A simple technique to estimate partition functions and equilibrium constants from Monte Carlo simulations

Michal Vieth
Department of Chemistry, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, California 92037

Andrzej Kolinski
Department of Chemistry, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, California 92037 and Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

Jeffrey Skolnick[a)]
Department of Molecular Biology, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, California 92037

(Received 14 November 1994; accepted 11 January 1995)

A combined Monte Carlo (MC) simulation-statistical mechanical treatment is proposed to calculate the internal partition function and equilibrium constant. The method has been applied to a number of one and multidimensional analytical functions. When sampling is incomplete, various factorization approximations for estimating the partition function are discussed. The resulting errors are smaller when the ratios of the partition functions are calculated (as in the determination of equilibrium constants) as opposed to the partition function itself. © *1995 American Institute of Physics.*

INTRODUCTION

In many chemical or biological systems,[1–3] one desires to calculate the equilibrium constants between different forms or aggregates of a given molecule. This could, of course, be obtained from the experimentally determined concentration dependence of the population of different multimeric states. Here, we provide an alternative combined approach [Monte Carlo (MC)—statistical mechanical] that is, in principle, exact for a given energy function. The underlying idea of this treatment is to use a computer simulation to provide the variables for the statistical treatment. In practice, the relationship to experiment depends upon the quality of the potential and exhaustiveness of the MC sampling procedure. In what follows, we describe the analytical treatment, along with an application of this formalism to the computation of the partition functions for a set of simple energy functions presented in Table I. The implications of the results, together with approximations that are used when computational resources limit the sampling, are also described.

METHOD

For the sake of simplicity, let us suppose that we want to calculate the equilibrium constant of an imaginary particle which may be in two entirely distinct energy minima A and B, separated by large barriers

$$A \leftrightarrow B. \tag{1}$$

From a statistical mechanical point of view, the equilibrium constant between A and B is described as[4–6]

$$K = Z_A / Z_B, \tag{2}$$

where Z_A, Z_B are the partition functions for forms A and B, respectively. Thus the problem of calculating the equilibrium constant reduces to the calculation of the partition functions for both subsystems. Following Mayer and Mayer,[4,5,7] a partition function for the molecule can be written as the product of the configurational term and the integration over the momenta degrees of freedom. The integration over configuration space can be expressed as the product of the volumes available to the atoms in the molecule. In subsequent discussion, we will concentrate on the calculation of the configurational partition function (configurational integral). In the calculation of the equilibrium constant, the integrals over momenta degrees of freedom in the numerator and denominator will cancel.[5,7]

Exact treatment

For a system, where Boltzmann statistics applies, the probability of seeing a particular conformation inside an N dimensional, infinitesmal volume element $dV_{\rm ac}$, centered around $\mathbf{r}_0 = \{x_i^0,\ i = 1 \cdots N\}$ is proportional to the energy of the state $E(\mathbf{r}_0)$ and is given by[4,8–10]

$$P(\mathbf{r}_0) = \frac{\exp[-E(\mathbf{r}_0)/kT]}{Z_{\rm int}}\, dV_{\rm ac} \tag{3}$$

$dV_{\rm ac}$ is the product of the volume elements accessible to each of the N coordinates (x_i) within a given tolerance db. $P(\mathbf{r}_0)$ is the probability that each coordinate x_i is in the region between $x_i^0 - 1/2\ db$ and $x_i^0 + 1/2\ db$ (and thus each coordinate has an infinitesmal volume element db accessible to it). $dV_{\rm ac}$ is connected to the definition of the $P(\mathbf{r}_0)$ and the discretization db in the following way:

[a)]To whom correspondence should be addressed.

$$P(r_0)=P\begin{pmatrix} x_1^0-0.5\ db<x_1^0<x_1^0+0.5\ db \\ x_2^0-0.5\ db<x_2^0<x_2^0+0.5\ db \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_N^0-0.5\ db<x_N^0<x_N^0+0.5\ db \end{pmatrix};$$
$$dV_{\rm ac}=db^N. \tag{4}$$

In what follows, we approximate db by Δb, the finite difference approximation.

In the Monte Carlo[11,12] method, the canonical distribution of states is obtained by a Markovian sequence in which the probabilities between two conformational states $\mathbf{r}$ and $\mathbf{r}'$ are given by

$$\frac{P(\mathbf{r}')}{P(\mathbf{r})}=\frac{\exp[-E(\mathbf{r}')/kT]}{\exp[-E(\mathbf{r})/kT]}. \tag{5}$$

Thus by calculating the fraction of time a system spends in a given state ($\mathbf{r}$), a dynamic Monte Carlo method provides $P(\mathbf{r})$. From Eq. (3), we get

$$Z_{\rm int}=\exp[-E(\mathbf{r}_0)/kT]\ \frac{dV_{\rm ac}}{P(\mathbf{r}_0)},$$
$$\mathbf{r}_0=(x_1^0,x_2^0,\ldots,x_N^0). \tag{6}$$

Note that $\mathbf{r}_0$ can denote any conformational state. However, in what follows, due to the better statistics, the most probable state is used.

Approximate treatments in the poor sampling case

The extraction of $P(\mathbf{r}_0)$, the probability that all N degrees of freedom are simultaneously in the state $\mathbf{r}_0$, is crucial to the evaluation of Eq. (6). As the number of degrees of freedom increases, the longer the simulation time is required for $P(\mathbf{r}_0)$ to converge. In the case of a macromolecular system dissolved in a solvent, where the number of degrees of freedom can be very large (on the order of 10^4), the convergence of the corresponding probabilities becomes prohibitively time consuming. Furthermore, the time required to complete one Monte Carlo cycle is proportional to the number of degrees of freedom. Thus for such systems, one needs a method to estimate $P(\mathbf{r}_0)$. In what follows, we propose three factorization approximations. The first estimates $P(\mathbf{r}_0)$ as the product of N independent probabilities $P_{i,\max}(x_i)$ that each coordinate is in its most probable state. This factorization approximation reduces Eq. (6) to

$$Z_{\rm int,1}\cong\exp[-E(\mathbf{r}_0)/kT]\ \frac{dV_{\rm ac}}{\Pi_{i=1}^{N\rm dim}P_{i,\max}(x_i^0)}. \tag{7}$$

Equation (7) rigorously holds for functions where all probabilities in each dimension are independent. Good examples are the functions f_1 and f_2 (see Tables I and III below). In general, however, this is not the case.

Two other ways of estimating $P(\mathbf{r}_0)$ are based on the approximation that the energy landscape is locally quasiharmonic. One then constructs a transformation matrix that transforms the initial coordinate set into a set of normal coordinates. For small oscillations (harmonic) around the equilibrium positions, the normal modes can be treated independently. First, one needs to construct the covariance matrix[13–15] for the system, and then diagonalize it. We define the covariance matrix, with respect to the most probable structure ($\mathbf{r}_0$), rather than with respect to the average structure

$$\sigma_{ij}=\langle(x_i-x_i^0)(x_j-x_j^0)\rangle. \tag{8a}$$

For a harmonic energy landscape, this definition is identical to the one based on the average structure, but use of the most probable state has the advantage that it places the reference state in an energy minimum rather than in a maximum for symmetric, bimodal distributions. The diagonal elements are the variances for each coordinate, and the off-diagonal elements are the covariances. The square root of the determinant of the covariance matrix, multiplied by $(2\pi)^{N/2}$ and $\exp[-E(\mathbf{r}_0)]$, gives the partition function if the energy landscape is harmonic [and is obtained by combining Eq. (6) and Eq. (8) in Ref. 14], that is,

$$Z_{\rm int,2}\cong[\det(\sigma)]^{1/2}(2\pi)^{N/2}\ \exp[-E(\mathbf{r}_0)/kT]. \tag{8b}$$

The final approximation simply uses the normal coordinate transformation to calculate the product of the independent possibilities. To obtain the normal coordinate $\{\xi\}$, we proceed as follows: The matrix of the energy second derivatives $\mathbf{F}$ is constructed from the covariance matrix

$$F_{ij}=kT[\sigma^{-1}]_{ij}. \tag{8c}$$

After diagonalizing $\mathbf{F}$ or σ, we get a set of N eigenvectors. The resulting eigenvector matrix is the desired transformation matrix. After transformation of the initial coordinate set $\{\mathbf{x}_0^i\}$ onto a normal coordinate set $\boldsymbol{\xi}$, then the independent probabilities $P_{i,\max}(\xi_i)$ are calculated in normal mode space. The resulting partition function in terms of coordinates in normal mode space is approximated by

$$Z_{\rm int,3}\cong\exp[-E(\mathbf{r}_0)/kT]\ \frac{dV_{\rm ac}}{\Pi_{i=1}^{N\rm dim}P_{i,\max}(\xi_i)}. \tag{8d}$$

Monte Carlo sampling

The Monte Carlo sampling procedure consists in the first stage of a random walk[12] with a step Δb equal to the discretization. In a random walk, the new value of a coordinate is generated from the old value by addition or subtraction of "Δb." The values of Z obtained by the random walk sampling are reported in the top rows of Table II. Random walk runs can be considered as a prescreening of the accessibility of the conformational space by each degree of freedom and provide an estimate of boundary values for the coordinates (the boundary values depend on the steepness of the energy function under consideration). In the next step, the sampling is uniform (new values of the coordinates are generated independently of old values) in between the boundaries for each coordinate. In both cases, the standard Metropolis criterion[11] was used to determine the transition probability.

RESULTS

Exact treatment

MC simulations were performed on a set of test functions summarized in Table I. In Table II, a comparison of the

TABLE I. Description of the test energy functions used in the simulations.

Abbreviation	Equation	Description
f_1	$f_1=\sum_{i=1}^{N} 0.5x_i^2$	One ($N=1$), two ($N=2$) or three ($N=3$) dimensional harmonic oscillator
f_2	$f_2=\sum_{i=1}^{N} 0.5(x_i^4-x_i^2)$	One ($N=1$), two ($N=2$) or three ($N=3$) dimensional *camel back* function
f_3	$f_3=\begin{cases} -e^{-\Sigma_{i=1}^{N}x_i^2} \quad -e^{-\Sigma_{i=1}^{N}(x_i-4)^2} \\ +\infty \quad \text{for } x<-3 \text{ or } x>6.5 \end{cases}$	Sum of two Gaussians in one ($N=1$) or two ($N=2$) dimensions
f_4	$f_4=\begin{cases} -2e^{-\Sigma_{i=1}^{N}0.25x_i^2} \\ +\infty \quad \text{for } x<-3 \text{ or } x>3 \end{cases}$	One wide Gaussian in one ($N=1$) dimension or in two ($N=2$) dimensions
f_5	$f_5=\begin{cases} -4e^{-\Sigma_{i=1}^{N}1.5x_i^2} \\ +\infty \quad \text{for } x<-1.5 \text{ or } x>1.5 \end{cases}$	One narrow Gaussian in one ($N=1$) dimension or in two ($N=2$) dimensions
f_6	$f_6=x^2+y^2+xy$	Two dimensional function with cross terms
f_7	$f_5=\begin{cases} -4e^{-\Sigma_{i=1}^{N}1.5x_i^2} \\ +\infty \quad \text{for } x<-3 \text{ or } x>3 \end{cases}$	One narrow Gaussian in one ($N=1$) dimension or in two ($N=2$) dimensions

MC simulation results is made with the direct numerical integration of the partition functions. This will be referred to as the ''exact'' values (except in the case of the harmonic oscillator function where the analytical solution is well known). kT is set to 1. For all of the test functions, the partition functions and average energies at most differ by 1% from the exact values. The inaccuracies come from the discretization of the conformational space (the choice of finite Δb). For the harmonic oscillator (whose force constant is $a=0.5$), the coordinate probability distribution is Gaussian with a maximum at $x=0$. The values of the partition functions and the average energies for one-, two-, and three-dimensional harmonic oscillators obtained from the MC simulations agree within a small error with the exact analytical values. The next test function, f_2, is the so-called ''*camel back*'' and has two minima in one dimension, four minima in two dimensions, etc. For the one-dimensional case the coordinate (x) probability distribution is shown in Fig. 1. In the one-dimensional case, there are two energy minima and two most probable states (corresponding to the two lowest energy states). The third case (f_3) is a sum of two Gaussians (with hard walls) in one and two dimensions. For the one-dimensional case, the coordinate probability distribution is presented in Fig. 2, and the two most probable conformational states lie at the centers of Gaussians and correspond to two energy minima. The last case is particularly interesting, because it shows that our approach can be used to calculate the partition function even for a rather complicated energy landscape.

For the remaining test functions f_4, f_5, f_6, and f_7 (meant to be test cases for the different approximations), the simulation based values of the partition functions agree with the exact values within 1.5%.

Factorization approximations

Table III presents a comparison of the exact values of the partition functions with those obtained using the three approximate methods. As expected, for factorizable functions (f_1 and f_2), the factorization in the initial coordinate set is exact for any number of degrees of freedom. The factorization in the normal mode space is exact for the harmonic functions f_1 and f_6 for any number of degrees of freedom, as is the approach based on the determinant of the covariance matrix. Surprisingly, for the function f_2, the results based on the determinant of the covariance matrix are very close to the exact values (with less than 7% error). Generally, if we have an energy landscape with one or multiple minima with not too many flat regions (functions f_1, f_2, f_4, f_5, f_6), the factorization approximation in the initial coordinate set works reasonably well (the errors are 20% or less). Unfortunately, if the function is anharmonic with many flat regions (f_3,f_7), this approximation introduces large errors (up to 86%), and the best approximation seems to be factorization in normal mode space (with errors of 50% or less). However, for functions such as f_2, the factorization in normal mode space can have large errors (up to 60%). The results based on the determinant of the covariance matrix are exact only in the case of the harmonic energy landscape. In cases where the energy landscape is flat with one or more well defined minima, the errors can be quite substantial, e.g., 800% in the case of f_7. In most cases, the factorization approximations overestimate the internal partition functions, but both the factorization in the initial coordinate space and in normal mode space are reasonable approximations to the partition function.

TABLE II. For the test energy functions comparison of the analytical (or numerical) integration values with the MC simulation results for the partition functions.[a,b]

Function	Parameters N	Parameters Δb	Parameters # of cycles	MC simulation results Z Eq. (6)	Analytical or numerical values Z
f_1	1	0.1	1.10^8	2.5063	2.5066
			2.10^8	2.5052	
f_1	2	0.1	1.10^8	6.2626	6.283
			1.10^9	6.2853	
f_1	3	0.1	2.10^8	15.478	15.749
			210^9	15.733	
f_2	1	0.1	2.10^8	2.8460	2.8467
			2.10^8	2.8453	
f_2	2	0.1	2.10^8	8.0709	8.1038
			2.10^8	8.0855	
f_2	3	0.1	8.10^8	22.787	23.059
			2.10^9	22.950	
f_3	1	0.1	2.10^8	14.776	14.729
			2.10^8	14.817	
f_3	2	0.05	2.10^8	99.984	98.529
			2.10^8	98.997	
f_4	1	0.05	2.10^8	22.397	22.354
			2.10^8	22.468	
f_4	2	0.05	2.10^8	80.019	80.548
			2.10^8	81.879	
f_5	1	0.05	2.10^8	45.634	45.591
			2.10^8	45.625	
f_5	2	0.05	2.10^8	45.932	45.842
			2.10^8	46.142	
f_6	2	0.1	2.10^8	3.593	3.627
			2.10^8	3.642	
f_7	2	0.05	2.10^8	48.687	48.647
			2.10^8	48.799	
f_7	1	0.05	2.10^8	73.604	73.002
			2.10^8	73.194	

[a]Top lines in the fourth and fifth columns show random walk results.
[b]Bottom lines in the fourth and fifth columns show uniform sampling results.

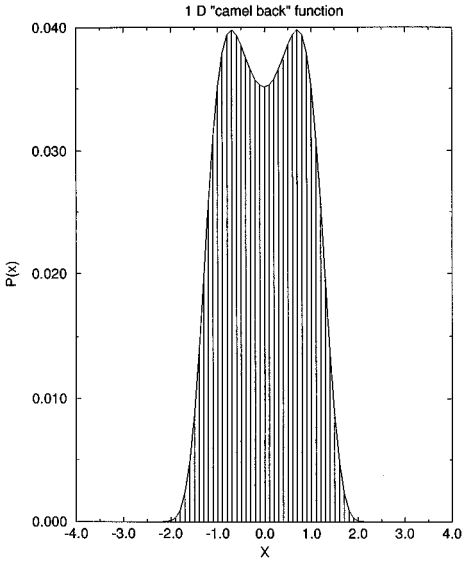

FIG. 1. Coordinate probability distribution for the one-dimensional *camel-back* function f_2. See the text for additional details.

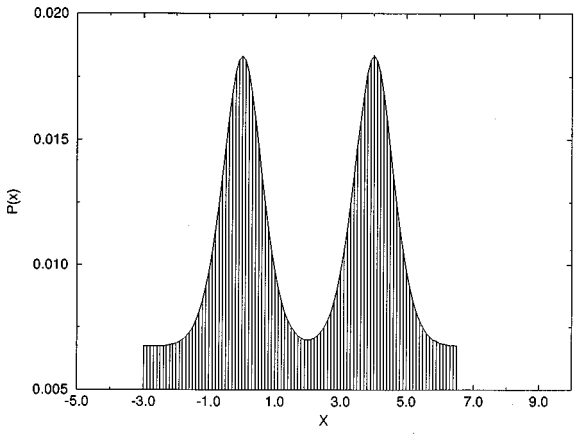

FIG. 2. Coordinate probability distribution for the one-dimensional version of f_3. See the text for additional details.

The effect of the factorization approximations on the equilibrium constant between an imaginary particle in two energy minima

The results for the equilibrium constant [Eq. (2)] between an imaginary particle experiencing the potential described by functions f_4 and f_5 in two different regions of the phase space are presented in Table IV. These functions have similar functional form, but differ by the width and the depth of their energy minima. For the one-dimensional case of functions f_4 and f_5, the equilibrium constant calculated from the simulation is within 0.5% of the exact value. In two dimensions, the factorization approximation in the initial coordinate space [Eq. (7)] is within 3% of the exact value, and differs from the exact result in three dimensions by only 14%. The factorization approximation in normal mode space [Eq. (8d)] gives results differing by roughly 25% from the exact values for the two-dimensional case, and by about 23% for the three-dimensional case. The results for the equilibrium constant based on the calculation of the determinant of the covariance matrix [see Eq. (8b)] have very large errors (up to 500%). Based on the above description, both coordinate factorization approximations work satisfactory, but it is difficult to tell which one is better.

TABLE III. Comparison of the numerical integration values for the partition functions with various factorization approximations.

Function	Analytical or numerical values Z	MC simulation results based on factorization Z_1 Eq. (7)	Z_2 Eq. (8b)	Z_3 Eq. (8d)	Parameters $N(\mathbf{a})$	Δb	# of cycles
f_1	6.283	6.2987	6.2903	6.2885	2(0.5)	0.1	2.10^8
f_1	15.479	15.737	15.731	15.764	3	0.1	2.10^8
f_2	8.1038	8.075	8.258	5.519	2(0.5)	0.1	2.10^8
f_2	23.059	23.010	21.468	14.50	3(0.5)	0.1	2.10^8
f_3	98.529	182.611	169.356	108.911	2	0.1	2.10^8
f_4	80.5475	97.603	96.951	90.981	2	0.05	2.10^8
f_5	45.8420	55.65	95.67	53.89	2	0.05	2.10^8
f_6	3.627	4.194	3.6302	3.631	2	0.1	2.10^8
f_7	73.0024	125.369	552.008	108.116	2	0.1	2.10^8

TABLE IV. Equilibrium constant for imaginary particle in two different energy minima corresponding to functions f_4 and f_5.

		MC simulation results			
Number of degrees of freedom (dimensions)	Analytical or numerical values	Exact MC treatment Eq. (6)	Factorization in the initial coordinate space Eq. (7)	Factorization in normal mode space Eq. (8d)	Determinant of the covariance matrix Eq. (8b)
2	1.757	1.742	1.754	1.688	1.01
3	5.548	5.216	4.522	3.568	1.645

SUMMARY

A straightforward approach to the calculation of partition functions has been presented. When the simultaneous probability of finding all degrees of freedom in the most probable state $P(\mathbf{r}_0)$ can be obtained, partition functions within 1% of the exact values can be calculated. For those cases where this is not possible, the most reasonable approximations are the factorization in the original coordinate space to estimate $P(\mathbf{r}_0)$, as well as the factorization approximation in normal mode space. The method is general and can be used for any bound system.[16]

ACKNOWLEDGMENTS

We wish to thank Dr. K. Olszewski and Dr. A. Godzik for their helpful discussions. This work was supported by NIH Grants No. GM38794 and FIRCA TW-00418A.

[1] P. B. Harbury, T. Zhang, P. S. Kim, and T. Alber, Science **262**, 1401 (1993).
[2] B. Lovejoy, C. Seunghyon, D. Cascio, D. K. McRorie, W. F. DeGrado, and D. Eisenberg, Science **259**, 1288 (1993).
[3] E. K. O'Shea, R. Rutkowski, W. F. Stafford III, and P. S. Kim, Science **245**, 646 (1989).
[4] J. E. Mayer and M. G. Mayer, *Statistical Mechanics* (Wiley, New York, 1963).
[5] N. Davidson, *Statistical Mechanics* (McGraw-Hill, New York, 1962).
[6] R. Fowler and E. A. Guggenheim, *Statistical Thermodynamics* (Cambridge University, Cambridge, 1960).
[7] D. R. Herschbach, J. Chem. Phys. **31**, 1652 (1959).
[8] T. L. Hill, *Statistical Mechanics* (McGraw-Hill, New York, 1956).
[9] T. L. Hill, *An Introduction to Statistical Thermodynamics* (Dover, New York, 1960).
[10] D. A. McQuarrie, *Statistical Mechanics* (Harper & Row, New York, 1976).
[11] N. A. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **51**, 1087 (1953).
[12] K. Binder and D. W. Heerman, Thesis, Institut für Physik, Johannes Gutenberg Universität Mainz, Federal Republic of Germany.
[13] C. L. Brooks, M. Karplus, and B. M. Pettit, Adv. Chem. Phys. **71**, 259 (1988).
[14] M. Karplus and J. N. Kushick, Macromolecules **14**, 325 (1981).
[15] R. M. Levy, A. R. Srinivasan, and W. K. Olson, Biopolymers **23**, 1099 (1984).
[16] M. Vieth, A. Kolinski, and C. L. Brooks III, J. Mol. Biol. (submitted, 1994).