjustment of M; H_0 and H_1 are Hamiltonians for the 0 and 1 states (either the initial, final, or intermediate states), respectively; and $\Delta H = H_1 - H_0$. In the FEP framework, the kinetic energy contribution to the Hamiltonian is often ignored, and we use potential energies U_0 and U_1 in place of H_0 and H_1, respectively, and $u \equiv (U_1 - U_0)$ is the energy change encountered in a perturbation trial. The weighting function w can be adjusted for a better combined FEP sampling in two opposite directions and to correct the asymmetric contributions from these samples. For an arbitrary $w(u)$, the free-energy difference between states 0 and 1 is given by

$$\exp(-\beta\Delta A) = \frac{Q_M}{Q_0}\Big/\frac{Q_M}{Q_1} = \frac{\langle w(u)\exp(-\beta u/2)\rangle_0}{\langle w(u)\exp(+\beta u/2)\rangle_1}, \tag{3}$$

where Q indicates the partition function of a canonical ensemble. Equation (3) gives the general working formula for the overlap sampling method.

A convenient feature of the calculation prescribed by eq. (3) is that it requires simulations of only the 0 and 1 systems—no simulation need be performed that samples configurations according to the Hamiltonian for the intermediate M. In this regard, the simulation requirement for OS is the same as that for the DA approach. A benefit of this feature is that it permits convenient optimization of the calculation, because $w(u)$ can be adjusted freely without imposing changes or a redesign of the simulation process. In fact, many choices for $w(u)$ can be examined simultaneously in a single pair of (0, 1) simulations. Here, we examine two special cases: the simplest, nonoptimized version, and a version with full optimization for both accuracy and precision. In its simplest form we can simply choose $w(u) = 1$ for all u, which reduces the working equation to

$$\exp(-\beta\Delta A) = \frac{\langle \exp(-u/2)\rangle_0}{\langle \exp(+u/2)\rangle_1}, \quad (4)$$

where, once again, u is always defined as $u = U_1 - U_0$ (and the designation of the 0 and 1 systems can be arbitrary). In this case all samples in both directions are accounted for uniformly in computing the ensemble averages; however, the formalism of eq. (4) is different from that of the direct FEP averaging method, as u is divided by 2 in the exponential operations. With this simple choice of $w(u)$, the OS calculation requires exactly the same amount of computation as the DA method. However, the reliability of calculation results may differ. We refer to eq. (4) as the simple overlap sampling (SOS) method.[30]

As mentioned before, Bennett has shown how to minimize the random error of a FEP calculation combining both the forward and the reverse perturbations. If we choose a Gaussian-like hyperbolic secant function for $w(u)$ in eq. (3), i.e.,

$$w(u) = 1/\cosh[\beta(u - C)/2], \quad (5)$$

we obtain the same working formula as Bennett's method:[31]

$$\exp(-\beta\Delta A) = \frac{\langle\{1 + \exp[\beta(u - C)]\}^{-1}\rangle_0}{\langle\{1 + \exp[-\beta(u - C)]\}^{-1}\rangle_1} \exp(-\beta C), \quad (6)$$

where C is constant, and u has the same definition as that in eq. (4). The parameter C can be used to optimize both the precision and the accuracy of the free-energy calculation. The optimal choice of $C = \Delta A$ (assuming that both forward and the reverse directions have the same sampling size) for the precision was shown to be best for accuracy.

The probability distribution functions of perturbations encountered in a FEP calculation, denoted as $f(u)$ and $g(u)$ for the forward and reverse directions, are useful for understanding the accuracy of OS calculation. These distributions obey[2,32,33]

$$f(u)\exp(\beta\Delta A) = g(u)\exp(\beta u). \quad (7)$$

Further, the fractional inaccuracy of $\exp(-\beta\Delta A)$ due to finite sampling can be expressed in terms of f and g,[30]

$$\delta_e \approx \frac{\int_{-\infty}^{u_f} w(u)[f(u)\,g(u)]^{1/2}du - \int_{u_g}^{\infty} w(u)[f(u)\,g(u)]^{1/2}du}{\int_{-\infty}^{\infty} w(u)[f(u)\,g(u)]^{1/2}du}, \quad (8)$$

where $\delta_e = 1 - \exp(-b\Delta A^{\text{sim}})/\exp(-b\Delta A^{\text{exact}})$, and u_f and u_g are "limit energies," which is a quantity used in the model assumption—all perturbations up to the limit energy ($u \geq u_f$ for a forward calculation and $u \leq u_g$ for reverse) are sampled perfectly but no sampling occurs at all beyond it.[11] Clearly, the systematic error of the OS calculation is related to the degree of overlap between the f and g distributions, which in turn, is determined by the perturbation magnitude and the sampling size.

At the crossing point of $f(u^*) = g(u^*)$, $u^* = \Delta A$ [cf. eq. (7)], and the f and g distributions have the greatest overlap. The particular choice of $C = \Delta A$ actually puts the maximal weight on u^*. By self-consistently solving the optimization parameter C (e.g., using a simple iteration process), the method is able to locate the correct free energy answer (the crossing point) from the perturbation data. Note that the computational time for solving C and thus ΔA is negligible ($\sim$1 s or less) when compared to simulation sampling (hours or days). A certain degree of overlap between f and g is needed for the algorithm to locate the optimal value of C. In eq. (8) the systematic error would become large if there were no overlap between f and g (the denominator goes to zero). Nevertheless, the degree of overlap required by the OS method is rather small, as we will note in the Results and Discussion section.

One might be interested in the performance of eq. (6) when $C = 0$—referred to here as the C0 method—whereby the procedure for solving the parameter is eliminated. We expect that the performance of such a method would be uneven, depending on the nature of the system and the perturbation. For example, for a perturbation with $\Delta A \sim 0$, its performance could be similar to that of the fully optimized Bennett's method; but when ΔA is far away from zero, this method would inappropriately overemphasize the data from FEP sampling in one direction by putting most of the weight there, and undercount those from the other direction at the same time, thus producing a free-energy result with marginal improvement over a single-direction FEP calculation. In the latter case, it would be much better to go with the (even simpler) SOS method. Considering its inconsistent performance, we do not recommend the C0 method as a general approach for computing free-energy differences.

Simulation Methodology

We test the methods discussed above for the evaluation of free-energy changes in alchemical transformations which is an important topic in molecular simulation. In summary, these methods are: FEP in forward direction, FEP in reverse direction, direct FEP averaging, simple overlap sampling, Bennett's method and the C0 method. We examine the performance of these methods side by side and keep all other factors the same. This is achieved by conducting the free-energy calculations as a postsimulation process; that is, we first perform the simulations and then collect the

entire set of perturbation data; different methods are then used to generate free-energy estimates using the identical set of sampled data. Thus, any difference in performance observed will be purely due to the free-energy postprocessing methodology. We are permitted to do so because we are investigating the relative performance of free-energy methodologies, rather than collecting free-energy values for particular systems. In addition, we have the freedom of choosing parameterization, perturbation, as well as simulation methods (e.g., dual topology or single topology, Monte Carlo, or molecular dynamics), as far as the comparison is based on the same grounds, which is guaranteed by our analysis.

The imprecision (standard error) of free energy results by different methods can be computed using techniques such as bootstrap and jackknife error estimate,[34–36] or from results of simulations repeated independently. In this study, we conduct multiple independent simulations for each testing system, and the variances of these results are evaluated to provide an estimate for the precision of all the methods. We have also performed bootstrap and jackknife estimates as a crosscheck for the consistency of this analysis. The inaccuracy (systematic error, or bias), on the other hand, cannot be estimated reliably using these methods,[37] and requires knowledge of the exact values of the free-energy difference, ΔA^{exact} as a reference. In this study, the "exact" value of free-energy differences for a given perturbation is estimated using a multistage perturbation calculation with a very long simulation length. When the perturbation magnitude is small enough and the simulation length is sufficiently long, all FEP techniques are supposed to be able to produce correct answers for ΔA. We crosscheck the condition of "small perturbation" and "long simulation" by examining the free-energy results produced by different methods—these estimates should be almost identical. We then take the average of results over different methods as the reference ΔA^{exact} for that particular substage. We repeat the process for all substages, and finally sum over these ΔA for a larger perturbation and treat this as the "exact" value for that perturbation.

Alchemical states are defined using λ parameter scaling. The alchemical degree of freedom λ can have a value between 0 and 1, with 0 and 1 representing the initial and the final states, respectively, and a value in between representing an intermediate. Specifically

$$U(\lambda) = U_0 + \lambda(U_1 - U_0) \tag{9}$$

For convenience, we adopt a set of fixed and equally spaced λ values for staging. Note, in principle, any two λ-states could be used to form a perturbation for the purpose of examining the performance of the methodology, as is warranted by our design. This feature permits a convenient way to study the effect of perturbation magnitude on the performance of different methods. To focus on the methodology comparison, we intentionally avoid simulations on states with λ equal or very close to 0 or 1 (except for the simple anion system, see below) to avoid possible end-state problems.[38]

Alchemical transformations are performed on three systems in explicit aqueous solutions: an anion system changing from one type to another, a mutation between methanol and ethane, and a transformation within an adenosine molecule. For all these simulations, the CHARMM molecular dynamics (MD) program[39] is used, and MD simulations are performed in an *NPT* ensemble with a pressure of 1 atm and a temperature of 300 K (controlled using the Nosé–Hoover[40,41] thermostat); nonbonded interactions are represented by a shifted-force potential[1] that diminishes to zero at 9 Å. Cubic periodic boundary conditions are used. Perturbation calculations are performed within the BLOCK array feature of the CHARMM package (where a dual-topology protocol is adopted). The TIP3P potential model[42] is used for water molecules. The systems are energy-minimized and preequilibrated before any production run is initiated.

Anion System

The anion potential parameters are based on the Lybrand et al.[43] parameterization and are taken to represent two particular anion states, rather than being a model of any particular anion type. They are close, in spirit, to other calculations of anion change (e.g., Br^- to Cl^- and vice versa). We use 13 λ values that are equally spaced between 0 and 1 to define the alchemical states, with 0 and 1 included because no special concern on the end-states is needed for this simple system. The periodic simulation box has an edge length of 18.856 Å, and each cubic box contains 216 water molecules besides the anion. The anion is fixed at the center of the simulation box during the course of the MD simulation. We use a step size of 2 fs for integration with the SHAKE[44] constraint applied to all bonds connected to a hydrogen atom. For each perturbation pair, a total of 15 repeated production runs with 1 ns each is conducted. System configurations are collected at intervals of 0.1 ps and saved for computing the free-energy differences in the postsampling stage. Thus there are 10,000 conformations stored in each trajectory. For computing the "exact" free energy difference, long production runs with 15 ns for each λ-state were conducted. Through this level of sampling we are able to comment on the relative statistical and systematic errors in the system.

Methanol–Ethane System

Methanol and ethane molecules are modeled in the united-atom representation: methanol comprises one united C atom as well as O and H atoms, while ethane has two united C atoms. The alchemical states along the reaction coordinate are defined by a set of λ values equally spaced with $\Delta\lambda = 0.05$. This system, as with most alchemical changes, has end-state problems when $\lambda = 0$ or 1. Thus, we simulate alchemical states with intermediate values of λ. The simulation box contains 216 TIP3P water molecules and the periodic edge is 18.856 Å. The MD production run for each state lasts 2 ns after full equilibration, and the step size is 2 fs with SHAKE constraint applied to all bonds with a H-atom directly connected. System conformations are saved for free-energy analysis with a period of 0.1 ps. We repeat the production runs independently 12 times for statistical analysis. Dynamics simulation of 20 ns is performed on each λ-state to produce the "exact" (FEP estimated) free-energy difference. Note in our simulation we fix the united C atoms of the methanol and ethane hybrid near the center of the simulation box. However, we do not fix the orientation (direction of C—O bond in methanol and C—C bond in ethane) of the molecules.

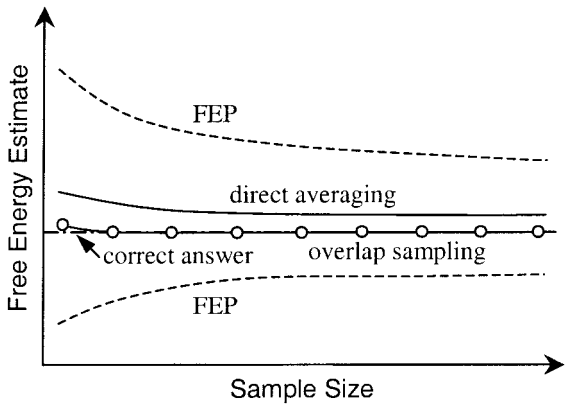

Figure 1. A schematic presentation of the general observation for the behavior of different free energy difference calculation methods. Key: dashed curves, FEP method in one single direction; solid curve, direct FEP averaging method; solid curve with open circles, overlap sampling method; dash-and-dotted horizontal line, the correct free-energy answer.

Adenosine System

The adenosine molecule is of biological interest itself.[45] The test alchemical transformation involves a mutation between a —CH_3 group and a —CH_2OH group ($C_{10}N_5O_4H_{13} \leftrightarrow C_{10}N_5O_3H_{13}$). We performed simulation tests with two different system sizes: one with 1038 water molecules, and the other with 510. The free-energy values for these two system sizes vary slightly, as expected; but the performance of the different free-energy methods shows the same kind of behavior. We conclude that the smaller system (with 510 water molecules) works well for our purpose and requires much shorter simulation time over the larger one; thus, in this article we report results only for the smaller adenosine system. The periodic edge of the smaller simulation box is 24.90 Å. The adenosine molecule is initially placed in the center of the simulation box, and a harmonic force constraint with force constant of 1.0 kcal/mol Å^2 is applied to the NN2 atom of the molecule to keep it centered during the simulation. The λ value is chosen to be equally spaced with value of 0.05 between 0 and 1; and again, end-states are not simulated. For each λ, the system is first energy minimized and equilibrated, then a MD production run for 750 ps is conducted. The MD time step is 1.5 fs with SHAKE constraints applied for all bonds having H-atom directly connected. 10,000 conformations are collected in one MD trajectory with an interval of 0.75 ps. We repeat the simulation 14 times for each perturbation to collect statistical information. The "exact" free-energy difference is estimated from a long MD run of 11.25 ns.

Results and Discussion

Before reporting the detailed results and discussion, we first present a schematic plot in Figure 1. This cartoon qualitatively demonstrates the general behavior of different free-energy calculation methods observed over the varieties of systems and perturbations being tested. Although the reliability of all the calculations is improved as the sampling size increases, a single direction FEP method usually converges too slowly, and fails to produce a proper free-energy estimate during the course of a simulation, except for small perturbation cases. The direct FEP averaging method usually improves the convergence, but its result may be systematically off from the correct answer due to the asymmetric behavior of the forward and reverse FEP calculations. The magnitude and the sign of this systematic error cannot be generally predicted because they are usually related to details of the systems, the perturbations, the sampling sizes, etc. In contrast, the overlap sampling method is able to produce a correct free-energy estimate with a relatively small set of sample data. In the following, we present our simulation results, and discuss important points including those outlined above in greater detail.

The accumulated free-energy changes along the reaction coordinate with long runs are presented in Figure 2 for all three systems. With current choices of multistaging, all free-energy methods produce roughly the same free-energy result for the anion and adenosine systems, as all curves collapse to a single one. The only exception is that for the methanol–ethane transformation, the accumulated free-energy change by the forward FEP and reverse FEP calculations starts to be slightly off around $\lambda = 0.5$. But these offsets appear to be roughly symmetric, due to the small perturbation ($\Delta\lambda = 0.05$) used in each stage. The ΔA^{exact} is obtained from these long simulation results according to the protocol described in the preceding section.

The direct FEP averaging method works reasonably well when the perturbation magnitude is small, as noted in Figure 2. In Figure 3, we give another set of examples, where results from small perturbations ($\Delta A \sim 0.5$ kcal/mol) between two λ-states of the methane–ethane and adenosine transformations are presented. In this figure, the free-energy estimates by different methods are presented in the lower half of one plot, and their random errors in the upper half, both as a function of sampling size (directly proportional to the MD production time). In short, one can examine the inaccuracy and free-energy convergence rate from the lower half of the plot, and imprecision from the upper half of the plot. For this particular figure, the free-energy error bar is computed using the error propagation formula from the block average energy-change statistics, representing the standard deviation of the mean of block averages; details of the blocking are described in the caption. For both systems, results of ΔA by the SOS, and Bennett's method are visually the same; but the forward and reverse FEP results clearly differ, and no convergence is reached within the test simulation length. The free energy estimate by the DA method is about the same as those by the SOS and Bennett's, because the FEP calculations are practically "symmetric" for this small perturbation. However, the difference in the imprecision of the calculations is noticeable: Bennett's method giving the smallest value, followed by the SOS method.

The difference in the performance of the methods becomes more prominent as the perturbation magnitude increases. We present larger perturbation examples in Figure 4: $\Delta A \sim 10.4$ kcal/mol for the anion system (plot a), 3.7 kcal/mol for the methanol–ethane mutation (plot b), and 8.5 and 3.5 kcal/mol for the adenosine transformations (plots c and d), respectively. Figure 3 used results of one MD production run; in contrast, Figure 4 shows the average results over 15 (for anion), 12 (for methanol–ethane),

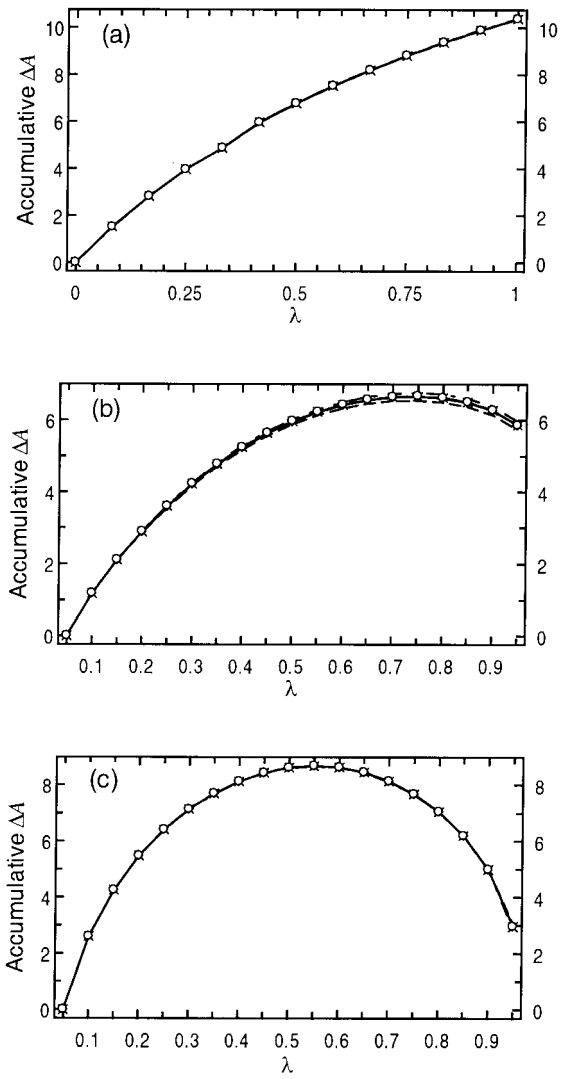

Figure 2. The accumulative free-energy difference ΔA along the λ coordinate for (a) the anion, (b) the methanol–ethane, and (c) the adenosine systems. Results by different FEP methods (forward FEP, reverse FEP, DA, SOS, and Bennett's) are shown, but they are nearly collapsed together.

or 14 (for adenosine) independently repeated MD runs. The error bars, representing the standard deviation of the mean, are computed from those repeated free-energy estimates at a given sampling size. The most likely behaviors are also investigated but not presented here because the observations are similar to those in Figure 4. The reference ΔA^{exact} is presented as a bold horizontal line.

Under these perturbation magnitudes, the results of the C0 method are very close to those by one of the single-direction FEP calculation, both in terms of precision, accuracy, and convergence rate. Clearly, the choice of $C = 0$ provides an overweighting of the sample data from one FEP direction, and the calculation is not optimal. For the purpose of clarity, those results are not presented. Regardless of noise, which will decrease with more repeated simulation data, common behavior can be clearly observed over all three systems. Not surprisingly, both of the single direction FEP calculations fail to converge to the correct answer (see the lower halves of plots). Also obvious is their asymmetric behavior. As a result, the DA method reports a free-energy answer offset from the correct value. The magnitude and the direction of the offset depends on the perturbation system, on the perturbation magnitude, as well as on the simulation length; and no general rule can be developed to tell when the method works well. In contrast, both the SOS and Bennett's methods work very well. They successfully converge to the correct free energy answer within the simulation course. The convergence rate of Bennett's method is especially amazing, as the right answer is reached almost at the very beginning of the simulation (10–200 ps) despite the large magnitude of the perturbation. The free-energy curves of the overlap sampling methods appear to be smoother than those of conventional FEP calculations (all with the same perturbation data), suggesting a robust formulation.

The precision of the free-energy calculations can be examined from the upper half of the plots in Figure 4. The difference in the random error of the different methods can be clearly seen. The precision of the overlap sampling methods is impressive, while Bennett's method consistently produces the lowest random error. It is able to reduce the random error by 3–10 times over that of the DA method, and usually even more over those by the single-direction FEP. For the examples shown in Figure 4, only 1% or less sample size is needed for the OS methods to achieve the same precision level as the conventional FEP method can reach at the end of the simulation. The random error decay rate as a function of sampling size also differs from method to method. For a better observation, the random errors are presented in a log-scale plot in the upper half of Figure 4d. The OS methods have a higher decay rate (with a slope $\sim -1/2$), indicating a more cost-effective way to reduce random error by increasing the simulation length.

We note that the behavior observed here is typical over different perturbation pairs (defined by the starting and ending states) with a similar perturbation magnitude. To demonstrate, in Figure 5 we present and compare the random error of Bennett's method and of direct FEP averaging for the adenosine system over the entire perturbation range examined. The random errors are computed in the same way as in Figure 4, i.e., as the standard deviation of the mean for 14 repeated simulation runs each with 750 ps MD production time. The imprecision in free energy difference for any λ_i–λ_j pairs can be seen in the contour plot, which is symmetric along the diagonal line of the x and y axes. Bennett's method consistently does a better job over the entire perturbation range. Similar (but slightly less) improvement is observed for the SOS method (plots not shown); and such observations are consistent over all three systems tested. The contour plot may be useful in other ways. For example, it could help to identify the "maximum" perturbation magnitude that can be handled in a one-stage FEP calculation for a given simulation length. In this particular case, the plot suggests that by using Bennett's method, starting from $\lambda = 0.05$, one could go to the $\lambda = 0.6$ state in a one-stage perturbation, and the free-energy result would be reasonably precise (with an error bar of 0.2 kcal/mol). In contrast, the DA method reaches the same level of imprecision for the $\lambda = 0.05 \leftrightarrow 0.2$ perturbation pair.

Clearly, by using Bennett's method, the working range can be largely extended.

Both the bootstrap and the jackknife analyses of standard error (imprecision) yield similar results (not presented here) as those observed above, and thus lead to the same conclusion regarding this aspect of the performance of different methods. Although the estimate of systematic error by the bootstrap/jackknife gives the correct relative ordering of the accuracy for different free energy methods, these techniques tend to underestimate the true systematic error (which could then mislead the interpretation of simulation results). As we pointed out,[37] this is because such techniques cannot account for the major systematic errors arising from inadequate sampling of important phase space. Instead, the best practice is to avoid the systematic error in the first place during the simulation, by using appropriate staged free-energy calculation schemes such as overlap sampling.

In summary, the overlap sampling techniques provide improved reliability, faster convergence rate, and extended working range for perturbation over the conventional FEP methods. These advantages no doubt will facilitate a more efficient free-energy calculation. For example, the OS technique will be able to produce an equally reliable or even better free-energy estimate than the conventional FEP methods using a shorter simulation with a small number of stages. In the following we examine the effectiveness of the OS methods in the context of multistage calculations for the anion, methanol–ethane, and adenosine mutations.

We employed various staging schemes to compute free energy difference for a given perturbation pair. More stages are used for the conventional FEP methods and less stages for the overlap sampling technique—this was done with the consideration of different working ranges for these methods. For simplicity, the same sampling size is used through all stages, so the relative computational effort needed by different methods can be easily compared by counting the number of stages. We then examined the reliability of free-energy results produced by these calculations. Note we choose intermediates only from the equally spaced λ-states previously used in computing ΔA^{exact}, with the principle to roughly form an equal perturbation magnitude for all stages.

For the anion system, the perturbation is chosen to be between the states of $\lambda = 0$ and 1. A single-stage calculation is adopted for the OS methods, and five-stage calculation for the conventional FEP methods (λ = 0.0, 0.083, 0.25, 0.42, 0.67, and 1.0). Com-

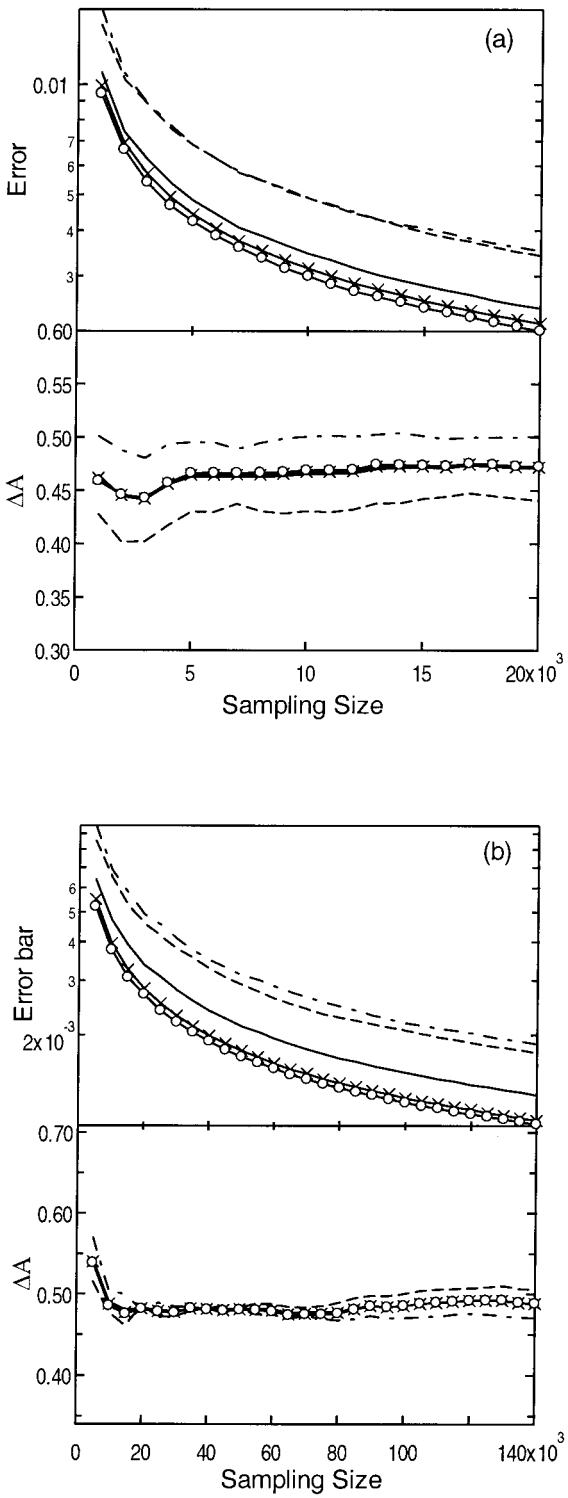

Figure 3. Free-energy results computed by different methods at small perturbation as a function of perturbation sampling size. Plot a: the mutation of methanol and ethane between two λ-states with $\lambda = 0.5$ and 0.6. Plot b: the alchemical transformation of the adenosine molecule between states having $\lambda = 0.4$ and 0.5. The upper and lower half of the plot gives the estimated free-energy difference and random error as a function of sampling size, respectively. These results are from a single MD production run whose simulation length is 24 ns for the methanol–ethane system and 10.5 ns for the adenosine system, respectively. The standard deviation of the mean (representing 68% probability) is computed using the error propagation formula from the blocked energy perturbation data. Each block contains 500 and 375 ps for plot a and b, respectively. Key: dashed curve—forward FEP, dash-dotted curve—reverse FEP, solid curve—direct FEP averaging, solid curve with crosses—simple overlap sampling, solid curve with open circles—Bennett's method, bold horizontal line in the lower half—the reference value of ΔA. Data have units of kcal/mol.

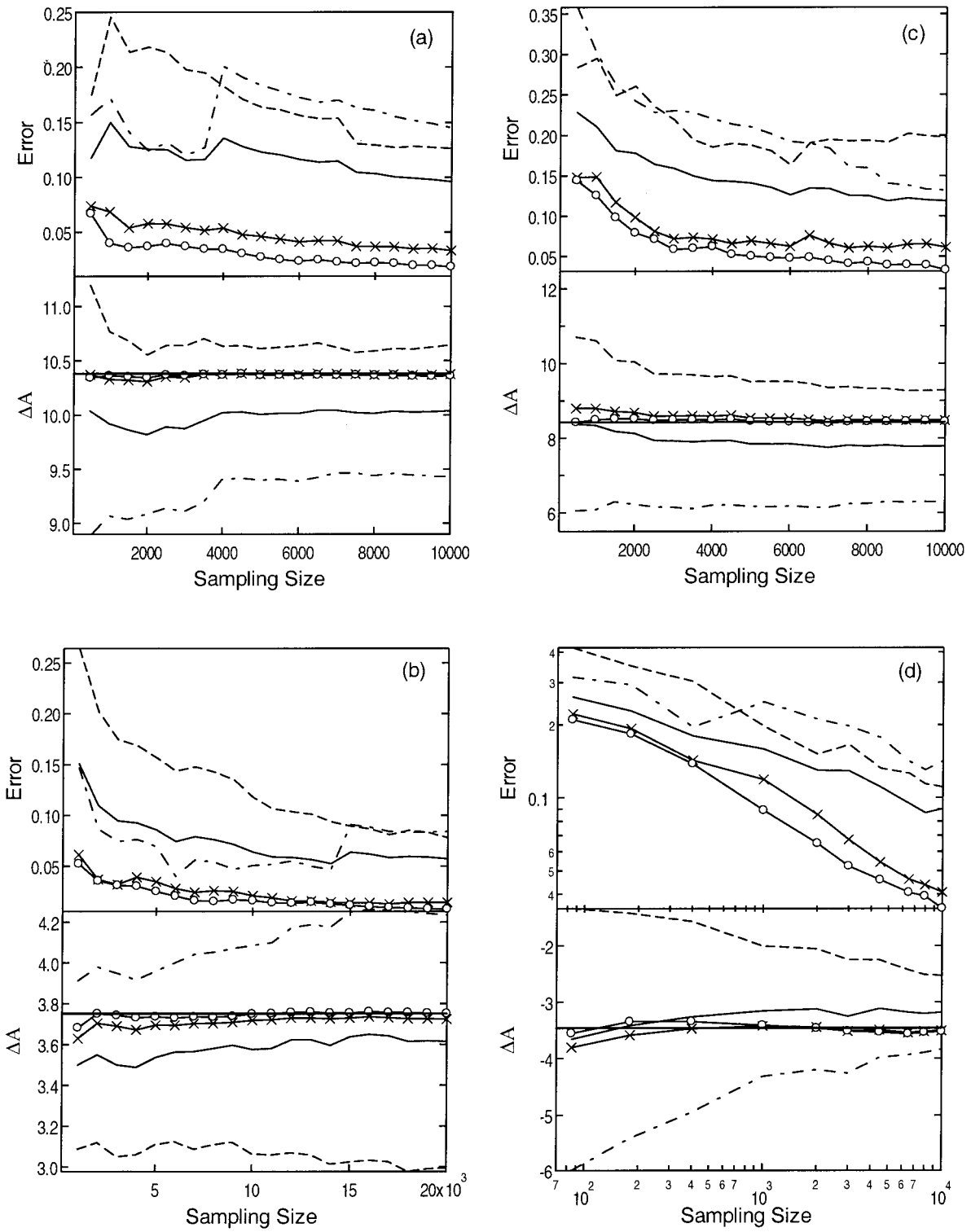

Figure 4. The average free-energy results of repeated MD simulations. (a) The anion system for the perturbation between $\lambda = 0.0$ and 1. (b) The methanol–ethane system, $\lambda = 0.2$ and 0.7. (c) The adenosine system, $\lambda = 0.05$ and 0.45. (d) The adenosine system, $\lambda = 0.45$ and 0.9. Units: kcal/mol. Same keys in Figure 3 are used.

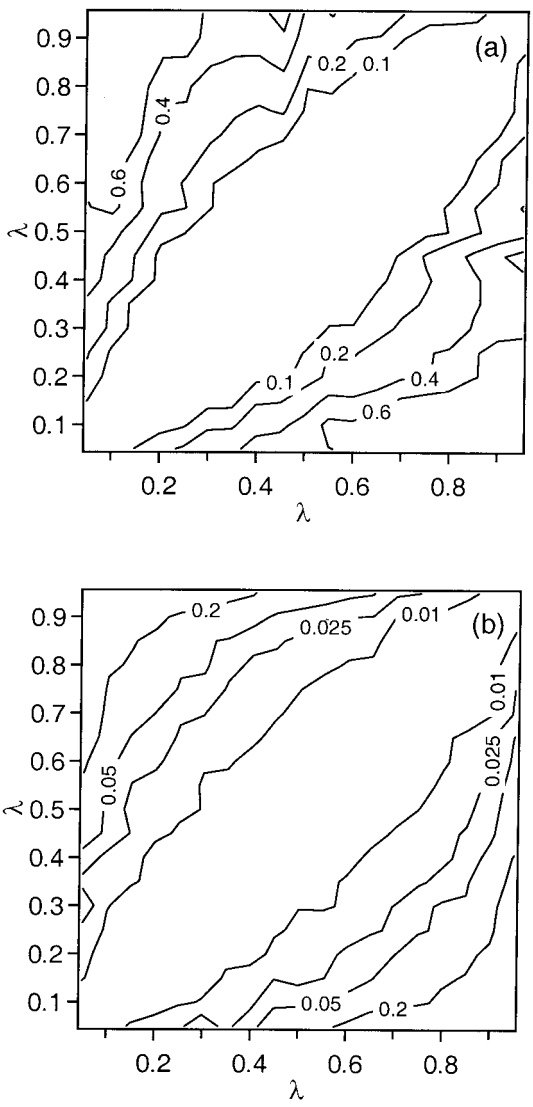

Figure 5. The free-energy difference random error for the adenosine mutation computed using (a) direct FEP averaging method, and (b) Bennett's method over the perturbation range of $\lambda = 0.05$ and 0.9.

pared to the OS calculations (sampling two states), the computational time used for the forward or reverse FEP is roughly 2.5× (sampling five states), and 3× for the direct FEP averaging (sampling six states). For the methanol–ethane transformation, we choose the perturbation to be between $\lambda = 0.2$ and 0.9, and compare performance of single-stage OS method with a four-stage calculation for conventional multistage FEP calculation (λ: 0.2, 0.3, 0.45, 0.75, and 0.9). For adenosine, the test perturbation is between $\lambda = 0.05$ and 0.9. Two different two-stage schemes are adopted for overlap sampling methods: sampling on states with $\lambda = 0.05$, 0.45, 0.9, and $\lambda = 0.05$, 0.6, 0.9, respectively. Three different multistage designs are used for the conventional FEP methods: (1) six-stage calculation with intermediates of $\lambda = 0.1$, 0.15, 0.25, 0.45, and 0.8, (2) nine-stage scheme with intermediate of $\lambda = 0.1$, 0.15, 0.2, 0.25, 0.35, 0.65, 0.8, and 0.85, and (3) 17-stage scheme with fixed λ increment of 0.05 for adjacent states.

The results of these computational errors are listed in Table 1. As one can see, overlap sampling methods, with much less overall computational cost, produce equivalent or better results than conventional multistage FEP calculations. For the adenosine system, both two-stage overlap sampling schemes produce impressive results; this implies the staging design may be less an issue in practice for free-energy calculation using overlap sampling calculations. As mentioned before, the staging for the OS method should ensure sufficient overlap between the f and g distributions to produce acceptable free energy results. However, we find that the needed overlap may be rather small. For example, in the perturbation of $\lambda = 0.05$ to 0.6 for adenosine transformation, we have $z = \int [f(u)\, g(u)]^{1/2} du \approx 0.001$ (visually almost no-overlap between f and g distributions) and the error is far less than 1 kT.

Concluding Remarks

The overlap sampling technique demonstrates excellent ability to reduce the finite time perturbation sampling errors for all three systems studied in this article. Compared to the conventional FEP techniques, both the fully optimized version (i.e., Bennett's method) and the simple version (SOS) are able to greatly improve the precision and accuracy of the free-energy calculation. They also feature fast convergence of the free-energy estimate, high random error decay rate, and extended perturbation working range. All of these make for a reliable and efficient free-energy evaluation using the overlap sampling technique. In addition, all the improvement comes with no additional simulation requirement other than conducting the conventional FEP calculation in two opposite directions.

Bennett's method demonstrates excellent capability in reducing both the systematic and random errors, and it is highly recommended as a general perturbation scheme for computing free-energy differences. For those who do not like the slightly additional iteration work required by Bennett's method, the simple overlap sampling is a good alternative; however, the method simply of choosing $C = 0$ for Bennett's parameter is not recommended unless the perturbation ΔA is known to be small. The direct FEP averaging method (or double-wide method) uses the same computational effort as the overlap sampling but produces much lower quality free-energy estimates; thus, we suggest avoiding this approach.

One should keep it in mind that all the perturbation techniques require a certain similarity between the two systems. Multistaging is an effective solution for free-energy calculation in systems with a significant difference. The number of stages required by an overlap sampling method will be much smaller than that by conventional FEP techniques, because of an extended working range for perturbations. The intermediate design in an FEP calculation is an important factor for an effective free-energy calculation. Some hints obtained from the current and previous studies may be useful for developing a detailed and easy-to-use principle for optimal staging (or switching) design. It has been argued that for an FEP calculation, the staging based on uniform entropy difference ΔS (rather than uniform ΔA or $\Delta\lambda$) for all the stages is the most

Table 1. Comparison of Reliability in Free-Energy Calculation Using Overlap Sampling and Conventional FEP Methods with Different Staging Schemes.

	Inaccuracy (%)			Random error		
Methods	Fwd FEP	Rev FEP	DA	Fwd FEP	Rev FEP	DA
A5	1.05	1.01	0.86	0.0081	0.011	0.0069
M4	11.43	13.69	1.40	0.023	0.022	0.016
D6	9.69	1.77	5.27	0.38	0.077	0.21
D7	2.40	2.66	2.49	0.045	0.047	0.014
D9	3.01	3.56	2.06	0.052	0.049	0.037
D17	1.59	1.87	0.25	0.020	0.029	0.019

	Inaccuracy (%)		Random error	
Methods	SOS	Bennett's	SOS	Bennett's
A1	0.10	0.10	0.024	0.012
M1	1.37	0.31	0.034	0.008
D2a	1.14	0.31	0.061	0.027
D2b	2.45	1.87	0.083	0.060

The numbers listed represent the average systematic error (in percentage) and random errors (kcal/mol; the standard deviation of the mean) for 15, 12, and 14 repeated MD simulations for the anion, methanol–ethanol, and adenosine systems, respectively.

For the anion perturbation between $\lambda = 0.0$ and 1.0: A1—a single stage calculation; A5—a five-stage scheme with $\lambda_i = 0.0, 0.083, 0.25, 0.42, 0.67$, and 1.0. For the methanol–ethane perturbation between $\lambda = 0.2$ and 0.9: M1—a single-stage calculation; M4—a four-stage scheme with $\lambda_i = 0.2, 0.3, 0.45, 0.75$, and 0.9. For the adenosine perturbation between $\lambda = 0.05$ and 0.9: D6—a six-stage calculation with $\lambda_i = 0.05, 0.1, 0.15, 0.25, 0.45, 0.8$, and 0.9; D7—a seven-stage scheme with $\lambda_i = 0.05, 0.1, 0.15, 0.25, 0.45, 0.6, 0.8$, and 0.9; D9—a nine-stage scheme with $\lambda_i = 0.05, 0.1, 0.15, 0.2, 0.25, 0.35, 0.65, 0.8, 0.85$, and 0.9; D17—a 17-stage scheme, λ_i has an increment of 0.05 from 0.05 to 0.9; D2a—a two-stage scheme with $\lambda_i = 0.05, 0.45$, and 0.9; D2b—a two-stage calculation with $\lambda_i = 0.05, 0.6$, and 0.9.

effective,[12] and principles based on it have been developed.[11,16] The probability density distributions *f* and *g* could provide an alternative route. As we show in this study, the reliability of free energy calculation can be related to the degree of overlap between these two distribution functions. Such a relationship could facilitate a practical way for determining the optimal multistaging design, and for assessing the reliability of calculation results without knowing the "true" answer. In addition, we note that the analysis of *f* and *g* distributions, including the numerical fittings of these distributions, is helpful in further improving the reliability the overlap histogram methods based on eq. (7),[46] similar improvement could be achieved for the overlap sampling calculation.

The effectiveness of free energy perturbation calculations makes it a good choice for computationally demanding simulations. We hope, with the reliability and efficiency it brings, that the overlap sampling scheme can relieve the difficulty in computing free energy by molecular simulation, and extend the capability for simulating computationally demanding, complex systems, especially biomolecular systems.

References

1. Allen, M. P.; Tildesley, D. J. Computer Simulation of Liquids; Clarendon Press: Oxford, 1987.
2. Frenkel, D.; Smit, B. Understanding Molecular Simulation: From Algorithms to Applications; Academic Press: San Diego, 2002, 2nd ed.
3. Kollman, P. Chem Rev 1993, 32, 2395.
4. Beveridge, D. L.; DiCapua, F. M. Annu Rev Biophys Chem 1989, 18, 431.
5. Leach, A. R. Molecular Modelling, Principles and Applications; Prentice Hall: London, 2001, 2nd ed.
6. Mark, A. E. In Encyclopedia of Computational Chemistry; Schleyer, P. v. R., Ed.; John Wiley & Sons: New York, 1998, p. 1070.
7. Zwanzig, R. W. J Chem Phys 1954, 22, 1420.
8. Wood, R. H.; Mühlbauer, W. C. F.; Thompson, P. T. J Phys Chem 1991, 95, 6670.
9. Kofke, D. A.; Cummings, P. T. Mol Phys 1997, 92, 973.
10. Kofke, D. A.; Cummings, P. T. Fluid Phase Equil 1998, 150, 41.
11. Lu, N.; Kofke, D. A. J Chem Phys 2001, 115, 6866.
12. Lu, N.; Kofke, D. A. J Chem Phys 1999, 111, 4414.
13. Zuckerman, D. M.; Woolf, T. B. Chem Phys Lett 2002, 351, 445.
14. Zuckerman, D. M.; Woolf, T. B. Phys Rev Lett 2002, 89, 180602.
15. Lu, N.; Kofke, D. A. In Foundations of Molecular Modeling and Simulation, AIChE Symp Ser; Cummings, P.; Westmoreland, P., Eds.; American Institute of Chemical Engineers: New York, 2001, p. 146.
16. Lu, N.; Kofke, D. A. J Chem Phys 2001, 114, 7303.
17. Radmer, R. J.; Kollman, P. A. J Comput Chem 1997, 18, 902.
18. Pearlman, D. A.; Kollman, P. A. J Chem Phys 1989, 90, 2460.
19. Lu, N.; Kofke, D. A.; Woolf, T. B. J Phys Chem B 2003, 107, 5598.
20. Powles, J. G. Chem Phys Lett 1982, 86, 335.

21. Parsonage, N. J Chem Soc Faraday Trans 1996, 92, 1129.
22. Jorgensen, W. L.; Ravimohan, C. J Chem Phys 1985, 83, 3050.
23. Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado–Rives, J. J Chem Phys 1988, 89, 3742.
24. Chipot, C.; Gorb, L. G.; Rivail, J.-L. J Phys Chem 1994, 98, 1601.
25. Blake, J. F.; Jorgensen, W. L. J Am Chem Soc 1990, 112, 7269.
26. Pearlman, D. A. J Comp Chem 1994, 15, 105.
27. Henchman, R. H.; Essex, J. W. J Comp Chem 1999, 20, 499.
28. Reynolds, C. A.; King, P. M.; Richards, W. G. Mol Phys 1992, 76, 251.
29. Li, J.; Platt, E.; Waszkowycz, B.; Cotterill, R.; Robson, B. Biophy Chem 1992, 43, 221.
30. Lu, N.; Singh, J. K.; Kofke, D. A. J Chem Phys 2003, 118, 2977.
31. Bennett, C. H. J Comput Phys 1976, 22, 245.
32. Shing, K. S.; Gubbins, K. E. Mol Phys 1982, 46, 1109.
33. Allen, M. P. In Proceedings of the Euroconference on Computer Simulation in Condensed Matter Physics and Chemistry; Binder, K.; Ciccotti, G., Eds.; Italian Physical Society: Como, Italy, 1996, p. 255.
34. Efron, B.; Tibshirani, R. J. An Introduction to the Bootstrap; Chapman & Hall/CRC: New York, 1993.
35. Deitrick, G. L.; Scriven, L. E.; Davis, H. T. J Chem Phys 1989, 90, 2370.
36. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. Numerical Recipes: The Art of Scientific Computing; Cambridge University: Cambridge, 1992.
37. In other situations bootstrap/jackknife methods are used to provide an estimate of systematic error from resampling the data set; however, such an estimate is not appropriate for characterizing the free energy systematic errors, which are caused by failing to sample the important phase space of the target system in a finite-length simulation. Such systematic error is not represented by the sample dataset itself, nor can it be "extrapolated" from it; thus, the bootstrap/jackknife will fail to render a good estimate.
38. Pearlman, D. A. J Phys Chem 1994, 98, 1487.
39. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187.
40. Nosé, S. Mol Phys 1984, 52, 255.
41. Hoover, W. G. Phys Rev A 1985, 31, 1695.
42. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. J Chem Phys 1983, 79, 926.
43. Lybrand, T. P.; Kollman, P. A. J Chem Phys 1985, 83, 2923.
44. Ryckaert, J. P.; Cicotti, G.; Berensden, H. J. C. J Comput Phys 1977, 23, 327.
45. Williams, M., Ed. Adenosine and Adenosine Receptors; Humana Press: Clifton, NJ, 1990.
46. Lu, N.; Woolf, T. B. Mol Phys 2003, submitted.