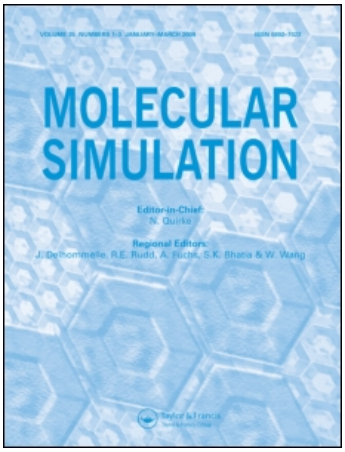

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Calculation of Solvation Free-Energy Differences for Large Solute Change from Computer Simulations with Quadrature-Based Nearly Linear Thermodynamic Integration

Mihaly Mezei[a]
[a] Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York, NY

To cite this Article Mezei, Mihaly(1993) 'Calculation of Solvation Free-Energy Differences for Large Solute Change from Computer Simulations with Quadrature-Based Nearly Linear Thermodynamic Integration', Molecular Simulation, 10: 2, 225 — 239
To link to this Article: DOI: 10.1080/08927029308022166
URL: http://dx.doi.org/10.1080/08927029308022166

Molecular Simulation, 1993, Vol. 10(2-6), pp. 225-239
Reprints available directly from the publisher
Photocopying permitted by license only

CALCULATION OF SOLVATION FREE-ENERGY DIFFERENCES FOR LARGE SOLUTE CHANGE FROM COMPUTER SIMULATIONS WITH QUADRATURE-BASED NEARLY LINEAR THERMODYNAMIC INTEGRATION

MIHALY MEZEI

Department of Physiology and Biophysics, Mount Sinai School of Medicine, Cuny, New York, NY 10029

(Received January 1993, accepted January 1993)

The free-energy simulation methodology is reviewed from the point of view of calculating large free-energy differences. The advantages of the nearly linear thermodynamic integration based on Gaussian quadrature are highlighted and its performance is characterized on systems ranging from the Lennard-Jones fluid to the A to B transition of DNA oligomers. A technique for optimizing the runlength at each quadrature point is given. Examples for the sensitivity of the calculated free energy to the atomic charges used are also presented.

KEY WORDS: Free-energy simulation, thermodynamic integration polynomial path, glycine, alanine dipeptide, DNA

INTRODUCTION

Free energy is the key quantity for the understanding of chemical equilibria, allowing the characterization of molecular associations, selectivities of associations and conformational preferences. Theoretical treatment of solvated macromolecules generally involve computer simulations, due to the multitude of effects determining their behaviour. The calculation of solvation free energies from computer simulations, however, has long been recognized as a computationally more demanding task than the calculation of structural properties or the internal energy. This critical status of the free energy is a consequence of its close relationship with the partition function as the general simulation methodologies developed over the last few decades owe their success to their ability to provide results without having to calculate the partition function.

However, the importance of the free energy in the understanding of chemical equilibria led to diverse and ultimately largely succesful attempts at its calculation. As the various methodologies were developed and tested, it became clear that reliable results with generally applicable methods require the adequate sampling of states that lie on a path connecting the initial and final systems in the configuration space. This feature brings an additional increase in the necessary effort for free energy calculations with increasing solvent complexity: not only does one need longer calculations for adequate estimation of the solute-solvent properties, the

number of such calculations will also increase as the path connecting the initial and final states in the configuration space becomes longer.

The purpose of this paper is to discuss methods that are aimed at interpolation over this path of ever increasing length and show that successful interpolation is indeed possible. As the state of the art of free energy simulations has been periodically reviewed in recent years [1-4] only a brief overview of the various methodologies will be given and emphasis will be placed on the nearly linear path and on the quadrature-based thermodynamic integration, a combination that has this interpolation capability. Finally, examples will be shown that demonstrate the interpolation techniques discussed.

FREE-ENERGY SIMULATION METHODOLOGY

Most free-energy simulation techniques are characterized by the path used to connect in the configuration space the two systems between which the free-energy difference is calculated and by the quantity chosen whose Boltzmann average is related to the free energy. For the understanding of the free-energy methodology it is important to keep in mind that these two issues are conceptually independent of each other although certain choices of paths are frequently associated with certain formalisms.

Choice of Path

As the free energy is a state function it is independent of the path over which it is calculated. This allows considerable latitude for the selection of the path over which the calculation is performed. The path is described by the introduction of a coupling parameter λ into the energy function. The various choices generally fall into the following two categories [1]:

$$E(\lambda, \mathbf{X}^N) = \lambda^k * E_1\ (\mathbf{X}^N) + (1 - \lambda)^k * E_0(\mathbf{X}^N) \text{ or} \tag{1}$$

$$E(\lambda, \mathbf{C}, \mathbf{X}^N) = E(\mathbf{C}(\lambda), \mathbf{X}^N). \tag{2}$$

Here E_0 and E_1 are the energy functions for the two systems of N atoms and λ is chosen in such away that $\lambda = 0$ and $\lambda = 1$ in (1) and (2) describe systems with energy function E_0 and E_1, respectively. $\mathbf{C}$ in (2) stands for the collection of potential parameters (including molecular geometries) that are continuous functions of λ with $\mathbf{C}(0)$ and $\mathbf{C}(1)$ representing systems 0 and 1, respectively. The path of (1) is operationally simpler than the path of (2) but it requires calculations on physically meaningless systems that are 'mathematical mixtures' of chemical systems. The path of (2) conforms better to chemical thinking since at each point along the path the calculation corresponds to a chemically meaningful (although not necessary stable or even existing) object.

Both choices can run into problems, though. The path of (1) has the potential problem of possible 'end-point catastrophe', i.e. the possible appearance of singularities at $\lambda = 0$ or 1 when particle creation or annihilation is involved. Such situations occur when the number of atoms in the two systems differ or an atom is mutated into an other of significantly different size or the conformation of the two systems are different enough that atoms are moved to positions previously unoccupied by the solute. As discussed below, this problem can be dealt with by

the use of $k > 1$ in (1) [1]. The path of (2), on the other hand, can lead through phase transitions causing numerical instabilities.

A further improvement has been recently found for paths of type (1): the polynomial path [5]. It generalizes (1) in such a way that different exponents will be used for different parts of the potential. A similar formalism was introduced in [6] and used to separate the creation of the cavity form the electrostatic charging process necessitated by the problems of the path of type (2) mentioned above. For a potential energy composed of terms of n different type, $t(i)$,

$$E(\mathbf{X}^N) = \sum_{i=1}^{n} E_{t(i)}(\mathbf{X}^N), \tag{3}$$

the polynomial coupling replaces (1) by

$$E(\lambda, \mathbf{X}^N) = \sum_{i=1}^{n} \lambda^{k(t(i))} * E_{1,t(i)} + (1 - \lambda)^{k(t(i))} * E_{0,t(i)}. \tag{4}$$

where $k(t(i))$ is the λ exponent to be used with the terms of type $t(i)$. A natural partiton of the energy is based on the exponent of the interatomic distances. The use of different λ exponents incurs negligible additional computational expense. The significance of (4) will be discussed below.

The independence of the free energy from the path over which it is calculated also imposes a numerical constraint over free-energy differences: their sum over a closed thermodynamic cycle has to be zero. This constraint can be used either for checking the numerical precison of the calculation (*vide infra*) or it can be used to eliminate the calculation of one step in the thermocycle. The latter choice led McCammon to the very succesful proposition to replace the calculation of the free energy of association with the calculation of free energies of mutations [7].

Choice of Formalism

Once the path is specified, the free-energy difference between the two states can be obtained in various ways, by exploiting various statistical thermodynamical identities.

Perturbation Method

The conceptually simplest formalism is the perturbation method (PM) [8–10] since it is based on a formula that follows directly from the partition function ratios:

$$\Delta A = -kT \ln \langle \exp[-(E_1 - E_0)/kT] \rangle_0 \tag{5}$$

where k is the Boltzmann constant, T is the absolute temperature and the symbol $\langle\ \rangle_0$ stands for the Boltzmann average of the quantity enclosed using E_0 as the energy in the Boltzmann factor. It is the method of choice for systems where the change is small and it is widely used indeed. The presence of the exponential in (5) warrants caution for larger changes since exponentiation drastically enlarges the range of the numbers to be averaged and therefore large numerical errors can occur [1, 11]. For example, if the energy difference $E_1 - E_0$ fluctuates in a range of 10 kcal/mol, the contributions to the averageing in (5) will vary over 6 orders of

magnitude. Note, that it is the *range* of energy differences that has to be kept small, not the differences themselves.

It should be pointed out that that the notion perturbation is used in several different contexts in statistical thermodynamics, a somewhat unfortunate situation giving rise easily to confusion. The point of departure of Zwanzig's paper [8] is the free-energy formula (5) but he then proceeds to an expansion into a power series of $1/kT$. For actual calculations, the expansion is truncated after a few terms. That usage is clearly different from the direct use of (5) since (5) is an exact expression. The perturbation notion has also been used in the free-energy literature in a general sense, to refer to calculations where a coupling parameter changes one system into an other in small steps, irrespective of the formalism used.

PM calculations have been enhanced in a variety of ways. For larger changes, the calculation can be broken up into a sequence of calculations between distinct states $E(\lambda_i)$ over some path (usually of type (2)) [12]. Non-Boltzmann sampling (usually called umbrella sampling), discussed below in conjunction with the probability ratio method, has been used by Torrie and Valleau [9] to enlarge the length of path sampled in a single simulation allowing the calculation of the free-energy difference between more different states in one step. A special case, called half-umbrella sampling, has been introduced by Scott and Lee [13] where the free-energy difference between states E_0 and E_1 is obtained from a run using $(E(\lambda_0) + E(\lambda_1))/2$ in the Boltzmann factor. An analogue of this, where a calculation with $E((\lambda_0 + \lambda_1)/2)$ is used to obtain the free-energy difference between $E(\lambda_0)$ and $E(\lambda_1)$ has been called double wide sampling. Based on a triangle inequality, it has been argued in [14] that the first version should be provide better sampling of the relevant part of the configurational space.

Thermodynamic Integration

Thermodynamic integration (TI) uses the expression of Kirkwood [15] that applies the fundamental theorem of calculus to replace the partition function with an expression that involves an additonal integration:

$$\Delta A = A_1 - A_0 = \int_0^1 \partial A(\lambda)/\partial\lambda \, d\lambda \tag{6}$$

$$= \int_0^1 \langle \partial E(\lambda)/\partial\lambda \rangle_\lambda \, d\lambda \tag{7}$$

where the symbol $\langle\rangle_\lambda$ stands for the Boltzmann average of the quantity enclosed using $E(\lambda)$ in the Boltzmann factor.

TI calculations can also provide directly the entropy change ΔS between the two states [1]:

$$T\Delta S = -\int_0^1 [\langle E(\lambda)\ \partial E(\lambda)/\partial\lambda \rangle_\lambda - \langle E(\lambda)\rangle_\lambda \langle \partial E(\lambda)/\partial\lambda \rangle_\lambda] \, d\lambda. \tag{8}$$

For a path of type (1) substitution of (1) into (8) yields

$$T\Delta S = -k \int_0^1 \{\lambda^{2k-1} [(\langle E_0^2 \rangle_\lambda - \langle E_0 \rangle_\lambda^2) + (\langle E_1^2 \rangle_\lambda - \langle E_1 \rangle_\lambda^2)] + \tag{9}$$

$$[\lambda(1-\lambda)]^{k-1}(\langle E_0\rangle_\lambda\langle E_1\rangle_\lambda) - (\langle E_0 E_1\rangle_\lambda)\}\, d\lambda$$

Evaluation of (8) or (9) requires negligible extra computational effort. A comparison between the $T\Delta S$ calculated with (8) or (9) and ΔE-ΔA, where ΔE is either calculated from separate simulations or extrapolated from the calculated $\langle E(\lambda)\rangle$'s to $\lambda = 0$ and 1, can give a consistency check on the simulation.

The integration can be carried out with a quadrature [16,17] or using the slow-growth method [18]. Numerical quadratures determine an integral by evaluating the integrand at a finite number of points. Well-known examples of this are the numerical integration with the trapezoid rule or with Simpson's rule. When the integral to be approximated is based on relatively few points, the best choice is the Gaussian quadrature that is known to minimize the quadrature error. Note that if an increase in the order of quadrature (i.e. number of quadrature points) is required, the higher order quadrature will use a new set of points (with the exception of $\lambda = 0.5$ for odd order quadratures). The work done for the lower order quadrature is not lost, though: it can be used to quantify the quadrature error by comparing the calculated integrand with the corresponding value of the Gaussian quadrature's fitting polynomial. Schlitter has introduced and tested an iterative quadrature scheme [19] that takes into account the variation in the statistical error of the integrand, but it has not yet been compared with the Gaussian quadrature. An alternative to quadratures, the slow-growth method varies λ (nearly) continuously during the simulation. Thus it essentially escapes quadrature error but at the expense of having larger errors in the integrand. This error can be reduced only by increasing the run length and varying the rate of change in λ along the path. The required lengths can be rather long. For example, Michell and McCammon showed that for a dipeptide-tripeptide transition 200 ps runs are the minimum [20] and the same minimum was found by Pearlman and Kollman for the methane-neopentane transition [21]. The qualification of the error incurred with the slow growth method is currently an active area [22–26]. Reliable estimate of the error is made difficult by the fact that the changes in the solvent environment are not necessarily spread out evenly over the transition: consider for example the mutation of an ion into an other that has different coordination number: the relaxation characteristics of the λ range where the solvation shell is being rearranged would be clearly quite different from the rest. The problem is compounded by the fact that the location of the λ-region where such transitions occur is not known beforehand.

TI with $k = 1$ in (1) is usually called linear TI and with $k > 1$ it has been referred to as nearly linear TI. The term 'nearly linear' has been introduced to distiguish it from the path of (2), referred to as a non-linear path. Nearly linear TI allows the calculation to avoid the endpoint catastrophe occurring when creation and/or annihilation of atoms is involved. In such cases linear TI leads to a so-called improper integral (i.e a definite integral where the integrand is singular at the endpoint(s)) for such systems and the other methods become numerically unstable. However, for a potential of the form $1/r^e$ the asymptotic behaviour of the integrand is known:

$$\langle \partial E(\lambda)/\partial\lambda\rangle_\lambda \quad \alpha \quad \lambda^{(kd/e)-1} \quad \text{as } \lambda \to 0 \tag{10}$$

where d is the dimensionality of the space [1] thus with a high enough k the integrand remains finite everywhere. In particular, a $1/r^{12}$ repulsion in three

dimensions requires $k \geqslant 4$. The special case for $d = 3$ was given earlier by Squire and Hoover [27] and by Mruzik, Abraham, Schreiber and Pound [16] and an integral transformation was derived from it to eliminate the singularity. The path described by (1) results in the same calculation as this integral transform when one of the systems is the ideal gas. However, the integral transform can not handle simultaneous creation and annihilation.

It has been suggested (and verified) in [5] that the polynomial path, described with (4), can partially counter the distorting effect of $k > 1$ in (1) by using the lowest possible exponent for each type of energy contribution. As a result, the integration will sample more evenly the [0, 1] λ interval and the result will be more precise with a given number of quadrature points or, equivalently, the same precision can be obtained with fewer quadrature points. Note, however that Schlitter has shown [19] that the statistical error at each quadrature point can be reduced by using higher than the minimum k value, indicating that there is a trade-off between quadrature error and statistical error and thus further improvement can possibly be obtained over the exponent-combination used in [5].

The other option for avoiding the endpoint catastrophe in TI is the use of a path of the form (2) as noted by Mezei and Beveridge [1] and by Cross [28]. Interestingly, Cross' parametrization of the Lennard Jones fluid reduced it to a polynomial path with $k = 13$ and 7 for the repulsion and dispersion term, respectively. The price paid is the generally less predictable behaviour of the integrand, requiring the use of more quadrature points.

Finite difference thermodynamic integration [29] combines TI and PM: the integral of (4) is evaluated by approximating the integrand at each quadrature points with a finite difference ratio over a small λ interval. The small change in the free energy, needed in the finite difference ratio is calculated with the perturbation method. As only small changes are required in λ, the perturbation method results will be reliable.

Probability Ratio Method

The solvation free energy can also be related to various distributions: the acceptance ratio method of Bennett [10], the overlap ratio method developed by Bennett [10] and Jacucci and Quirke [30] (shown to perform well in aqueous systems too [31]) and the probability ratio method all fall into this category. Voter introduced a variant of the acceptance ratio method that calculates large free-energy differences in one step, as long as the energy difference fluctuations are small (*vide supra*) [32].

The probability ratio method that exploits the relation between the Boltzmann factor and probability of occurrence, was first developed for the calculation of the free-energy profile (i.e. potential of mean force) along a given path [33], but also applied to the determination of free-energy differences by Mezei, Mehrotra and Beveridge [34]:

$$\Delta A = -kT \ln \left[(P(\lambda)_{\lambda=1}/V_1)/(P(\lambda)_{\lambda=0}/V_0) \right], \tag{11}$$

where $P(\lambda)$ is the Boltzmann probability of the system to be at the intermediate stage λ when λ is also a variable during the simulation and V_0, V_1 represent the configuration space volume corresponding to the $\lambda = 0$ and 1 state, respectively. Valleau, Patey and Torrie have recognized that (11) translates small free-energy

differences into large ratios in the probability of sampling and thus this method requires non-Boltzmann ("umbrella") sampling (US) with a modified Hamiltonian, $E'(\mathbf{X}^N, R(\lambda))$ [9, 33] to sample λ values whose probability is small:

$$E'(\mathbf{X}^N, R(\lambda)) = E(\mathbf{X}^N, R(\lambda)) + E_w(\lambda). \tag{12}$$

The Boltzmann average $\langle Q \rangle_B$ of any quantity Q can be recovered as

$$\langle Q \rangle_B = \langle Q\ w(\lambda) \rangle_w / \langle w(\lambda) \rangle_w \text{ where} \tag{13}$$

$$w(\lambda) = \exp\ [E_w(\lambda)/kT] \tag{14}$$

and $\langle\ \rangle_w$ implies configurational average using the modified Hamiltonian given by (12). Notice that here the use of US is essential for the calculation, not just a performance enhancer, as with the PM. The determination of $E_w(\lambda)$ proved to be a serious obstacle, though. This obstacle was reduced significantly by the introduction of adaptive techniques, based on the fact that the best choice for $E_w(\lambda)$ is $W(\lambda)$. This adaptive umbrella sampling was used by Paine and Scheraga [35] to obtain the gas-phase conformational free-energy map of the alanine dipeptide and Mezei recalculated the free-energy difference between the C_7 and α_R conformations of the alanine dipetide in aqueous solution [11, 36]. The adaptive umbrella sampling method proved to be significantly more reliable than the use of the harmonic weighting function on the dimethyl phosphate anion [37]: the closure error over a thermodynamic cycle consisting of three distinct solute conformations was 8 kcal/mol with empirically determined weighting functions and 0.6 kcal/mol with the adaptive method. For the aqueous system several additional problems have to be dealt with: matching of iterations with large statistical noise, recognition of equilibration phase, guiding the simulation to undersampled regions and others. The method not only provides improved computational efficiency but is inherently self-checking.

Methods Without Coupling Parameter

The free energy can also be obtained from grand-canonical ensemble simulations [38, 39]. Here the chemical potential is kept fixed and the density is obtained at the end of the calculation as an ensemble average. It has been successful in obtaining the excess free energy of liquid water with a cavity-biased insertion technique [40, 41]. For systems involving larger molecules the modeling of the fluctuating number of molecules becomes increasingly difficult as the liquid will not contain molecular size cavities. It has been recently proposed in the context of the related Gibbs ensemble simulations [42] that "growing" the molecule fragment by fragment could overcome this.difficulty [43], although this approach is a tacit reintroduction of the coupling parameter idea. Widom's particle insertion method [44] is closely related to the use of the grand-canonical ensemble and it faces similar difficulties at higher densities.

Jayaram and Beveridge have recently introduced an approximate formula that relates the solvation free energy to the energy range sampled during the simulation and obtained quite good results [45].

Digression

An analysis of the methods described above reveals that both the perturbation method and the probability ratio method intrinsicly require the complete sampling of the path connecting the two systems in the configuration space and the same is true for the slow-growth variant of thermodynamic integration while the quadrature-based thermodynamic integration requires knowledge only about states lying at the quadrature points. The difficulty in obtaining adequate assessment of the error of free energies culculated with PM or slow-growth TI has been compounded by the fact that the agreement between two calculations performed in the opposite direction (i.e. low hysteresis) is not enough to ensure that the error is actually low. Newer techniques [22–26] for estimating the error of free energies calculated with slow-growth TI require simplifying assumptions that are not necessarily met. On the other hand, the error estimates on the individual quadrature points provide a reliable error estimate on the free energy calculated. Thus the crucial question for the success of quadratures is the magnitude of the quadrature error. It has been shown [46] that using the path of (1) with $k = 1$ the integrand in (5) is monotonous. Similar arguments can show that for $k > 1$ both $\langle E_0 \rangle_\lambda$ and $\langle E_1 \rangle_\lambda$ are monotonous functions of λ, thus it is reasonable to expect that the integrand of (7) is still going to be "well behaved" (i.e. not have many unexpected extrema). Paths described by (2), however, may contain an unspecified number of oscillations and inflexions. The difficulty of adequate sampling over the path of (2) has been recently highlighted by Pearlman and Kollman [21] who showed that free-energy differences over transitions including bond-streching can incur unexpectedly large errors (although they showed how to correct for them with additional calculations). An additional advantage of the path of (1) is that when the reference system is the ideal gas it describes a transcritical path and thus avoids any possible numerical instability that may arise when a phase transition is encountered along the path – a distinct possibility with the use of (2).

The arguments presented above – interpolation capability, reliable error estimates, nearly monotonous transcritical path – lead to the conclusion that Gaussian quadrature based TI over the path of (1) or (4) is the method of choice for large changes and that its advantage increases with the system size. Calculations described below will illustrate some of these points.

CALCULATIONS USING NEARLY LINEAR TI

Lennard-Jones Fluid

Calculation of the free energy of the Lennard-Jones fluid near its triple point over the path of (1) showed that 5-point Gaussian quadratures can give the excess free energy of the fluid to good precision and largely independently of the k value (as long as $k \geqslant 4$) [14]. The estimated errors (2 S.D.) were 0.004% or less.

Liquid Water

The nearly linear TI with $k = 4$ was applied to the SPC [47] and MCY [48] water models using 64 waters under face-centered cubic periodic boundary conditions at room-temperature experimental density [46, erratum]. The integration was based on

Table 1 Free-energy differences calculated for liquid water.

Runlength: Sys_1	Sys_0	n_q	*k(1)*	*k(6)*	*k(12)*	*300K* ΔA	*1000K* ΔA
SPC	IG	5	4	4	4	−5.67 ± 0.06	−5.65 ± 0.05
SPC	IG	8	4	4	4	−5.68 ± 0.06	−5.70 ± 0.03
MCY	IG	5	4	4	4	−4.03 ± 0.05	−4.04 ± 0.03
SPC	MCY	3	1	1	1	−1.66 ± 0.04	−1.63 ± 0.04
SPC	IG	3	4	4	4	−5.61 ± 0.12	−5.61 ± 0.04
SPC	IG	3	2	2	4	−5.73 ± 0.06	−5.69 ± 0.05
MCY	IG	3	4	4	4	−3.97 ± 0.03	−4.04 ± 0.16
TIP4P	IG	3	4	4	4	−5.43 ● 0.08	−5.42 ± 0.06
TIP4P	IG	3	3	2	4	−5.46 ± 0.07	−5.45 ± 0.03

Legend: a. The symbol K refers to 1000 force-biased [63] Metropolis Monte Carlo [64] steps; b. ΔA represents the free-energy difference between Sys_1 and Sys_0 in kcal/mol; c. n_q is the number of quadrature points; d. $k(i)$ are the λ exponents used with terms involving $1/r^i$; e. The results in the top half are from Reference 46 and in the bottom half from Reference 5; f. Energies are in kcal/mol; g. The SPC-MCY run was only 600K long.

5-point quadratures. An independent calculation of the MCY-SPC free-energy difference using linear TI based on a 3-point quadrature allowed a consistency check on a thermodynamic cycle. The results, given in Table 1, show that the calculations converge very quickly and the error estimates from the individual runs are confirmed by the small closure error on the IG-MCY-SPC-IG cycle: under 0.1 kcal/mol. The error estimates on individual free energies were obtained by the method of batch means [49, 50] and represent 95% confidence intervals (2 S.D.).

These systems were also used recently to demonstrate the improvements possible with the polynomial path [5]. Table 1 also shows the results of 3-point quadrature calculations with the polynomial path and with the path of (1) on the SPC and TIP4P [51] water: the polynomial path results agreed with the 8-point quadrature results showing that precise results can be obtained with 3-point quadrature. The curvature of the integrand was also significantly reduced with the polynomial path for both models, explaining the success of the low-order quadrature.

Methane Solvation in Water

Fleischman and Zichi have studied the perfomance of the nearly linear TI on the solvation of methane in water [52, 53]. Several exponents were tried but the result was found insensitive to the choice of exponents (as long as it was at least 4). The feasibility of using only a small number of quadrature points has been verified: a 5-point Gaussian quadrature reproduced the TI integrand calculated over 40 points. As the calculated solvation free energy is rather small (1.23 kcal/mol calculated, 2.00 kcal/mol experimental) rather long runs were required to reduce the statistical noise below the calculate value. An increase in the statistical error of the integrand was noted around $\lambda = 0.6$.

Glycine Zwitterion Formation

Glycine forms zwitterion in water at neutral pH. In a recent work this free energy of the zwitterion formation was decomposed into two principal contributions: the

difference between the energy of formation of the two molecules and the difference between the free-energy contribution of the solvent in solution [54]. The calculations used prefixed geometries for both the neutral and the zwitterionic form. The gas-phase energy difference has been estimated by ab initio calculations up to 2nd order Moller-Plesset level as 22.8 kcal/mol, favoring the neutral species, in good agreement with eariler work of Langlet, Caillet, Evleth and Kassab [55] at the 6-31G* level with geometry optimization (about 20 kcal/mol). Next the difference between the solvation free energies was calculated by thermodynamic integration over the path of (1). The solute-solvent interactions were described by the AMBER force field [56] and the TIP4P water model [51] was used for the water-water interactions and atomic (partial) charges were derived from STO-3G population analysis. A 5-point quadrature and λ exponent $k = 4$, was used, resulting in a solvation free-energy difference of 32.0 ± 2.3 kcal/mol, favoring the zwitterion. Table 2 gives the calculated quadrature point results. Combined with the ab-initio calculations, this yields −9.2 ± 2.3 kcal/mol for the free energy of zwitterion formation. To test the effect of the atomic charge selection procedure, a new set was also calculated, this time based on 6-311G** population analyis and for both forms the solvation free-energy differences of between the two different charge models were calculated with linear TI using 3-point quadratures. The new charge sets reduced the free energy of solvation of both species, by, 3.8 ± 0.2 kcal/mol and by 9.2 ± 0.2 kcal/mol for the neutral and zwitterionic form, respectively. Thus the 6-311G** charge set gives −14.6 ± 2.3 kcal/mol for the free energy of zwitterion formation. While it is encouraging that the final results with the two charge sets encompass the experimental estimate of −11 kcal/mol [54] the 5.4 kcal/mol difference beteen the two charge sets is surprisingly large, given the fact that the charges were obtained with a 'reasonable' procedure and the comparison of neutral and zwitterionic forms could have been expected to provide better cancellation of errors.

Alanine Dipeptide Conformational Free Energies

The C_7 and α_R conformations of the alanine dipeptide were used to compare the nearly linear TI, finite difference TI and the probability ratio method with adaptive umbrella sampling [11], the latter two over a path of type (2) involving the linear movement of the dipeptide atoms. The dipeptide was modeled with the OPLS potential [57] surrounded by 215 TIP4P water [51] in a FCC simulation cell, at 298K and experimental density. The free-energy differences, all favoring α_R, were obtained as 11.5 ± 3.0 kcal/mol with a 5-point nearly linear TI (the 12.6 kcal/mol quoted in [11] was the result after only 3000K long runs), 8.9 ± 4.6 kcal/mol with a 5-point finite difference TI and 12.8 ± (~1.5) kcal/mol with the probability ratio method. Table 2 gives the calculated quadrature point results for the nearly linear TI. The finite difference TI shows the largest statistical error and its deviation from the other two indicates that the quadrature error is also likely to be large - a demonstration of the 'unpredictable' behaviour of the integrand over paths of type (2).

These free-energy differences are much larger than what is expected experimentally and what was calculated earlier [34, 36] using the potential library of Clementi and coworkers [58] and the MCY water model [48]. In the Clementi model the charges are separately calculated for each conformation while the OPLS-based calculations described above used conformation independent charges. To test the importance of changing the charges with conformation, the Clementi-based calculations were

Table 2 TI integrands over the path of (1) with $k = 4$

λ	$c(\lambda)$	*Glycine* $N \rightarrow +/-$	*Dipeptide* $C_7 \rightarrow \alpha_R$	*DNA (GGCC)* $A \rightarrow B$	*DNA (AATT)* $A \rightarrow B$
0.04691	0.118463	40.2 ± 1.6	−205.0 ± 7.	−5809 ± 63	1542 ± 20
0.23076	0.239314	14.0 ± 0.4	−62.5 ± 3.	−2640 ± 33	483 ± 15
0.5	0.284444	−2.5 ± 0.6	−0.9 ± 2.	−105 ± 10	12 ± 6
0.76924	0.239314	−52.7 ± 1.9	48.1 ± 3.	2417 ± 31	−404 ± 14
0.95309	0.118463	−52.6 ± 1.4	134.1 ± 4.	5523 ± 33	−1486 ± 30
ΔA:		−32.0 ± 3.2	−11.5 ± 3	−118 ± 35	29 ± 16
Number of waters:		215	215	793	793
Runlength per λ:		3000K	4000K	4000K	4000K

Legend: a. $c(\lambda)$ is the quadrature coefficient for the quadrature point λ; b. energies are in kcal/mol; c. errors represent 95% confidence interval (2 S.D.); d. ΔA is the calculated free-energy difference.

repeated with the charges fixed at the α_R values. The new calculations (adaptive US in two steps, 6000K total runlength) increased the free-energy difference from 1.8 kcal/mol to 7.4 kcal/mol, showing that keeping the charges fixed can indeed change the result significantly. The size of the increase is all the more remarkable since charge differences averaged only 0.011 *e* and the largest change in the charges was 0.033 *e*. The difference in the change of the dipole moments when the c_7 structure is transformed into the α_R structure is not large enough to explain these differences either: the dipole moment of the c_7 structure was 0.116 eÅ and the dipole moment of the α_R structure was 1.74 eÅ and 2.44 eÅ, using the α_R and C_7 charges, respectively.

Nucleic Acid Conformational Free Energies

The solvation free-energy differences between the canonical *A* and *B* conformations of two nucleic acid tetramers have been also calculated using the nearly linear TI with five-point quadratures [59].

The first calculation involved the DNA duplex 5′-GGCC-3′ described with the potential library of Clementi *et al.* [58] and the MCY water model [48] while the second calculation was done on the 5′-AATT-3′ tetramer described by the AMBER force field [56] in TIP3P water [51]. For both calculations the charge on the phosphate group was decreased by 0.75e, to represent the effect of condensed counterions according to the theory of Manning [60] and no explicit counterions were modeled. The simulations used periodic boundary conditions in a hexagonal prizm containing 800 waters at experimental density and 298K temperature. The solute-water interactions were calculated with the minimum image convention while the water-water interactions were truncated with a 7.00 Å spherical cutoff. The results are shown in Table 2. As the calculated free-energy differences are too large it is clear that both models require further refinements. It is interesting to note that on this system the model that used different charges for the two conformations gave the larger free-energy difference, contrary to the dipeptide case.

Discussion of the Precision

These calculations show that nearly linear TI is capable of treating changes that are significantly larger than the ones currently being attempted. As the run lengths

were kept about the same the calculated error estimates increased with the size and complexity of the system. The DNA calculations have especially large errors but these calculations were really quite short in view of the systems involved (the computational effort spent at each quadrature point is roughly equivalent to 5 ps molecular dynamics runs each). The particularly large error of the Clementi-based calculation can be partially attributed to the excessive ($\sim$7000 atm) pressure of the MCY water [61] since that results in more solute-solvent overlap during the creation phase. Also, it should be kept in mind that these errors are only a few percent of the solute-solvent interaction energies, thus they basically represent the precision of a simulation involving such a large solute and such a large number of solvent molecules.

However, as TI is based on the ensemble average of energies at a given state, the error can be expected to decrease at the usual $1/N^{1/2}$ with the length of the run N. This behaviour of the error has recently been confirmed on the IG – $[Na^+]_{aq}$ – $[Li^+]_{aq}$ – IG thermocycle [62] where both the calculated error estimate and the closure error decreased steadily with the increase of the length of the calculations.

The error estimates at the quadrature points, shown in Table 2, are consistently larger at λ regions where creation/annihilation is occurring, in accord with Schlitter's result [19]. Thus improved precision with a given computational effort can be obtained if the runlengths at each quadrature point are optimized - see the Appendix.

The precision of the nearly linear TI can also be increased without significant increase in the computational effort by the use of the polynomial path, as shown on liquid water.

CONCLUSIONS

It has been demonstrated that the nearly linear TI with Gaussian quadratures is a robust method for the calculation of large solvation free energies with reliable error estimates. Its efficiency can be further improved by the use of polynomial path and by a new runlength optimization technique at each quadrature point, as described in the Appendix. The calculations discussed have also provided two examples where the calculated free energy differences have shown large dependence on the atomic chaerges used. These results underscore the importance of the proper selection of the atomic charges in the potential.

Acknowledgements

This work was supported by NIH grant #GM24914 and NSF grant #CHE-8203501 to Prof. D.L. Beveridge, by RCMI grant #SRC5G12RRO307 from NIH to Hunter College, by several CUNY/PSC grants while the author was working at Hunter College, and by NIH grant #R55-GM43500. Computing resources were provided by the City University of New York, University Computing Center. Conversations with Drs. M. Pettitt and S. Fleischman are also acknowledged and thanks are also due to Dr. Fleischman for a copy of Reference 53.

Note added in Proof:

Recent calculations showed that the polynomial TI works equally well for solvation free energy calculations (H. Resat and M. Mezei, in preparation).

References

[1] M. Mezei and D.L. Beveridge, "Free energy simulations," *Ann. Acad. Sci. N.Y.*, **482**, 1 (1986).

[2] D. Frenkel, "Free-energy computations and first-order phase transitions," in *Molecular dynamics simulation on statistical mechanical systems*, Soc. Italiana di Fisica, Bologna, (1986).

[3] D.L. Beveridge, and F.M. DiCapua, "Free energy via molecular simulation: applications to chemical and biomolecular systems," *Annu. Rev. Biophys. Chem.*, **18**, 431 (1989).

[4] T.P. Straatsma, and J.A. McCammon, "Computational Alchemy," *Annu. Rev. Phys. Chem.*, **43**, 407 (1992).

[5] M. Mezei, "Polynomial path for the calculation of liquid state free energies from computer simulations tested on liquid water," *J. Comp. Chem.*, **13**, 651 (1992).

[6] J. Hermans, A. Pathiaseril, and A. Anderson, "Excess free energy of liquids from molecular dynamics simulations. Application to water models," *J. Am. Chem. Soc.*, **110**, 5982 (1988).

[7] A.J. McCammon, "Computer-aided molecular design," *Science*, **238**, 486 (1987).

[8] R.W. Zwanzig, "High-temperature equation of state by perturbation method. I. nonpolar gases," *J. Chem. Phys.*, **22**, 1420 (1954).

[9] G.M. Torrie, and J.P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling," *J. Comp. Phys.*, **23**, 187 (1977).

[10] C.H. Bennet, "Efficient estimation of free-energy differences from Monte Carlo data," *J. Comp. Phys.*, **22**, 245 (1976).

[11] M. Mezei, "Evaluation of the adaptive umbrella sampling Method," *Molecular Simulation*, **3**, 301 (1989).

[12] W.L. Jorgensen, and C. Ravimohan, "Monte Carlo simulation of differences in free energies of hydration," *J. Chem. Phys.*, **83**, 3050 (1985).

[13] H.L. Scott, and C.Y. Lee, "The surface tension of water: a Monte Carlo calculation using an umbrella sampling algorithm," *J. Chem. Phys.*, **73**, 4591 (1980).

[14] M. Mezei, "Direct calculation of excess free energy of the dense Lennard-Jones fluid with nonlinear thermodynamic integration," *Molecular Simulation*, **2**, 201 (1988).

[15] J.G. Kirkwood, "Statistical mechanics of fluid mixtures," *J. Chem. Phys.*, **3**, 300 (1935).

[16] M.R. Mruzik, F.F. Abraham, D.E. Schreiber, and G.M. Pound, "A Monte Carlo study of ion-water clusters,", *J. Chem. Phys.*, **64**, 481 (1876).

[17] M. Mezei, S. Swaminathan, and D.L. Beveridge, "Ab initio calculation of the free energy of liquid water," *J. Am. Chem. Soc.* **100**, 3255 (1978).

[18] H.J.C. Berendsen, J.P.M. Postma, and W.F. van Gunsteren, "Statistical mechanics and molecular dynamics: the calculation of free energy," in *Molecular Dynamics and Protein Structure*, J. Hermans, ed., Polycrystal Book Service, Illinois, p. 43 (1985).

[19] J. Schlitter, "Methods for minimizing errors in linear thermodynamic integration," *Molecular Simulation*, **7**, 105 (1991).

[20] M.J. Mitchell, and J.A. McCammon, "Free-energy difference calculation by thermodynamic integration: difficulties in obtaining a precise value," *J. Comp. Chem.*, **12**, 271 (1991).

[21] D.A. Pearlman, and P.A. Kollman, "The overlooked bond-stretching contribution in free-energy perturbation calculations," *J. Chem. Phys.*, **94**, 4532 (1991).

[22] R.H. Wood, "Estimation of errors in free-energy calculations due to the lag between the Hamiltonian and the system configuration," *J. Phys. Chem.*, **95**, 4838 (1991).

[23] R.H. Wood, W.C.F Muhlbauer, and P.T. Thompson, "Systematic errors in free-energy-perturbation calculations due to finite sample size: sample size hysteresis," *J. Phys. Chem.*, **95**, 6670 (1991).

[24] J. Hermans, "A simple analysis of noise and hysteresis in free-energy simulations," *J. Phys. Chem.*, **95**, 9029 (1991).

[25] J. Hermans, R.H. Yun, and A.G. Anderson, "Precision of free energies calculated by molecular dynamics simulations of peptides in solution," *J. Comp. Phys.*, **13**, 429 (1992).

[26] J. Schlitter, and D. Husmeier, "System relaxation and thermodynamic integration," *Molecular Simulation*, **8**, 285 (1992).

[27] D.R. Squire and W.G. Hoover, "Monte Carlo simulation of vacancies in rare-gas crystals," *J. Chem. Phys.*, **50**, 701 (1969).

[28] A. Cross, "Influence of Hamiltonian parameterization on convergence of Kirkwood free energy calculations," *Chem. Phys. Letters*, **128**, 98 (1986).

[29] M. Mezei, "The finite difference thermodynamic integration, tested on calculating the hydration free-energy difference between acetone and dimethylamine in water," *J. Chem. Phys.*, **86**, 7084 (1987).
[30] G. Jacucci and N. Quirke, "Monte Carlo calculation of the free-energy difference between hard and soft core diatomic liquids," *Molec Phys.*, **40**, 1005 (1980).
[31] M. Mezei, "Test of the overlap ratio method on the calculation of the aqueous hydration free-energy difference between acetone and dimethyl amine." *Molec. Phys.*, **65**, 219 (1988).
[32] A.F. Voter, "A Monte Carlo method for determining free-energy differences and transition state theory rate constants," *J. Chem. Phys.*, **82**, 1890 (1985).
[33] G. Patey, and J.P. Valleau, "Monte-Carlo method for obtaining the interionic potential of mean force in ionic solution," *J. Chem. Phys.*, **63**, 2334 (1975).
[34] M. Mezei, P.K. Mehrotra and D.L. Beveridge, "Monte Carlo determination of the free energy and internal energy of hydration for the ala dipeptide," *J. Am. Chem. Soc.*, **107**, 2239 (1985).
[35] G.M. Paine and H.A. Scheraga, "Prediction of the native conformation of a polypeptide by a statistical-mechanical procedure. I. Backbone structure of enkephalin," *Biopolymers*, **24**, 1391 (1985).
[36] M. Mezei, "Adaptive umbrella sampling: self-consistent determination of the non-Boltzmann bias," *J. Comp. Phys.*, **68**, 237 (1987).
[37] B. Jayaram, M. Mezei and D.L. Beveridge, "Conformational stabiliy of dimethylphosphate anion in water: liquid state free-energy simulations," *J. Am. Chem. Soc.*, **110**, 1691 (1987).
[38] G.E. Norman, and V.S. Filinov, *High. Temp. U.S.S.R.*, **7**, 219 (1969).
[39] D.J. Adams, "Grand canonical ensemble Monte Carlo for a Lennard-Jones fluid," *Mol. Phys.*, **29**, 307 (1975).
[40] M. Mezei, "A cavity-biased (T,V,μ) Monte Carlo method for the computer simulation of fluids." *Molec. Phys.* **40**, 901 (1980).
[41] M. Mezei, "Grand-canonical ensemble Monte Carlo study of dense liquids: Lennard-Jones, soft spheres and water," *Molec. Phys.* **61**, 565 (1987); erratum: **67**, 1207 (1989).
[42] A.Z. Panagiotopoulos, "Direct determination of phase coexistence properties of fluids by Monte Carlo simulation in a new ensemble," *Molec. Phys.*, **61**, 813 (1987).
[43] J.J. De Pablo and J.M. Prausnitz, "Phase equilibria for fluid mixtures from Monte Carlo simulations," *Fluid Phase Equil.*, **56**, 177 (1989).
[44] B. Widom, "Some topics in the theory of fluids," *J. Chem. Phys.*, **39**, 2808 (1963).
[45] B. Jayaram, and D.L. Beveridge, "A simple method to estimate free energy from a molecular simulation: renormalization on the unit interval," *J. Phys. Chem.*, **94**, 7288 (1990).
[46] M. Mezei, "Excess free energy of different water models computed by Monte Carlo methods," *Molec. Phys.*, **47**, 1307 (1982); Erratum: **67**, 1205 (1989).
[47] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren and J. Hermans, "Interaction models for water in relation to protein hydration," in *Intermolecular Forces*, B. Pullman, ed., Reidel (1981).
[48] O. Matsuoka, E. Clementi and M. Yoshemine, "CI study of the water dimer potential surface," *J. Chem. Phys.*, **64**, 1351 (1976).
[49] R.B. Blackman, and J.W. Tuckey, "The Measurement of power spectra," Dover (1958).
[50] W.W. Wood, "Monte Carlo studies of simple liquid models," in *Physics of Simple Liquids*, H.N.V. Temperley, J.S. Rowlinson, and G.S. Rushbrooke, eds., North Holland, (1968).
[51] W.L. Jorgensen, J. Chandrashekar, J.D. Madura, R. Impey and M.L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, **79**, 926 (1983).
[52] S.H. Fleischman, and D.A. Zichi, "Free-energy simulations of methane solvation: a study of integrand convergence properties using thermodynamic integration," *J. Chim. Phys.*, **88**, 2617 (1991).
[53] S.H. Fleischman, and D.A. Zichi, "Growing methane: simulations of aqueous methane solvation," preprint.
[54] M. Mezei, "Free-energy computer simulations for the study of proton transfer in solutions," in *Proton transfer in hydrogen bonded systems*, T. Bountis, ed., Plenum, NY (1992).
[55] J. Langlet, J. Caillet, E. Evleth, and E. Kassab, "Theoretical study of intermolecular proton transfer in glycine," in *Modeling of molecular structures and properties*, J.L. Rivail, Ed., (*Stud. Phys. Theor. Chem.*), **71**, 345 (1990), Elsevier, Amsterdam.
[56] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, Jr., and P.K. Weiner, "A new force field for molecular mechanical simulation of nucleic acids and proteins," *J. Am. Chem. Soc.* **106**, 765 (1984).

[57] W.L. Jorgensen and C.J. Swenson, "Optimized intermolecular functions for amides and peptides. Hydration of amides," *J. Am. Chem. Soc.*, **107**, 1489 (1985).
[58] G. Corongiu, and E. Clementi, "Intramolecular and intermolecular interactions for deriving chemical formulae and for simulate complex chemical systems," *Gazz. Chim. Ital.*, **108**, 273 (1978).
[59] M. Mezei, and D.L. Beveridge, unpublished results (1986).
[60] J. Manning, "Counterion binding in polyelectrolyte theory," *Acc. Chem. Res.*, **12**, 443 (1979).
[61] M. Mezei, and D.L. Beveridge, "Further quasicomponent distribution function analysis of liquid water. Temperature dependence of the results," *J. Chem. Phys.*, **76**, 593 (1982).
[62] P.V. Maye and M. Mezei, to be published.
[63] N.A. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, "Equation of State Calculation by Fast Computing Machines," *J. Chem. Phys.*, **21**, 1087 (1953).
[64] M. Rao, C.S. Pangali and B.J. Berne, "On the Force Bias Monte Carlo Simulation of Water: Methodology, Optimization and Comparison with Molecular Dynamics," *Mol. Phys.*, **37**, 1779 (1979).

APPENDIX

Optimization of the runlength at the quadrature points

If the standard deviation Δ_i of each quadrature contribution depends on the runlength N_i *as*

$$\Delta_i = p_i / N_i^{1/2} \tag{15}$$

then the error of the free energy Δ can be obtained from

$$\Delta^2 = \sum_i c(\lambda_i) * p_i^2 / N_i \tag{16}$$

where $c(\lambda_i)$ is the quadrature coefficient for the i-th point since standard deviation squares are additive. The proportionality constant p_i can be inferred from initial data or, near the creation/annihilation region, as

$$p_i \alpha \lambda_i^{\{kd/2e\}-1}. \tag{17}$$

(17) follows from the assymptotic formula of Schlitter [19] according to which the divergence of Δ_i^2 is one order stronger than the divergence of the TI integrand. Using the Lagrange multiplier method, minimization of (16), subject to the constraint

$$N_t = \sum_i N_i, \tag{18}$$

leads to the optimal condition

$$N_1 \alpha c(\lambda_i)^{1/2} * p_i. \tag{19}$$

An important feature of this approach is that it allows the retention of the Gaussian quadrature, resulting in a procedure that minimizes the statistical error and the quadrature error at the same time. Its performance, however, has to be assessed with actual calculations and compared with other approaches, such as in [19].