

Correspondence of potentials of mean force in proteins and in liquids

Yibing Shan and Huan-Xiang Zhou^{a)}

Department of Physics, Drexel University, Philadelphia, Pennsylvania 19104

(Received 25 October 1999; accepted 21 June 2000)

The concept of potential of mean force (PMF) is now widely used in predicting protein structures. Proteins notably differ from liquids by their inhomogeneity and chain connectivity. Does meaningful correspondence exist between PMFs in proteins and PMFs in liquids? This question was addressed in this article. We constructed “proteins” each with 90 residues selected from a system of 500 hard spheres. The residues were of two types, N and P. They interact among themselves (with energies E_{NN} , E_{PP} , E_{NP}) and the 410 “solvent” spheres (with energies E_{NS} and E_{PS}). Out of the 500 hard spheres, we first identified all chains consisting of 90 residues that have appropriate distances between nearest neighbors. The conformation of a protein was selected as the one having the lowest total energy among the 3.7 million chains. A corresponding liquid system was constructed without imposing distance constraints among solute spheres. The PMFs obtained from the proteins and the liquid system show remarkable similarities. For eleven sets of the energy parameters, the first minima of the PMFs in the proteins agree with their counterparts in the liquid state to within a constant. © 2000 American Institute of Physics. [S0021-9606(00)51235-5]

I. INTRODUCTION

Knowledge of how frequently different types of amino acids are found near each other in known protein structures has been widely used in predicting protein structures.^{1–11} A number of procedures have been proposed to extract potentials of mean force (PMFs) from known protein structures.^{12–19} The guiding principle is that, in liquids, the pair correlation function $g(r)$ (equivalent to the pair frequency in proteins) is related to the PMF $w(r)$ through $g(r) = \exp[-\beta w(r)]$, where $1/\beta = k_B T$ is the product of the Boltzmann constant and the temperature. However, proteins notably differ from liquids by their inhomogeneity and chain connectivity. Does meaningful correspondence exist between PMFs in proteins and PMFs in liquids?

In this article we address the above question by working with model proteins constructed from a system of 500 hard spheres. The hard-sphere system was chosen because it is simple and well understood as a liquid. If 90 of the 500 hard spheres are chosen as protein residues, then there are $500!/410! \approx 10^{239}$ possible conformations. The distances between two nearest C_{α} s in proteins are always very close to 3.8 Å. To mimic this chain connectivity, we imposed the restriction that the distances between nearest neighbors are less than 3.83 Å (and of course greater than the diameter σ of a hard sphere, which is set to 3.33 Å). For each such chain, we allowed for four “bad” bonds—those with nearest-neighbor distances up to $1.5\sigma = 5$ Å. A total of 3.7 million chains were found. For simplicity we considered two types of residues, N (for nonpolar) and P (for polar). The residues of a protein, 90 in total, interact among themselves and with the remaining 410 solvent, or S, hard spheres. The interaction energies between two spheres are denoted E_{AB} , where

$A, B = N, P, \text{ or } S$. For a protein with a particular sequence of residues (i.e., a series of N’s and P’s), the structure was selected to be the one with the lowest total energy among the 3.7 million chains.

We are faced with two basic questions. From the structures of the proteins one can obtain the pair frequency of any two types of residues, but how does one convert this pair frequency into a PMF? In the liquid state, the PMF and the pair correlation function are connected by a Boltzmann relation. In proteins the situation is complicated by the chain connectivity. In particular the size of a protein scales with N_{res} (the number of residues) roughly as $N_{\text{res}}^{1/3}$, thus the number of pairs with large distances will decrease to zero. In essence, relating the pair frequency to a PMF is a matter of defining a reference state. In the liquid state, the reference state is one in which all the spheres are randomly distributed. This obviously is a bad choice for proteins since the randomly distributed spheres would violate chain connectivity. In our earlier work¹⁹ we introduced a reference state in which chain connectivity and inhomogeneity are specifically accounted for. This reference state will be used here.

The second basic question is, how is the PMF related to the elementary interaction energies (i.e., E_{AB})? This question is difficult to answer even in the liquid state. Consider the pure hard sphere liquid in which there is no interaction between the spheres except for excluded volume. The pair correlation function is complicated, with peaks at full multiples of σ and valleys in between. Thus, a simple relation between a protein PMF and E_{AB} is very unlikely. This is precisely the reason why we seek to find correspondence between proteins and the liquid state. Hopefully the insight on the liquid state can be imparted to proteins.

This article is organized as follows. In Sec. II we describe the generation of the protein structures and the calculation of the PMF. The results for the PMFs in the proteins

^{a)} Author to whom correspondence should be addressed; electronic mail: hxzhou@einstein.drexel.edu.

TABLE I. Parameter sets and the contact minima of the protein PMFs.

Set	ENN	EPP	ENP	ENS	EPS	N-N pair	P-P pair	N-P pair
1	-1.0	-0.8	0.2	0.5	0.0	-0.73	-0.28	0.33
2	-1.0	-1.6	0.2	0.5	0.0	-0.59	-0.60	0.33
3	-1.0	-2.4	0.2	0.5	0.0	-0.53	-0.75	0.33
4	-1.0	-3.2	0.2	0.5	0.0	-0.44	-0.78	0.33
5	-0.5	-1.6	0.2	0.5	0.0	-0.51	-0.65	0.30
6	-1.5	-1.6	0.2	0.5	0.0	-0.69	-0.47	0.38
7	-2.0	-1.6	0.2	0.5	0.0	-0.70	-0.39	0.37
8	-1.0	-1.6	0.2	0.5	-0.5	-0.67	-0.42	0.38
9	-1.0	-1.6	0.2	0.5	0.5	-0.49	-0.69	0.31
10	-1.0	-1.6	0.2	0.0	0.5	-0.51	-0.75	0.38
11	-1.0	-1.6	0.2	1.0	0.5	-0.57	-0.52	0.18

are compared to those in the liquid state in Sec. III. In Sec. IV we make some remarks on the physical basis of pair frequencies in proteins.

II. METHODS

A. Generation of protein structures

We used one particular configuration of the system of 500 hard spheres for selecting protein conformations. This is the one at the end of 1 billion collisions after starting the hard spheres on a face-centered cubic lattice. The simulation of the movement of the hard spheres was carried out by a program provided in the book of Allen and Tildesley.²⁰ The 3.7 million 90-residue chains with constrained nearest neighbor distances were generated by constructing a tree starting from each of the 500 hard spheres. The second level of the tree consists of all the neighbors of the starting sphere. A neighbor is defined as one that is within $1.5\sigma=5 \text{ \AA}$. Each subsequent level consists of the neighbors of the spheres in the preceding level. A sphere is excluded from a level if it has already appeared in any previous level. A tally of bad bonds is kept for each branch. When the tally reaches four, that branch is pruned. This procedure is expanded to the 90th level.

The total energy \mathcal{E} of each protein is the sum of all the pairwise interactions involving the protein residues:

$$\mathcal{E} = \frac{1}{2} \sum_{i \neq j=1}^{90} E_{ij} + \sum_{i=1}^{90} \sum_{l=1}^{410} E_{il}, \quad (1)$$

where i and j refer to spheres on a chain and l refers to the rest 410 solvent spheres. For any two spheres, the interaction energy is a constant E_{AB} ($A, B = N, P, S$) if the distance r is between $\sigma=3.33 \text{ \AA}$ and $1.5\sigma=5 \text{ \AA}$. It is zero if $r > 1.5\sigma$. Eleven sets of energy parameters were studied. These are listed in Table I. For a given sequence of N and P residues, the energy function in Eq. (1) was applied to each of the 3.7 million chains. The one with the lowest total energy was selected as the actual conformation.

Two hundred and twelve protein sequences were studied. These were based on a set of 243 actual proteins compiled in our previous study.¹⁹ These proteins were nonhomologous and monomeric. From each of these proteins, we selected 90 consecutive residues that had the smallest radius of gyration (the 31 proteins with less than 90 residues were

discarded). The residue types were then converted to N or P. The former case consists of Leu, Ile, Val, Phe, Met, Cys, Pro, Gly, and Ala and the latter consists of the other eleven natural amino acids. The conformation of one particular model protein is shown in Fig. 1. Among the 212 sequences the average numbers of Ns and Ps were 39 and 51, respectively.

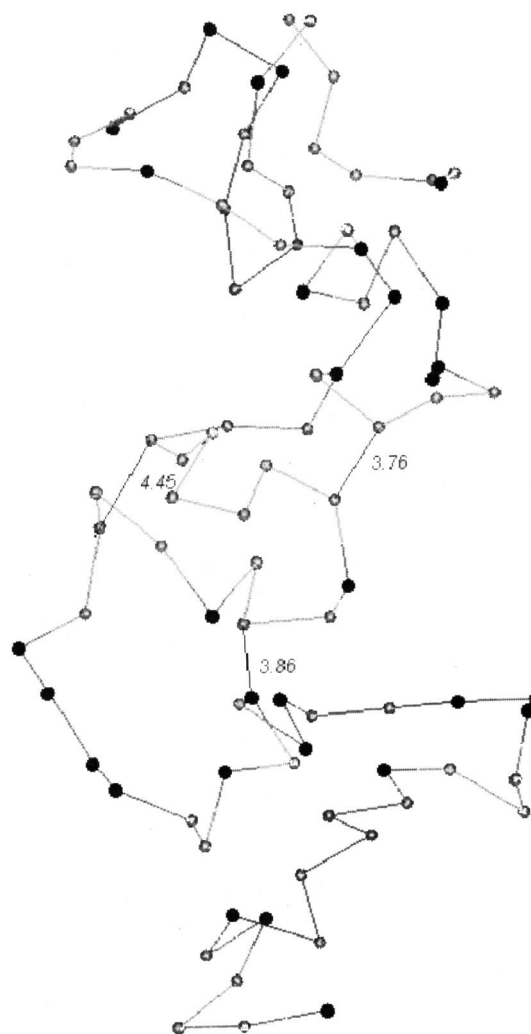


FIG. 1. Structure of a model protein. The black and gray balls represent P and N residues, respectively. Three close contacts are indicated by thin lines and distances in \AA . They are separated by 16, 14, and 11 bonds, respectively.

B. Calculation of protein PMF

The PMF is obtained by comparing a particular pair's (e.g., N-P's) frequency in the model proteins and that in a reference state. Let these be $PF(r)$ and $PF^0(r)$, respectively. The potential of mean force, $w_p(r)$, is then given by

$$PF(r) = \exp[-w_p(r)]PF^0(r). \quad (2)$$

In principle a temperature factor $k_B T$ should be introduced in Eq. (2). We just interpret $w_p(r)$ as being measured in units of $k_B T$.

The reference state should mimic the model proteins in every aspect except that the interactions between the residues are eliminated. In the model proteins (as in real proteins), the pair frequency is also influenced by the inhomogeneous distributions of residues within the proteins (due to residue-solvent interactions) and the chain constraints. The influence of the latter two factors should be kept intact in the reference state. This led us to the following procedure for generating the reference state.¹⁹ First, the inhomogeneous distributions of residues were accounted for by keeping the radial distances at their values in the model proteins but selecting their polar and azimuthal angles randomly. The chain connectivity was then modeled by imposing a Gaussian constraint on the distance between each residue pair. The means of the Gaussian constraints were 5, 17, 9, 25, and 32 Å for pairs separated by 1–6, 7–11, 12–21, 22–51, and 52–89 bonds, respectively. The standard deviations were 5, 7, 17, 26, and 20 Å, respectively. These values were very similar to those used for actual proteins in our previous work. The only notable difference is the mean for pairs separated by 12–21 bonds, which decreased from 24 to 9 Å. This decrease can be understood by the observation that, in the model proteins, residues separated by 12–21 bonds tend to form close contacts (see Fig. 1). In generating the reference state, residues were “grown” one at a time, each time chain constraints with all preceding residues were accounted for simultaneously. This procedure is very similar to the classical construction of a random chain. To avoid local effects, the distances of pairs that are separated by less than seven bonds were excluded in binning the pair frequencies.

C. PMF in the liquid state

For the liquid state, 90 of the 500 hard sphere were assigned to be solutes. Of these, 39 were assigned to be N and 51 were assigned to be P. We took a set of 5 million samples (out of $500!/410!/39!/51! \approx 10^{126}$ possibilities) and selected the 10 000 with the lowest total energies among them. The solute and solvent spheres interacted exactly the same way as in the model proteins, but we now introduced periodic boundary conditions with the minimum image convention. The PMF $w_1(r)$ was calculated by Eq. (2), but the reference state was changed to be one in which the 500 spheres are randomly dispersed in the basic simulation box.

Though the above approach for obtaining the PMFs in the liquid state mirrors that used for the model proteins, it is more natural to obtain the PMFs in the liquid state by taking into account all the 5 million samples of solute assignments,

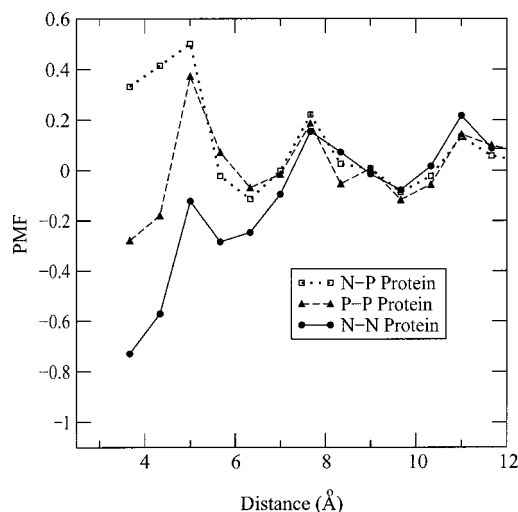


FIG. 2. PMFs of the N-P, P-P, and N-N pairs in the model proteins.

each according to its Boltzmann weight. This approach was also used to obtain the PMFs in the liquid state.

III. RESULTS

Figure 2 shows the protein PMFs of the N-N, P-P, and N-P pairs for parameter set 1 (i.e., $E_{NN} = -1.0$, $E_{PP} = -0.8$, $E_{NP} = 0.2$, $E_{NS} = 0.5$, and $E_{PS} = 0.0$). All the three PMFs exhibit a minimum at contact distance (i.e., $r = \sigma$), but the minima of the N-N and P-P pairs are negative whereas the minimum of the N-P pair is positive. These are consistent with the negative interaction energies of the N-N and P-P pairs and the positive interaction energy of the N-P pair. The maxima at $r = 1.5\sigma = 5$ Å are clearly visible. Beyond $r = 2\sigma$ the three PMFs fluctuate around zero. For comparison, we show the PMFs of the N-P and N-N pairs in the liquid state in Fig. 3. The liquid-state PMF of the N-P pair is nearly the same as the PMF of the pure solvent (not shown) and exhibits a minimum at $r = \sigma$ and a maximum at

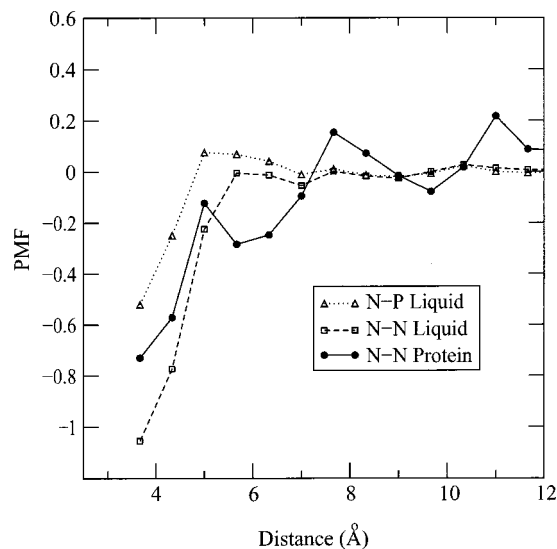


FIG. 3. PMFs of the N-P and N-N pairs in the liquid state. Also shown is the PMF of the N-N pair in the model proteins.

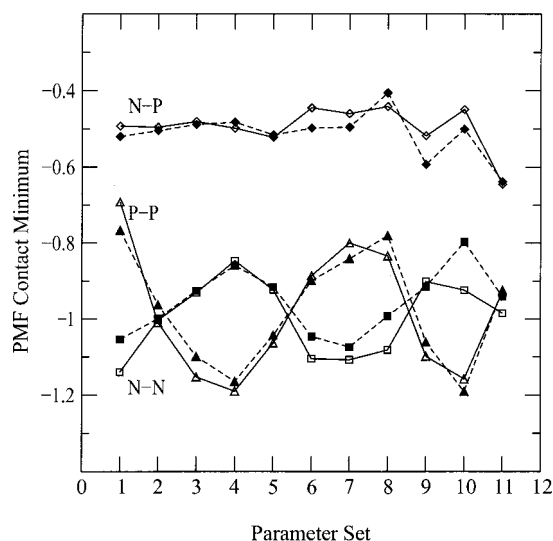


FIG. 4. Contact minima of the PMFs. Solid and dashed curves are the results from the proteins and those in the liquid state, respectively. The former results are shifts downward (by 0.405, 0.405, and 0.811).

$r = 1.5\sigma$. The liquid-state PMF of the N–N pair has a lower minimum at $r = \sigma$ than its counterpart in the proteins. Interestingly, the liquid-state PMF of the N–N pair does not exhibit a peak at $r = 1.5\sigma$. This indicates that the peak which otherwise would appear is suppressed by the strong attraction between two N residues at this distance in the liquid state. Beyond $r = 2\sigma$ the liquid-state PMFs fluctuate much closer to zero than their counterparts in the proteins. The better statistics are obviously a result of more sampling (10 000 vs 212).

We studied eleven sets of the E_{AB} parameters to gain a better understanding of their influence on the PMFs. The minima at contact distance are listed in Table I and plotted in Fig. 4. It is immediately clear that the parameter changes have very little effect on the PMF of the N–P pair. It is also interesting to note that, even when only E_{NN} (or E_{PP}) is changed, both the PMF of the N–N pair and the PMF of the P–P pair are affected.

Figure 4 plots the first minima of the PMFs in the liquid state, obtained on the 10 000 samples with lowest energies (out of 5 million). The PMFs of the N–N, P–P, and N–P pairs in the liquid state closely track their counterparts in the proteins upon shifting the latter results downward by 0.405, 0.405, and 0.811, respectively. The contact minima of the N–P pair also show little variation in the liquid state. In fact these contact minima are nearly the same as that found for the pure solvent, and hence are hardly affected by the pair interactions introduced through E_{AB} .

The PMFs in the liquid state calculated on the low-energy samples are compared to those calculated by weighting all the 5 million samples by the Boltzmann factor in Fig. 5. Reasonable consistency between the two approaches can be observed.

IV. DISCUSSION

We have demonstrated that the PMFs from proteins show remarkable similarities to the PMFs in the liquid state.

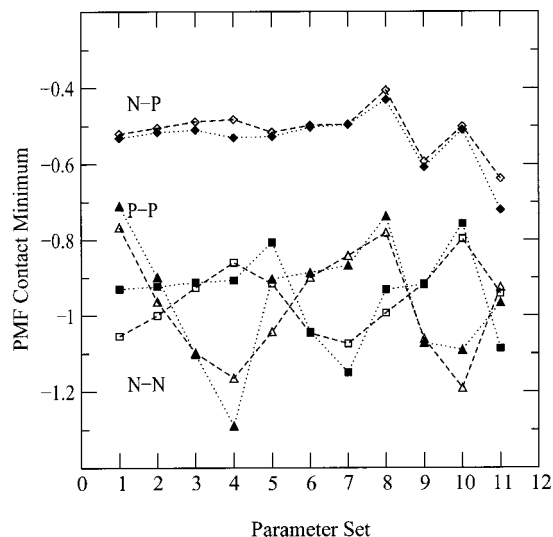


FIG. 5. Comparison of the PMFs in the liquid state calculated in two approaches: on the 10 000 low-energy samples (dashed curves) and by weighting all the 5 million samples by the Boltzmann factor (dotted curves). The latter results were obtained by using a temperature factor $k_B T = 4$.

For eleven sets of the energy parameters, the first minima of the PMFs in the proteins agree with their counterparts in the liquid state to within a constant. The correspondence comes after complications due to proteins' chain connectivity and inhomogeneity are properly accounted for in the PMF calculation. This correspondence provides insight to the physical basis of pair frequencies and PMFs in proteins.

The contact minima of the PMFs of the N–N and P–P pairs are about equal for parameter set 2. In this case $E_{NN} = -1.0$ and $E_{PP} = -1.6$. One may have expected that equality of the contact minima is obtained when E_{NN} and E_{PP} are equal. Why is a lower E_{PP} value required? The major difference between N and P residues lies in their interactions with the solvent. Relative to P, N dislikes S. Other things being equal, two N's will more likely be near each other than two P's in order to stay away from the solvent. Thus it is the interaction with the solvent that accounts for the apparent stronger affinity between two N's. In actual proteins we indeed observed much stronger affinities between nonpolar residues than between polar residues.¹⁹

That contact minima of the PMFs in the proteins and in the liquid state do not match exactly is hardly surprising. The contact minimum measures the change in free energy when a pair of residues are brought into contact from a large (or effectively infinite) separation. In that initial state each partner should be in an "average" environment. In the proteins, because of chain connectivity, a N (or P) residue in an average environment will be around other N (or P) residues. Hence in proteins there are residual favorable interactions with the environment, which would reduce the magnitude of the free energy change upon forming a contact pair. This explains why the contact minima of the N–N and P–P pairs are shallower in the proteins.

This study was inspired by the work of Thomas and Dill,¹⁴ who raised fundamental questions about the relationship between pair frequencies and PMFs in proteins. One major concern was the effect of the burial of nonpolar resi-

dues. The burial of nonpolar residues, due to unfavorable interaction with the solvent, if not properly treated can dominate PMFs of other pairs. In our definition of the reference state for calculating protein PMFs, we specifically account for the burial of nonpolar residues by preserving their radial distances. That the protein PMFs thus obtained track those in the liquid state demonstrates that our procedure is appropriate.

It has been suggested^{14,16} that the PMFs extracted from the proteins should be exactly the same as the pair interaction energies ("true energies" in the terminology of Ref. 14 and E_{AB} here). This cannot be the case. As noted earlier, the PMF of the ordinary hard sphere liquid is nonzero and indeed exhibits minima at $r=n\sigma$ and maxima at $r=(n+0.5)\sigma$, even though the pair interaction energy is zero (i.e., besides excluded volume). The PMFs are affected by coupling between different residue-residue and residue-solvent pairs. We have seen this coupling effect both in the protein PMFs and in the liquid-state PMFs. A consequence is that the PMFs will depend on the residue compositions of the proteins. This is not a major concern, since the residue compositions of proteins are relatively stable.

The hard sphere model for proteins introduced in the present study has certain advantages over the popular lattice model. The hard sphere system is well understood as a liquid. In addition, solvent can be explicitly included without difficulty.

In summary, we have demonstrated correspondence be-

tween PMFs in proteins and in the liquid state. This correspondence should strengthen the physical basis of proteins PMFs.

ACKNOWLEDGMENT

This work is supported in part by NIH Grant No. GM58187.

- ¹M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl, *J. Mol. Biol.* **216**, 167 (1990).
- ²D. T. Jones, W. R. Taylor, and J. M. Thornton, *Nature (London)* **358**, 86 (1992).
- ³S. H. Bryant and C. E. Lawrence, *Proteins* **16**, 92 (1993).
- ⁴J. P. Kocher, M. J. Rooman, and S. J. Wodak, *J. Mol. Biol.* **235**, 1598 (1994).
- ⁵S. E. DeBolt and J. Skolnick, *Protein Eng.* **9**, 637 (1996).
- ⁶L. A. Mirny and E. I. Shakhnovich, *J. Mol. Biol.* **264**, 1164 (1996).
- ⁷L. A. Mirny and E. I. Shakhnovich, *J. Mol. Biol.* **283**, 507 (1998).
- ⁸M. Rooman and D. Gilis, *Eur. J. Biochem.* **254**, 135 (1998).
- ⁹J. Skolnick and A. Kolinski, *Science* **250**, 1121 (1990).
- ¹⁰S. Sun, *Protein Sci.* **2**, 762 (1993).
- ¹¹A. Kolinski and J. Skolnick, *Proteins* **18**, 338 (1994).
- ¹²S. Miyazawa and R. L. Jernigan, *Macromolecules* **18**, 534 (1985).
- ¹³S. Miyazawa and R. L. Jernigan, *J. Mol. Biol.* **256**, 623 (1996).
- ¹⁴P. D. Thomas and K. A. Dill, *J. Mol. Biol.* **257**, 457 (1996).
- ¹⁵F. Melo and E. Feytmans, *J. Mol. Biol.* **267**, 207 (1997).
- ¹⁶C. Zhang, *Proteins* **31**, 299 (1998).
- ¹⁷A. Rojnuckarin and S. Subramaniam, *Proteins* **36**, 54 (1999).
- ¹⁸S. Miyazawa and R. L. Jernigan, *Proteins* **36**, 357 (1999).
- ¹⁹M. Vijayakumar and H.-X. Zhou *J. Phys. Chem.* (in press).
- ²⁰M. P. Allen and D. J. Tildesley, *Computer Simulations of Liquids* (Oxford University Press, Oxford, 1987).