Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 4904–4907, April 1999
Biophysics

Statistical mechanics of protein-like heteropolymers

RUXANDRA I. DIMA*§, JAYANTH R. BANAVAR*, MAREK CIEPLAK†, AND AMOS MARITAN‡

*Department of Physics and Center for Materials Physics, 104 Davey Laboratory, Pennsylvania State University, University Park, PA 16802; †Institute of Physics, Polish Academy of Sciences, 02-668 Warsaw, Poland; and ‡International School for Advanced Studies (S.I.S.S.A.), Via Beirut 2-4, 34014 Trieste, Istituto Nazionale di Fisica Nucleare and the Abdus Salam International Center for Theoretical Physics, Trieste, Italy

Edited by Peter G. Wolynes, University of Illinois, Urbana, IL, and approved February 15, 1999 (received for review November 30, 1998)

ABSTRACT **A strategy is outlined for obtaining the free energy of a typical designed heteropolymer. The design procedure considers the probability that the target conformation is occupied in comparison with all the other conformations that could house the given sequence. Numerical calculations on lattice heteropolymer models are presented to illustrate the key physical principles.**

Protein folding is an important example of the general class of problems involving conflicting constraints and, thence, a rugged energy landscape (1–6). The microscopic approaches of polymer theory and the results obtained in the study of spin glasses potentially add up to a powerful framework for the study of a variety of systems including proteins, polyampholytes, imprinted copolymers, and gels (1–13). Recently, many papers have been published that have attempted to use this framework for a sophisticated, analytic study of the phase diagram of designed heteropolymers (7, 14–19).

The functionality of a protein is mainly controlled by its structure in its native state (commonly assumed to be its ground state). An optimal tailoring of the structure of the native state of a protein by altering its amino acid sequence will enable the creation of proteins with desired functionality and will have applications in drug design. In general, a randomly chosen sequence will not have protein-like properties of a thermodynamically stable native state and rapid kinetic accessibility to it. The evolution of naturally occurring proteins with useful functionality and with native structures that are stable against mutations and small changes in solvent properties is the hallmark of a selection procedure or a design process.

The original idea for protein design (20–23) consists of running through sequences of amino acids to determine which sequence (or sequences) has the lowest energy in a target conformation. In this approach, a constraint is usually placed on the composition of amino acids in the test sequences, to avoid populating a sequence with just those amino acids that are most attractive to each other.

This idea has been formalized by noting that the probability for a particular sequence s to occur in a specific target conformation Γ is proportional to

$$P_s(\Gamma,T_{\rm des}) = \exp[-H_s(\Gamma)/T_{\rm des}], \qquad [1]$$

where $H_s(\Gamma)$ is the energy of the sequence s in conformation Γ, and $T_{\rm des}$ is a temperature at which the design is thought to occur. When $T_{\rm des} = 0$, the design is thought to be perfect and the sequence (or sequences) with the lowest energy in the target conformation are chosen, whereas for an infinite value of $T_{\rm des}$, there is no design at all and all sequences have the same P_s corresponding to what is known as the random heteropolymer case. In this problem, the annealed variables are the conformations, whereas the role of quenched random variables is played by the sequences.

We note that Eq. 1 is merely an approximation that requires modification to carry out the selection procedure for protein-like heteropolymers rigorously. P_s is affected by the probability of the sequence to be in other competing conformations and, thus, the design should maximize the relative probability of the sequence being in the target conformation. Such a correct design procedure thins out the competing low-lying energy states thereby inducing a funnel topography in the energy landscape (24, 25).

The free energy of a given sequence s at temperature T is given by

$$F(s,T) = -T\log\sum_{\Gamma}\exp[-H_s(\Gamma)/T], \qquad [2]$$

where the sum is over all the conformations of a self-avoiding walk.

The free energy of a typical random heteropolymer at a temperature T is obtained by averaging over the free energies of all sequences (with equal weight or corresponding to the infinite $T_{\rm des}$ limit) (26):

$$\langle F(T)\rangle_{\rm random} = \frac{\Sigma_s F(s,T)}{\Sigma_s 1}. \qquad [3]$$

This equation was generalized (9, 14–19) to designed sequences by postulating that the ensemble-averaged free energy was obtained as a weighted average over all sequences and given by

$$\langle F(T,T_{\rm des})\rangle = \frac{\Sigma'_\Gamma\Sigma_s P_s(\Gamma,T_{\rm des})F(s,T)}{\Sigma'_\Gamma\Sigma_s P_s(\Gamma,T_{\rm des})}, \qquad [4]$$

where the primed sum is over selected conformations that house the designed sequences. In refs. 14–19, these conformations were selected to be the compact ones.

To obtain the quenched average involving the logarithm in Eq. 2, one introduces n replicas. Further, because the sum over conformations also appears in Eq. 4, one needs yet another replica that couples to the previous n replicas after the summation over sequences which are the quenched variables.

In summary, the conventional analysis attempts to interpolate between two limits: a trivial one at infinite $T_{\rm des}$, in which all sequences have equal weight and no selection procedure is employed, and the second at $T_{\rm des} = 0$, in which only certain special sequences contribute. The goal is to have these special sequences be ones that are protein-like in the sense of having large thermodynamic stability. From Eq. 1, in the $T_{\rm des} = 0$ limit, nonzero weights are assigned only to those sequences whose ground-state energies have the lowest value among the ground-state energies of all sequences. Such sequences do not necessarily correspond to ones that are thermodynamically

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

This paper was submitted directly (Track II) to the *Proceedings* office.
§To whom reprint requests should be addressed at: Department of Physics, 104 Davey Laboratory, University Park, PA 16802. e-mail: dima@phys.psu.edu.

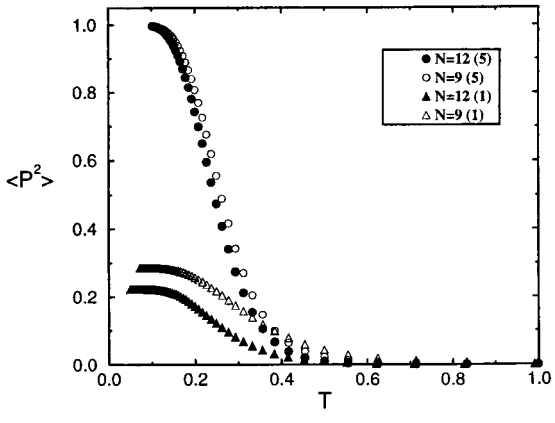

FIG. 1. Plot of the order parameter $\langle P^2 \rangle$ versus T at $T_{des} = 0$ for the HP model applied to chains of $N = 9$ and $N = 12$ beads. The number 1 or 5 in parentheses denotes the use of the corresponding equation in the text.

stable, but instead may very well be characterized by high degeneracies or low-lying excited states, thus making them physically uninteresting. As a consequence, the conventional analysis possibly may not interpolate between physically relevant limits.

We now proceed to the rigorous way of implementing this problem. Eqs. 2 and 4 are correct, whereas Eq. 1 will be replaced. It is still true that the free energy of a typical random heteropolymer at a temperature T is obtained by averaging over the free energies of all sequences (with equal weight or corresponding to the infinite T_{des} limit) (25). The key error in the analysis is Eq. 1, which should be replaced by

$$P_s(\Gamma, T_{des}) = \exp\{-[H_s(\Gamma) - F(s, T_{des})]/T_{des}\}, \qquad [5]$$

where $F(s, T_{des})$ is defined in Eq. 2. Physically, Eq. 4 arises from the observation that the probability that a sequence s is in the conformation Γ at a temperature T_{des}, depends not only on the energy of the sequence in the conformation, but also involves the partition function in the denominator as a normalization (27–30). Thus, it does not suffice to merely consider the target conformation energy but the probability that the target conformation is occupied in comparison with all the other conformations that could house the given sequence. The correct procedure is clearly more cumbersome than the previous approaches and will entail the introduction of more replicas.

It is important to note that Eq. 1 is a special case of the correct Eq. 5, when the free energies of the protein-like sequences are essentially the same independent of sequence, i.e., self-averaging. It would be interesting to assess whether the free energies of protein-like sequences do become sequence-independent in the thermodynamic limit.

To illustrate the difference between the sequences selected by the two procedures, one based on Eq. 1 and the other on Eq. 5, we studied some representative quantities of an ensemble of sequences with the aid of numerical calculations with T_{des} set equal to zero.

We begin by studying the behavior of an effective order parameter defined as

$$\langle P^2(T, T_{des}) \rangle = \frac{\Sigma'_\Gamma \Sigma_s P_s(\Gamma, T_{des}) P_s^2(T)}{\Sigma'_\Gamma \Sigma_s P_s(\Gamma, T_{des})}, \qquad [6]$$

where

$$P_s^2(T) = \sum_\Gamma \left(\frac{\exp(-\beta H_s(\Gamma))}{\Sigma_\Gamma \exp(-\beta H_s(\Gamma))} \right)^2,$$

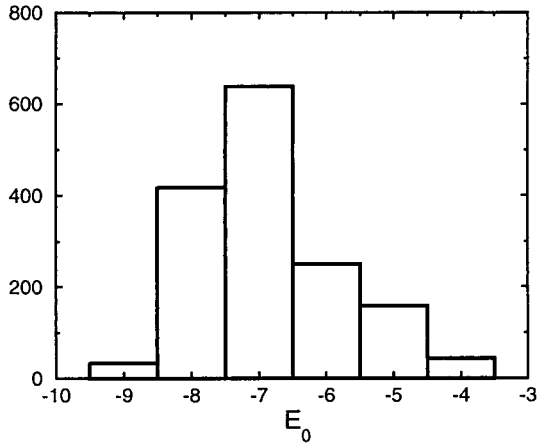

FIG. 2. The histogram of the ground-state energies of the 1,569 good sequences of 16 beads with two types of amino acids (HP model) on a square lattice.

the unprimed sum is over all conformations and the primed sum is again over selected target conformations that house the designed sequences. The selected conformations in Eq. 1 are the maximally compact ones (conformations having the largest number of contacts), whereas those in Eq. 5 are the good conformations—conformations that are the unique native state of at least one sequence (some maximally compact conformations may be good as also ones that are not maximally compact).

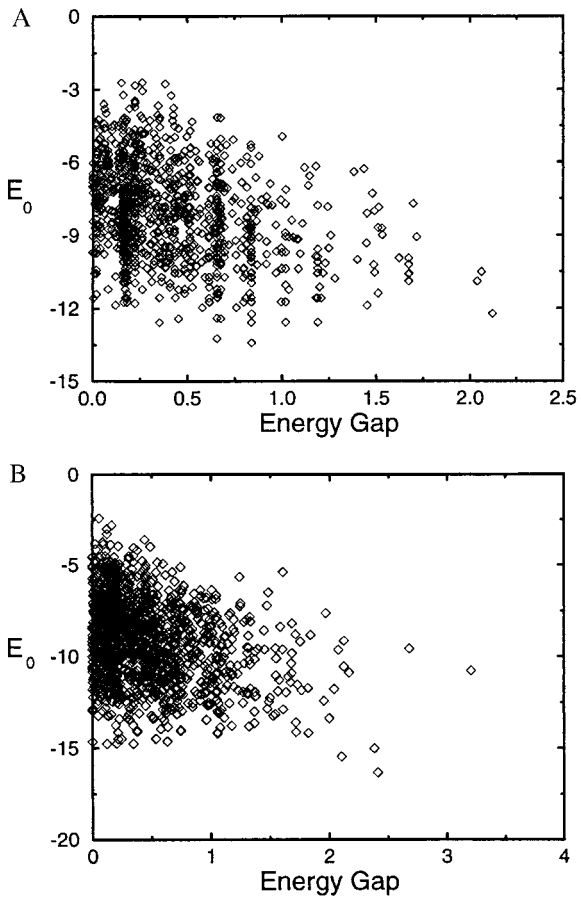

FIG. 3. (*A*) The plot of the ground-state energies of 1,146 randomly picked good sequences of 16 beads with 4 types of amino acids versus the energy gaps. (*B*) The plot of the ground-state energies of 1,517 randomly picked good sequences of 16 beads with 8 types of amino acids versus the energy gaps.

Physically, as $T \to \infty$, one expects that any sequence will have an equal probability to be in any of the numerous conformations available, thus making the order parameter small. In the $T \to 0$ limit, if the selected sequences have a nondegenerate ground state, then the order parameter approaches 1. If the selected sequences have degenerate ground states with a degeneracy $g(s)$, then $\langle P^2(0, 0)\rangle$ is less than 1 and given by

$$\langle P^2(0,0)\rangle \to \frac{\Sigma_s 1/g(s)}{\Sigma_s 1}, \qquad [7]$$

where the sum is over these sequences.

tions. Each location of the chain was assigned an amino acid selected at random with equal probability and the interaction energy matrix was taken to be

$$\begin{pmatrix} -0.485 & -1.677 & 0.697 & -0.223 \\ -1.677 & -0.837 & 0.182 & -0.656 \\ 0.697 & 0.182 & -0.676 & 0.160 \\ -0.223 & -0.656 & 0.160 & 0.515 \end{pmatrix}.$$

(The results are qualitatively the same for other sets of interaction parameters). The corresponding matrix for the 8 amino acid model (H1, H2, H3, H4, P1, P2, P3, P4) was

$$\begin{pmatrix} -0.522 & 0.144 & -0.966 & -1.679 & 0.638 & 0.745 & -0.378 & 0.395 \\ 0.144 & -1.078 & -2.380 & -1.682 & -0.002 & 1.407 & -0.947 & 0.037 \\ -0.966 & -2.380 & -0.516 & -0.451 & -0.120 & -0.309 & 0.172 & -1.506 \\ -1.679 & -1.682 & -0.451 & -1.543 & 0.160 & 0.998 & -1.044 & -0.697 \\ 0.638 & -0.002 & -0.120 & 0.160 & -1.370 & -0.448 & -0.219 & -0.193 \\ 0.745 & 1.407 & -0.309 & 0.998 & -0.448 & -0.210 & 1.007 & 0.050 \\ -0.378 & -0.947 & 0.172 & -1.044 & -0.219 & 1.007 & 0.950 & 0.781 \\ 0.395 & 0.037 & -1.056 & -0.697 & -0.193 & 0.050 & 0.781 & -0.184 \end{pmatrix}.$$

Let us now specialize to the HP (hydrophobic H and polar P) model introduced by Lau and Dill (31). It has been demonstrated (4, 31) that the properties of real proteins are mimicked reasonably well by those of chains of N beads made of only two types of amino acids (hydrophobic H and polar P) with the conformational space consisting of all self-avoiding walks on a two-dimensional square lattice. The advantage of the model is that, for moderate values of N, one may exactly enumerate both the sequences and the conformations. The interaction energies between the two types of amino acids are set to the values $\varepsilon_{HH} = -1$, $\varepsilon_{HP} = 0$, and $\varepsilon_{PP} = 0$.

Our numerical calculations using the HP model for chains with $N = 9$ and $N = 12$ are summarized in Fig. 1. The two sets of data represented by circles were determined using Eq. 5 for $P_s(\Gamma, 0)$, whereas the two sets of data represented by triangles were determined using Eq. 1 for $P_s(\Gamma, 0)$. Strikingly, the behavior at low temperature is qualitatively different in the two cases. For the HP model, the great majority of sequences with the ground state in maximally compact conformations with the lowest possible energy are degenerate with $g(s)$ that increases with the length N of the chain. This accounts for the low value of the order parameter as measured from Eq. 1. Furthermore, the size dependence (from Fig. 1) is also different for the two cases, with Eq. 5 leading to the correct behavior.

For the HP model, one may define protein-like sequences as those that have a unique ground-state conformation (and an associated energy gap between the ground state and the first excited state that is at least 1). Fig. 2 shows a histogram of the ground-state energies of the 1,569 such sequences for $N = 16$. The key point is that the selection procedure based on Eq. 1 would pick out all sequences (including the trivial HHHHHHHHHHHHHHHH sequence) that have a ground-state energy of -9 (corresponding to a maximally compact conformation with the maximum number of 9 HH possible contacts) irrespective of the ground-state degeneracy. The sequences thus selected would not be representative of the 1,569 protein-like sequences, which have a range of ground-state energies.

To assess the role of the number of types of amino acids in possibly removing the degeneracy, we now proceed to consider the $N = 16$ model, but with 4 and then 8 types of amino acids. The first model consisted of an ensemble of chains with 16 beads made of 4 types of amino acids (H1, H2, P1, P2) mounted on all possible 802,075 two-dimensional conformations. Approximately 60% of the former sequences have a unique ground state, whereas this number increases to 76% for the case with 8 amino acids.

Two measures (2, 6, 20, 33, 34) of the thermodynamic stability of a sequence in its native state are the energy gap, defined as the difference between the first excited state and the native state energies, and the z-score z_s given by

$$z_s(s) = \frac{\langle E\rangle - E_0}{\sigma}. \qquad [8]$$

Here, $\langle E\rangle$ and σ are the average energy of a sequence s over all alternative conformations and the corresponding SD, respectively. E_0 represents the ground-state energy of that sequence.

For each sequence with 16 beads the alternative conformations were taken to be all conformations with 6, 7, 8, and 9 contacts (a total of 30,169 conformations), but the native one. The graphs of the ground-state energies of these sequences versus the corresponding energy gaps (Fig. 3) indicate a broad distribution of ground-state energies. Notably, protein-like heteropolymers with a high thermodynamic stability characterized by large energy gaps do not necessarily have the lowest ground-state energy.

We turn now to a three-dimensional lattice model that has been considered standard for heteropolymer freezing studies (7). The sequences have 27 beads made up of all 20 types of amino acids and the space of conformations is restricted to the 103,346 maximally compact conformations that fit on a $3 \times 3 \times 3$ lattice. Such a situation is expected to occur in this coarse-grained model of a protein when there is an overall attractive interaction between the amino acids. Each location of the chain was assigned an amino acid generated according to its frequency of occurrence in nature (35) and the 210 interaction energies between the amino acids were taken from table 3 of Miyazawa and Jernigan (36). For such a model, the great majority of sequences (approximately 90%) have nondegenerate ground states, so that a protein-like sequence might be defined as one having a thermodynamically stable ground state. Fig. 4*A* represents a plot of the ground-state energies of good sequences versus their energy gaps, and Fig. 4*B* is a graph of the ground-state energies versus z_s. Here, as alternative conformations for each sequence, we took all maximally compact conformations but the native one. There are two notable features: first, the sequences having the lowest value of

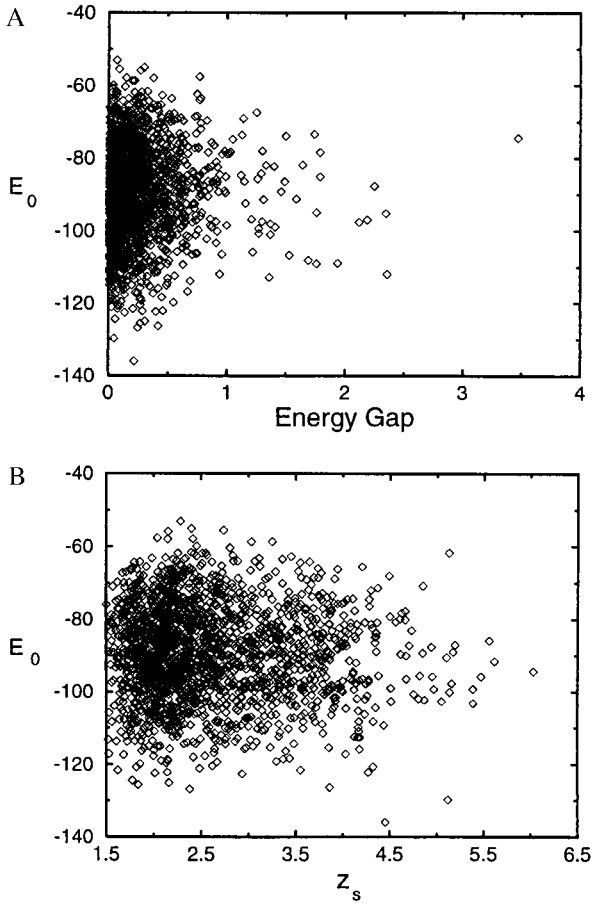

FIG. 4. (*A*) The plot of the ground-state energies of 2,012 randomly picked good sequences of 27 beads with 20 types of amino acids versus the energy gaps. (*B*) The plot of the ground-state energies of the same sequences as in A versus z_s.

the ground-state energy (which would be the ones selected using Eq. 1) do not have the highest energy gap or z_s; second, even if these could be considered protein-like sequences, there are many other equally good sequences that are not taken into account by Eq. 1 simply because their ground-state energies are not equal to the lowest one in the ensemble.

We conclude with some general observations. First, the rigorous approach does not require any constraints on the composition of amino acids. Second, an improper modification of Eq. 4 by allowing the sum over Γ to extend over *all* conformations would lead to a wrong result in which the average free energy $\langle F \rangle$ becomes trivially independent of $T_{\rm des}$ because $\Sigma_\Gamma P_s(\Gamma, T_{\rm des}) = 1$. Third, as pointed out before, it is useful to consider the $T_{\rm des} = 0$ case explicitly. There are two possible scenarios for the ground state of typical sequences. For simple models, such as the HP model of Lau and Dill (31), some sequences have a unique ground state, whereas most of them have degenerate ground states. The more generic situation is one in which, because of 20 kinds of amino acids and a more realistic interaction matrix, virtually each sequence has a unique ground state. However, the "well-designed" sequences have high thermodynamic stability with a small density of low-lying excited energy states. In such a scenario, for small values of $T_{\rm des}$, one obtains a sensible $\langle F \rangle$ as an average over predominantly those sequences that have a stability gap larger than or equal to $T_{\rm des}$. The order of taking limits of $T_{\rm des} \to 0$ and the system size going to ∞ do not commute and the correct approach would be to allow $T_{\rm des} \to 0$ after the thermodynamic limit is taken. Indeed, if one were to consider $T_{\rm des} = 0$ explicitly for finite systems, one would get the same result as for $T_{\rm des} = \infty$.

We are indebted to Sasha Grosberg and Vijay Pande for useful correspondence. This work was supported by Istituto Nazionale di Fisica Nucleare (Italy), Komitet Badan Naukowych Grant 2P03B-025-13, National Aeronautics and Space Administration, North Atlantic Treaty Organization, and the Petroleum Research Fund administered by the American Chemical Society.

1. Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. (1995) *Science* **267**, 1619–1620.
2. Bryngelson, J. D., Onuchic, J. N., Socci, J. N. & Wolynes, P. G. (1995) *Proteins* **21**, 167–195.
3. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528.
4. Dill, K. A., Bromberg, S. Yue, S., Fiebig, K., Yee, K. M., Thomas, D. P. & Chan, H. S. (1995) *Protein Sci.* **4**, 561–602.
5. Camacho, C. J. & Thirumalai, D. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 6369–6372.
6. Klimov, D. K. & Thirumalai, D. (1996) *Phys. Rev. Lett.* **76**, 4070–4073.
7. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1999) *Rev. Mod. Phys.,* in press.
8. Wolynes, P. G. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 6170–6175.
9. Plotkin, S. S., Wang, J. & Wolynes, P. G. (1997) *J. Chem. Phys.* **106**, 2932–2948.
10. Wang, J., Plotkin, S. S. & Wolynes, P. G. (1997) *J. Phys. France I* **7**, 395–421.
11. Derrida, B. (1980) *Phys. Rev. Lett.* **45**, 79–82.
12. Pande, V. S., Grosberg, A. Y., Joerg, C. & Tanaka, T. (1996) *Phys. Rev. Lett.* **76**, 3987–3990.
13. Bryngelson, J. D. & Wolynes, P. G. (1989) *J. Phys. Chem.* **93**, 6902–6915.
14. Ramanathan, S. & Shakhnovich, E. I. (1994) *Phys. Rev. E* **50**, 1303–1312.
15. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 12976–12979.
16. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1995) *Macromolecules* **28**, 2218–2227.
17. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1995) *J. Chem. Phys.* **103**, 9482–9491.
18. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1995) *J. Phys. A* **28**, 3657–3666.
19. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1997) *Physica D* **107**, 316–321.
20. Shakhnovich, E. I. & Gutin, A. M. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 7195–7199.
21. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1994) *J. Phys. France II* **4**, 1771–1784.
22. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1994) *J. Chem. Phys.* **101**, 8246–8257.
23. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1995) *Phys. Rev. E* **51**, 3381–3392.
24. Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z. & Socci, N. D. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 3626–3630.
25. Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 1282–1286.
26. Shakhnovich, E. I. & Gutin, A. (1989) *Biophys. Chem.* **34**, 187–199.
27. Kurosky, T. & Deutsch, J. M. (1995) *J. Phys. A* **28**, l387–l393.
28. Deutsch, J. M. & Kurosky, T. (1996) *Phys. Rev. Lett.* **76**, 323–326.
29. Mirny, L. A. & Shakhnovich, E. I. (1996) *J. Mol. Biol.* **264**, 1164–1179.
30. Seno, F., Vendruscolo, M., Maritan, A. & Banavar, J. R. (1996) *Phys. Rev. Lett.* **77**, 1901–1904.
31. Lau, K. F. & Dill, K. A. (1989) *Macromolecules* **22**, 3986–3997.
32. Chan, H. S. & Dill, K. A. (1993) *Phys. Today* (Feb.), 24–32.
33. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4918–4922.
34. Bowie, J. U., Luthy, R. & Eisenberg, D. (1991) *Science* **253**, 164–170.
35. Creighton, T. E. (1993) *Proteins: Structures and Molecular Properties* (Freeman, New York), p. 4.
36. Miyazawa, S. & Jernigan, R. (1996) *J. Mol. Biol.* **256**, 623–644.