Design of force fields from data at finite temperature

J. M. Deutsch and Tanya Kurosky
University of California, Santa Cruz, California 95064
(Received 18 April 1997; revised manuscript received 20 June 1997)

We investigate the problem of how to obtain the force field between atoms of an experimentally determined structure. We show how this problem can be efficiently solved, even at finite temperature, where the position of the atoms differs substantially from the ground state. We apply our method to systems modeling proteins and demonstrate that the correct potentials can be recovered even in the presence of thermal noise. [S1063-651X(97)04510-8]

PACS number(s): 87.15.By, 34.20.Gj

I. INTRODUCTION

In many cases it is possible to determine, quite precisely, the structure of a physical system. X-ray crystallography has made it possible to determine structures of a myriad of different compounds. Among the most complicated of these are protein crystals, where thousands of atoms appear in the unit cell. The structures of many hundreds of proteins have been determined in this way. The forces between these atoms are of great importance in predicting the interaction of proteins with other molecules and also in enabling one to do protein folding numerically. Therefore there has been a great deal of effort to determine the forces between submolecules in these systems.

One approach has been to determine the forces from *ab initio* quantum calculations of small molecules and additional data obtained from experiments on small molecules giving, for example, resonant frequencies of vibration of certain bonds. This has led to a number of force fields. For a review of these, see Ref. [1]. These have been used extensively in computational studies of biological molecules. Such potentials involve many hundreds of parameters, all of which are quite difficult to determine. These force fields are still evolving.

Another approach which is the subject of this paper has been to try to extract the values of parameters in the force field from the experimentally determined structures. This approach has some advantages to it over a direct *ab initio* approach. First, the *ab initio* approach has assumed, for the most part, two-body potentials and has ignored higher-body terms. At a microscopic level these other terms should be important. One would like to develop effective potentials that mimic the higher-body terms as well as possible. By extracting potentials from experimental structures and fitting them to an effective two-body form, an optimum two-body force field which includes higher-body effects can then in principle be calculated.

Second, the *ab initio* approach is intended to describe the interactions of all the atoms of a protein. One would like to believe, however, that such detail is not necessary in order to predict the overall structure [2]. Coarse grained force fields that consider interactions only between amino acids can be computed from the experimental structures [3–9]. This may be too crude an approximation for many applications but has the advantage that it greatly reduces the complexity of a protein folding simulation.

One of the most practical approaches along these lines has been an attempt to derive the energy of interaction of entire amino acids from their pairing frequency [4]. To do so, one treats the protein as a dilute gas of amino acids, which gives a simple analytical relation between the pairing frequency and potentials. Despite the approximate nature of such an approach, this has led to some success in predicting protein structure [10–12]. There have been some recent criticisms of the approximations used [6] along with improvements to the method [7].

The purpose of this work is as follows. We devise and test a method for determining parameters of a force field from experimental data on molecular structures. This method finds the set of parameters that will be most likely to fold the molecules into their observed structures. Our method is general enough that it can determine the parameters of a force field of arbitrary complexity, such as the *ab initio* off-lattice approaches mentioned above. This method works correctly even at finite temperature. This is important from a practical standpoint since the positions of the atoms are only defined to within a few angstroms. The problem at finite temperature is very different than at zero temperature and we will see that it is a much harder problem. Our solution is very efficient, and appears on test cases to work remarkably well.

The method used has two features that make it very powerful. First it is an iterative scheme at finite temperature. The basic idea is to start with the wrong values of parameters and perform a simulation which gives parameters closer to the true ones. Then again perform a simulation, but this time with the new estimated parameters. This is done repeatedly until satisfactory convergence is obtained. Second, this method introduces the idea of clamping. The simulations are initially performed with a clamping potential applied. This is an artificial constraint that has been added to the system to prevent it from straying too far from the known experimental structure. Without this constraint the simulation would be hopelessly slow for a real system such as a protein, because folding a protein is an extremely slow process. However, this clamping constraint does not allow the protein to explore other metastable states, meaning that the dynamics will be exponentially faster than the folding problem. As the parameters converge the clamping potential is diminished, and therefore has no effect on the final results.

A caveat that we mention is that enough experimental data must be available in order to determine the correct values of parameters. Even at finite temperature, we will see that good results can be obtained for quite small data sets. It then seems feasible that our method could be used to determine the force field of real proteins.

II. THE PROBLEM

A. Terminology

Consider a system of N atoms with coordinates $\Gamma \equiv \{\mathbf{r}_i\}$, $i=1,\ldots,N$. The atoms are of different types s, and the chemical sequence can be denoted $S=\{s_i\}$, $i=1,\ldots,N$. The Hamiltonian for the system depends on m parameters $P=\{p_i\}$, $i=1,\ldots,m$, for example, the charge and van der Waals radius. We denote the Hamiltonian as $H(\Gamma,S,P)$.

The problem is then as follows. Given experimental data on N_{mol} molecules, at finite temperature, with sequences S_i and configurations Γ_i^*, what value of parameters P will maximize the probability that these molecules have these experimentally determined structures?

Very often the parameters can be redefined in such a way that the Hamiltonian depends on them linearly,

$$H(\Gamma,S,P)=\sum_i^m p_i h_i(\Gamma,S). \tag{1}$$

For example, the van der Waals repulsion between two atoms separated by a distance r can be written as $K(a/r)^{12}$, where K and a are parameters. Both of these can be absorbed into a single parameter $p=Ka^{12}$.

B. Zero temperature

If the molecules are in their ground states, then the force on any atom must be zero. Thus minimizing the sum of the squares of the forces on all atoms with respect to the parameters P should give a solution to this problem. Indeed, numerical tests using the model presented in Sec. IV confirm that this method works very well and precisely recovers the values of all parameters, up to an overall multiplicative constant. However, at any finite temperature this method fails quite dramatically. In this case the sum of the squares of all forces can never be chosen to be truly zero. As a result the minimum is obtained by setting many parameters, such as the charge and van der Waals radius, equal to zero. At finite temperature, it is crucial to consider entropic effects and a more fundamental approach to this problem is required. We will see in the next section that at finite temperatures an optimum estimate of parameters can be obtained that includes the overall multiplicative constant.

For lattice models, the above approach will also not work even at zero temperature, since the concept of a force is more difficult to define. For a dense system, it is impossible to make small displacements, as atoms in the middle of the molecule are already surrounded by occupied sites. Thus other methods must be employed.

C. The method

The formalism used previously to analyze the problem of sequence design also applies here [13]. We want to minimize

$$\Delta F \equiv \sum_{i=1}^{N_{\mathrm{mol}}} H(\Gamma^*_i,S_i,P)-F(S_i,P) \tag{2}$$

with respect to the parameters P. ΔF is the difference between the energies of the molecules in their experimentally determined conformations, and their free energies

$$F(S_i,P)=-T\ \ln\sum_{\Gamma}\exp[-\beta H(\Gamma,S_i,P)]. \tag{3}$$

The parameters thus found are optimal in the sense that the molecules will be more likely to be in their experimentally determined structures Γ_i when they interact with these parameters than with any other choice of parameters. The present work attempts to find the solution to a well defined problem. Other recent work [9] chooses a more arbitrary criterion for optimizing the potential, and will not work at finite temperature.

In practice however, the calculation of the free energy is a formidable task, thus we must devise an efficient method to minimize ΔF.

We start by observing that if we have an approximate solution P_0, the free energy can then be expanded around that point. For notational simplicity, we will omit the summation over different molecules, as a single molecule can be redefined to be composed out of N_{mol} molecules. Corresponding to the parameters P_0, we introduce the Hamiltonian $H_0(\Gamma)\equiv H(\Gamma,S,P_0)$.

$$\Delta F \approx H(\Gamma^*)-F_0-\langle H-H_0\rangle_0 + \frac{1}{2}\beta[\langle (H-H_0)^2\rangle_0-\langle (H-H_0)\rangle_0{}^2]. \tag{4}$$

The averages $\langle\cdots\rangle_0$ are performed with respect to H_0. Since F_0 is independent of P, the minimum of this expression is much easier to determine than that of the exact one because it involves calculating averages, which is much easier than calculating free energies. The averaging can be done numerically, say by molecular dynamics or Monte Carlo simulation. A further simplification can be made for the class of Hamiltonians that are writable in the form of Eq. (1). In this case ΔF is *bilinear* in the parameters P. That is, it can be written as

$$\Delta F=\sum_i^m N_i p_i+\frac{1}{2}\sum_{i,j=1}^m p_i M_{ij} p_j+\text{const}, \tag{5}$$

where N_i and M_{ij} are constants that are determined by calculating the average above. Because of this, the minimum values of the parameters can be calculated by solving the matrix equation $Mp=-N$.

If P_0 is not too far from the true minimum, this procedure gives a better approximation to the minimum of ΔF than P_0. We can redefine P_0 to be about this new point and then repeat this procedure iteratively, until the values of parameters have converged. If P_0 is too far, the procedure will not converge, however, we have seen that the radius of convergence is greatly increased by taking fractional steps in the direction of P. If we regard P as a vector of parameters, then we can take our new set of parameters to be $\epsilon P+(1-\epsilon)P_0$.

Very interesting recent work [8] using an iterative procedure should give similar results at zero temperature. We do not expect other recent work [9] to give similar results even at zero temperature.

D. Clamping

Calculating the above averages is still quite difficult because it involves folding entire molecules with parameters P_0, to obtain their statistical properties in equilibrium. Even if we start the molecule off in the experimentally determined conformation Γ^*, it will not stay close to there if parameters P_0 are quite different than their true values. Folding real proteins is still impossible with current computers, so at first sight, the above method would appear impractical. However, we can circumvent this problem by adding a *clamping* term to H_0.

Folding proteins is difficult because of the many local minima in the energy landscape, however, if we add a clamping term to the Hamiltonian,

$$H_C = C\sum_i^N |\mathbf{r}_i - \mathbf{r}_i^*|^2, \tag{6}$$

this localizes the molecule to configurations near the experimentally determined values Γ^*. Therefore equilibrating molecules is many orders of magnitude faster than without this term, even if the value of C is rather small, allowing the atoms to explore their local environments.

So in Eq. (4), we add H_C to H_0:

$$H_0(\Gamma) = H(\Gamma_0, S, P) + H_C. \tag{7}$$

As long as C is small, the second order expansion should still be a useful approximation.

Once approximate values of parameters have been determined with the clamping potential on, it can be gradually turned off. With the correct parameters for P_0, a clamping potential is not necessary because the initial configuration we start the molecule in, Γ^*, is already correctly folded.

This trick works because, unlike the problem of protein folding, we know the tertiary structure of the molecule and can use that fact to speed up the averaging.

III. APPLICATION TO LATTICE SYSTEMS

We apply our method to lattice systems, such as the HP model [14]. Consider a two dimensional square lattice with a self-avoiding chain interacting with its nearest neighbors. We assume that there are two species of monomers σ that define the sequence of the chain, of types $\sigma = 1$ and $\sigma = 2$.

$$H(\{\sigma_i\},\{r_i\}) = \frac{1}{2}\sum_{i,j}^N V_{\sigma_i\sigma_j}\Delta(\mathbf{r}_i - \mathbf{r}_j). \tag{8}$$

$\Delta(\mathbf{r})$ is 1 if $\mathbf{r}$ is nearest neighbor displacement, and zero otherwise. In the HP model [14], the interaction between type i and j, V_{ij}, is especially simple: $V_{11} = V_{12} = 0$ and $V_{22} = -1$. For a given sequence, the ground state may be degenerate. For $N = 14$, there are 386 sequences with unique ground states, so-called ''good sequences.'' We randomly chose 37 of these ground state sequences as input to our algorithm which gave predictions for the V_{ij}'s [15].

We chose H_0 to be zero if there was one or more nearest neighbor contact, and otherwise, it was infinite. This confines all our averaging to conformations that have a chance of being a ground state. A conformation with no contacts cannot be in a unique ground state. We did not use Monte Carlo simulation, but instead calculated the averages using exact enumeration. This is quite efficient as the averages in Eq. (4) can be written in terms of second and fourth order correlation functions, $C_{ij} \equiv \langle \Delta(\mathbf{r}_i - \mathbf{r}_j)\rangle$ and $D_{ijkl} \equiv \langle \Delta(\mathbf{r}_i - \mathbf{r}_j)\rangle\langle \Delta(\mathbf{r}_k - \mathbf{r}_l)\rangle$. These correlation functions are only computed once and so the design code runs very quickly, over an order of a few seconds on an Intel 586 machine.

Minimizing Eq. (4) gives the values [15] $V_{11} = 0.057$, $V_{12} = 0.14$, and $V_{22} = -1$. This might seem to be quite far off from the original values, however, refolding the 37 chains using these new values gives precisely the same ground states for all the chains. In other words this potential gives the same ground state as the original.

For a commonly used variant [16,17] of the Dill and Lau model, there are 1619 good sequences. In this case, the values found are [15] $V_{11} = -0.89, V_{12} = 0.28$, and $V_{22} = -1$. Again, this correctly refolds all 37 conformations considered to the correct ground states.

In both cases, the method reproduces the correct ground states immediately, so that an iterative method need not be considered. We now turn to a continuous system at finite temperature.

IV. APPLICATION TO AN OFF-LATTICE SYSTEM

We now consider an off-lattice system containing much of the essential physics of a real protein. We consider a sys-

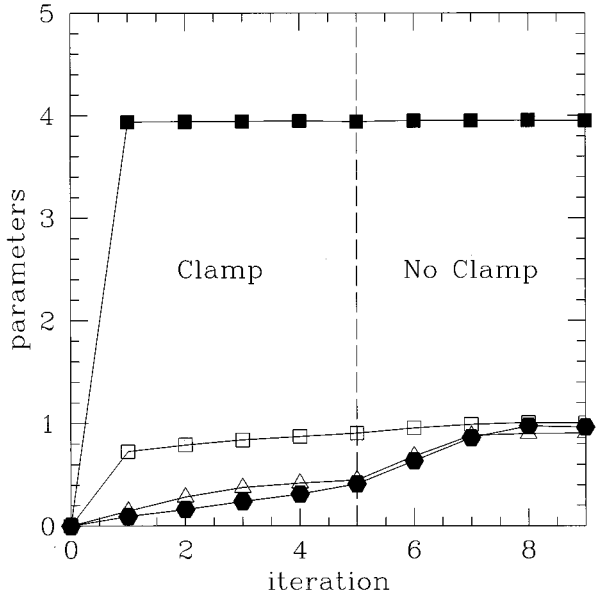

FIG. 1. The computed values of the parameters as a function of the number of iterations for the off-lattice model considered in the text. The spring constant k is denoted by the open triangles, the equilibrium spring length r_0 by solid squares, the charge Q by solid hexagons, and the van der Waals radius a by open squares

tem of atoms connected by springs with an equilibrium length r_0, and spring coefficient k. We also say that there are two types of atoms with charge q_i of either Q or $-Q$. Finally we include a van der Waals repulsion $(a/r)^{12}$. The Hamiltonian is then

$$H=\sum_i^N \frac{k}{2}(r_i-r_0)^2+\sum_{i<j}^N \frac{q_i q_j}{|\mathbf{r}_i-\mathbf{r}_j|}+\left(\frac{a}{|\mathbf{r}_i-\mathbf{r}_j|}\right)^{12}. \quad (9)$$

which depends on the parameters k, r_0, Q, and a. This Hamiltonian can be rewritten in the form of Eq. (1).

To test our method, we first made a database of 12 8-mer structures, with 12 different sequences of the q_i. We chose some fixed values for the parameters, $k=1$, $r_0=4$, $Q=1$, and $a=1$. We cooled the atoms using simulated annealing down to a temperature where they had collapsed to well defined structures, $\beta=20$. Then we fed these structures into our program, which uses a Monte Carlo calculation to estimate the averages in Eq. (4). We applied a moderate clamping potential with $C=2.5$ for five iterations and then turned it off and continued to iterate four more times. The program is supposed to determine the parameters k, r_0, Q, and a from only the database of these 12 structures. The results are displayed in Fig. 1. The results took about five minutes on an Intel 586 microprocessor. The computed parameters are within 12% of the real values.

V. CONCLUSIONS

We have presented a relatively simple method for determining forces between atoms from their structure at finite temperature. We have applied this to several model systems, on lattice and off lattice, and have found that it gives accurate results very efficiently. Our approach expands ΔF introduced earlier [13] to second order about some approximate parameters. ΔF is again minimized, and the procedure is repeated iteratively until satisfactory convergence is obtained. Because of the efficiency of this method, it appears computationally feasible to us to apply our method to real protein data bases. This is currently under investigation.

ACKNOWLEDGMENTS

We wish to thank Douglas Williams for useful discussions. This work is supported by NSF Grant Number DMR-9419362 and acknowledgment is made to the Donors of the Petroleum Research Fund, administered by the American Chemical Society for partial support of this research.

[1] V. Daggett and M. Levitt, Annu. Rev. Biophys. Biomol. Struct. **22**, 353 (1993).
[2] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[3] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **227**, 876 (1992).
[4] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[5] M. Hendlich *et al.*, J. Mol. Biol. **216**, 167 (1990).
[6] P. D. Thomas and K. A. Dill, J. Mol. Biol. **257**, 457 (1996).
[7] P. D. Thomas and K. A. Dill, Proc. Natl. Acad. Sci. USA **93**, 11 628 (1996).
[8] M. H. Hao and H. A. Scheraga, Proc. Natl. Acad. Sci. USA **93**, 4984 (1996).
[9] L. A. Mirny and E. I. Shakhnovich, J. Mol. Biol. **264**, 1164 (1996). .
[10] C. Wilson and S. Doniach, Proteins **6**, 193 (1989).
[11] S. Sun, Protein Sci. **2**, 762 (1993).
[12] A. Monge, E. J. Lathrop, J. R. Gunn, P. S. Shenkin, and R. A. Friesner, J. Mol. Biol. **247**, 995 (1995).
[13] T. Kurosky and J. M. Deutsch, J. Phys. A **27**, L387 (1995).
[14] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[15] Note that because we are feeding in states at zero temperature, the potential is only determined to within an overall multiplicative factor.
[16] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993).
[17] E. I. Shakhnovich and A. M. Gutin, Protein Eng. **6**, 793 (1993).