Eur. Phys. J. B **8**, 323–326 (1999)

THE EUROPEAN
PHYSICAL JOURNAL B
EDP Sciences

Hydrophobicity and unique folding of selected polymers

M. Vendruscolo[a]

Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel

Received: 4 June 1998

Abstract. In a suitable environments, proteins, nucleic acids and certain synthetic polymers fold into unique conformations. This work shows that it is possible to construct lattice models of foldable heteropolymers by expressing the energy only in terms of individual properties of monomers, which model exposure to the solvent and the steric factor.

PACS. 87.15.By Structure, bonding, conformation, configuration, and isomerism of biomolecules – 87.10.+e General, theoretical, and mathematical biophysics (including logic of biosystems, quantum biology, and relevant aspects of thermodynamics, information theory, cybernetics, and bionics)

It is generally believed that the hydrophobic interaction plays a major role in protein folding [1–9]. Under physiological conditions, non polar amino acids are buried inside the core of the native state of a protein to avoid contact with water molecules. A long standing question is to what extent other non-covalent forces, such as hydrogen bonding, electrostatic and van der Waals interactions contribute to stabilize the folded state [1–3].

Unraveling the different roles played by these interactions will have a considerable impact in different areas of research in biophysics, such as the prediction of protein structures [1,2,4–6] the design of synthetic drugs [5,10–12], and the production of self-assembling non-biological polymers [13] and other polymeric materials [14].

With the advent of genome projects [15] a wide gap is opening between the number of known protein sequences and their correspondent structures [16]. The bottleneck in protein structure prediction is at present largely due to the incorrect treatment of the interactions [17–19]. Various techniques to single out the native state of a protein from a library of alternative structures are typically carried out by assigning an energy-like function that incorporates the compatibility of each amino acid to its local environment [20]. Compatibility is described in terms of charge, polarity and secondary structure content, within a given conformation. Details in the local environment play a crucial role also in RNA folding. A key ingredient in this case is given by metal ion coordination numbers [21]. Likewise and rather surprisingly, a non-biological polymer (an aromatic hydrocarbon) has been recently designed which is able to fold into a unique helical structure having a large cavity, supposedly under the effect of the hydrophobic interaction [13].

This study contributes to the development of a rational treatment of interactions in heteropolymers at the single monomer level. We show that it is possible to construct minimalistic lattice models of foldable heteropolymers, by introducing an energy function that depends only on individual residues' environments. A model will be called "foldable" if there are sequences, either randomly chosen or selected, with a unique, thermodynamically stable and kinetically reachable ground state [2,3,5–8,19,22–24]. We adopt a simple approximation for the energy which accounts both for the propensity to be exposed to the solvent and for the excluded volume effects due to the different sizes of the monomers. Although naturally existing or synthesized polymers, such as proteins, nucleic acids and tailored hydrocarbons, are characterized by much more complex interactions, the main focus here is on the fact that the unifying feature is the tendency to avoid contact with the solvent by some species of monomers. Previous theoretical studies concentrated mainly on the treatment of pairwise contact energies [2,17,19,25–29]. This is in contrast with the present study, in which the hydrophobic effect is investigated at the individual particle level.

Lattice models, although often criticized [30], have been recognized to capture some of the most relevant thermodynamic features of the folding process [31], such as the existence of a unique ground state, amenable to exact computations, and the cooperativity of the transition. Even key dynamical processes, such as the nucleation-condensation mechanism [32], have been validated with the help of lattice models [33].

On a lattice, a polymer is represented as a connected chain of N monomers. Hydrophobicity and steric factors can be modeled as the tendency of a monomer to have a specific number of non-bonded nearest-neighbors. We define the hydrophobic model HM_1 by expressing the energy

[a] e-mail: femichel@wicc.weizmann.ac.il

E_1 as,

$$E_1 = \sum_{i=1}^{N} |n_i - \overline{n}(a_i)| \ , \qquad (1)$$

where n_i is the number of non-bonded nearest-neighbors of monomer of species a_i in position i along the chain, and $\overline{n}(a_i)$ is the ideal value of n_i. This expression was first proposed by Hao and Scheraga, who considered it together with a pairwise energy term [34]. They presented a method to optimize energy parameters to obtain lattice models of foldable polymers. Other previous work has been devoted to the hydrophobic interaction, although without specifically disentangling it from other interactions. Mirny and Domany [29] introduced explicitly an hydrophobic term in the energy function and they performed various tests of fold recognition and dynamics. In a recent work Li, Tang and Wingreen [35] discussed the "designability principle" [8] in terms of a "binary" model with two species of amino acids, where the energy is expressed in terms of the exposure to the solvent only. The model proposed here is more general and no major modification is required to extend it to the treatment of realistic models of foldable heteropolymers, as for example, in a "contact map" representation of protein structure [29,36].

As a preliminary step, we first explore 2D lattice models, restricting our attention to the existence of a unique ground state structure, disregarding the thermodynamic and the kinetic issues and we compare our results with those obtained using the standard HP model [2]. The success of the HP model is due to the fact that in 2D a 2% fraction of all the possible sequences has a unique ground state [2,7,8,12], although probably the folding transition is of the wrong order [6]. In the HP model the energy is written in the pairwise contact approximation

$$E_{pair} = \sum_{j>i} U(a_i, a_j) \Delta_{ij} \ , \qquad (2)$$

where a_i can be either H (hydrophobic) or P (polar) and Δ_{ij} is a contact matrix, which is defined to be 1 if two monomers are non-bonded nearest-neighbor and 0 otherwise. Typical values for the interaction parameters are $U(H,H) = -1$ and $U(H,P) = U(P,P) = 0$ [2]. A chain of $N = 16$ monomers is amenable to complete enumeration of all 802 075 possible symmetry-unrelated conformations, either compact or not [2,11,12]. For the above mentioned choice of contact energy parameters, there are 1539 (2%) sequences among the $2^{16} = 65\,536$ possible ones which have a unique ground state [11,12]. We compare this result with those obtained by using equation (1), setting $\overline{n}(1) = 1.5$ (hydrophobic-like) and $\overline{n}(2) = 0.4$ (polar-like). A larger number, 10 178 (16%), of sequences was found to have a unique ground state.

We have also explored the case of three species of monomers, a number which, within a contact approximation of the interactions (as in Eq. (2)), is believed to epitomize the essential features of the interplay between folding and glass transitions in random heteropolymers [37]. We chose at random 20 177 sequences among the 2 018 016 possible ones with fixed composition $N(1) = 6$, $N(2) = N(3) = 5$, where $N(a)$ is the number of monomers of species a. Choosing energy parameters $\overline{n}(1) = 0.4$, $\overline{n}(2) = 1.1$ and $\overline{n}(3) = 1.8$, 9439 (47%) sequences were found with a unique ground state.

In the spirit of Mirny and Domany [29], a more accurate form for the hydrophobic energy is given by

$$E_2 = \sum_{i=1}^{N} \beta(a_i) \left[n_i - \overline{n}(a_i)\right]^2 \ . \qquad (3)$$

This expression will be referred to as hydrophobic model HM$_2$. The parameters $\beta(a_i)$ capture the various degrees with which the different species of monomers tend to attain the preferred number $\overline{n}(a_i)$ of contacts. In the case of 2 species of monomers, we repeated the same calculation as for the HM$_1$ model. Letting here $\beta(1) = \beta(2) = 1$, we found 9821 (14%) sequences with a unique ground state. We observe that, at least from the above calculations in 2D, the approximation of the hydrophobic interaction proposed in this work is capable of yielding foldable sequences.

We now turn to the calculations in 3D which represent the essential part of this work, and further illustrate the extent to which the present model embodies foldability. It is known that the HP model is pathological in 3D, since it is rather uncommon to have a sequence with a unique ground state with a large gap above it [6,38], although the situation can be different with a choice of parameters favoring more collapsed structures [8,19]. We discuss here the general case of 20 species of monomers in the HM$_2$ model. We compare these results with those obtained by using a common parametrization of the pairwise contact interaction matrix $U(a_i, a_j)$ in equation (2), due to Miyazawa and Jernigan (MJ) [25], although other choices would be possible [25–29]. For the HM$_2$ model, we derived the 40 parameters $\overline{n}(a)$ and $\beta(a)$ from a statistical analysis of the non redundant set of 246 protein structures reported by Hinds and Levitt [27]. The procedure, similar to that of Mirny and Domany [29], is straightforward. For each amino acid species a, we computed the average, $\overline{n}(a)$, and the standard deviation, $\beta(a)$, of the number of contacts it forms in the set of experimentally known crystal structures (see Tab. 1). Two amino acids are said to be in contact if their C_α atoms are closer than 8.5 Å in the native structure [36].

On the cubic lattice, the 103 346 symmetry-unrelated maximally compact conformations of a polymer of length $N = 27$ can be enumerated in a manageable computer time [8,40]. If it is guaranteed that the ground state is maximally compact, exact enumeration can be used to demonstrate its uniqueness. We adapted the energy parameters to the cubic lattice by matching the average number of contacts that a monomer forms on the $3 \times 3 \times 3$ cube with the average of the ideal number of contacts, $(1/20) \sum_{a=1,20} \overline{n}(a)$. This result is obtained by rescaling the energy parameters in Table 1 by a factor 3.315.

To characterize foldability, we first investigate the thermodynamic stability of the ground states of random

Table 1. Mean $\overline{n}(a)$ and standard deviation $\beta(a)$ of the number of contacts of amino acids, obtained from a statistical analysis of a non redundant set of 246 protein structures.

ALA	GLU	GLN	ASP	ASN	LEU	GLY	LYS	SER	VAL	ARG	THR	PRO	ILE	MET	PHE	TYR	CYS	TRP	HIS
7.56	5.62	6.23	5.51	6.02	7.63	5.55	5.86	6.31	8.29	6.58	6.73	5.73	8.07	7.72	7.58	7.45	8.81	7.67	6.59
2.98	2.17	2.36	2.41	2.58	2.25	3.51	2.08	3.07	2.53	2.33	2.77	2.85	2.35	2.37	2.31	2.51	2.42	1.36	2.43

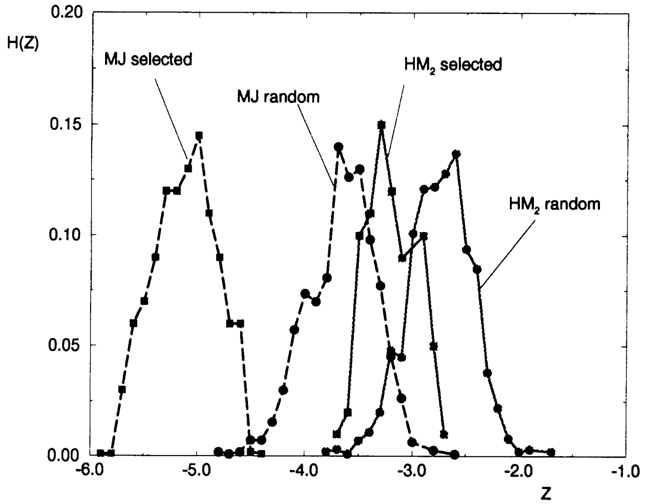

Fig. 1. Normalized histograms of the Z scores for the HM_2 (full lines) and MJ (dotted lines) models. Circles refer to random sequences and squares to designed ones.

sequences. A typical measure of thermodynamic stability is given by the Z score [19,28], which is defined by $Z = (E_n - \langle E \rangle)/\sigma$, where E_n is the energy of the ground state, $\langle E \rangle$ is the average energy, and σ the standard deviation in the distribution of the energy around the average. We measure the distribution of the Z scores for a set of 1000 random HM_2 sequences. We found that only 2% of them had a unique lowest energy state (the "ground state" among maximally compact conformations). Moreover, on average the degeneracy was 22.

For comparison, 99% of the 1000 random MJ sequences that we considered had a non degenerate lowest energy state, and the remaining ones had a very small degeneracy. The result of the comparison of the Z scores is shown in Figure 1.

The analysis of kinetic accessibility of the identified lowest energy states, together with the low values of the Z score and the large degeneracy associated to them mark a shortcoming of enumerating only maximally compact conformations. By using other simulation techniques, such as the standard lattice Monte-Carlo (SMC) [22,39], and the prune-enriched Rosenbluth method (PERM) [42], we easily found non-compact lower energy states for most of the considered HM_2 sequences.

The former analysis was carried out on random sequences, whereas foldability is believed to be a property of *selected* sequences [4–6,9–12,23,24]. A way to demonstrate that the HM_2 model is foldable is to show that it is possible to select sequences whose ground states are both unique *and* maximally compact. The usual design procedure [10], introduced to study pairwise interactions, prescribes to choose a target conformation and then to search in sequence space for the sequence with minimal energy onto such conformations. This procedure delivers a better Z score for the 1000 designed MJ sequences that we considered, as can be seen from Figure 1. However, in the case of the HM_2 model, we found that such technique is not sufficiently effective in designing out alternative conformations. A sequence design procedure similar to those proposed in references [11,12] proved to be more effective. Sequences selected in this way were found to have a unique ground state by exact enumeration among maximally compact conformations. More crucially, in no cases we have been able to reach lower energy states using dynamical simulation techniques such as the SMC and the PERM algorithms. The histogram of the Z score of the 100 HM_2 sequences selected in this way is shown in Figure 1.

We summarize our results for the HM_2 model. We first showed that 2% of randomly selected sequences have a unique lowest energy state when all maximally compact conformations are enumerated. Second, we showed that sequence design is effective in identifying foldable sequences. The sequence selection procedure is more effective in improving the Z score of the MJ model than that of the HM_2 model, as shown in Figure 1.

How should these results be interpreted? In a recent paper, Li *et al.* [41] showed that the MJ interaction matrix can be accurately expressed by using only 20 hydrophobicity-related parameters each one associated with an amino acid species. This paper is consistent with Li *et al.* results about the importance of the hydrophobic effect, but it proposes a different way to encode it. The considerably smaller fraction of sequences with a unique ground state found for the HM_2 model is probably an artifact of the lattice model used. We found that in all the cases of degeneracy considered, a sequence has the same energy on two different structures if the 20 parameters n_i $(i = 1, \ldots, 20)$ are identical for the two structures. This situation is unlikely to extend to the off-lattice case. Furthermore, results by Mirny and Domany [29] indicates that an energy term like equation (3) gives a good correlation between the energy gap and the conformational distance from the ground state.

Different ways to encode hydrophobicity can help increasing specificity of the model. Work in this direction is timely, since it has been proved that pairwise contact Hamiltonians alone (such as the MJ model) are unsuitable for folding real proteins [43]. It is therefore of critical importance to improve the energy function used for protein folding. In this spirit we have analyzed the behavior of an energy term which expresses the hydrophobic and the steric interactions at the level of individual monomers. We have shown that this term alone is capable to give rise to foldable models. The message we get from this conclusion is that it is very promising to undertake a study of realistic models of off-lattice proteins adopting a combination of pairwise and hydrophobic terms.

It is a pleasure to thank E. Domany and P. Grassberger for discussions.

Note added

After submission of the manuscript, we become aware of a related independent study by Micheletti *et al.* [44].

References

1. *Protein folding*, edited by T.E. Creighton (W.H. Freeman and Company, New York, 1992).
2. K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, H.S. Chan, Protein Sci. **3**, 561 (1995).
3. J.D. Bryngelson, P.G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
4. T. Garel, H. Orland, D. Thirumalai, in *New Developments in Theoretical Studies of Proteins*, edited by R. Elber (World Scientific, Singapore, 1996).
5. V.S. Pande, A.Yu. Grosberg, T. Tanaka, Rev. Mod. Phys (in press).
6. E.I. Shakhnovich, Curr. Opin. Struct. Biol. **7**, 29 (1997).
7. C. Camacho, D. Thirumalai, Proc. Natl. Acad. Sci. USA **90**, 6369 (1993).
8. H. Li, R. Hellig, C. Tang, N. Wingreen, Science **273**, 666 (1996).
9. A. Irbäck, C. Peterson, F. Potthast, Proc. Natl. Acad. Sci. USA **93**, 9533 (1996).
10. E.I. Shakhnovich, A.M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).
11. J.M. Deutsch, T. Kurosky, Phys. Rev. Lett. **76**, 323 (1996).
12. F. Seno, M. Vendruscolo, A. Maritan, J.R. Banavar, Phys. Rev. Lett. **77**, 1901 (1996).
13. J.C. Nelson, J.G. Saven, J.S. Moore, P.G. Wolynes, Science **277**, 1793 (1997).
14. M. Muthukumar, C.K. Ober, E.L. Thomas, Science **277**, 1225 (1997).
15. D. Duboule, Science **278**, 555 (1997) and references therein.
16. J.L. Sussman, Nature Struct. Biol. **4**, 517 (1997).
17. V.S. Pande, A.Yu. Grosberg, T. Tanaka, J. Chem. Phys. **103**, 9482 (1995).
18. A. Finkelstein, Curr. Opin. Struct. Biol. **7**, 60 (1997).
19. M. Vendruscolo, A. Maritan, J.R. Banavar, Phys. Rev. Lett. **78**, 3967 (1997).
20. J.U. Bowie, R. Luethy, D. Eisenberg, Science **253**, 164 (1991).
21. J.A. Doudna, E.A. Doherty, Fold. Des. **2**, R65 (1997).
22. A. Sali, E.I. Shakhnovich, M. Karplus, Nature **369**, 248 (1994).
23. J.D. Bryngelson, J.N. Onuchic, N.D. Socci, P.G. Wolynes, Proteins **21**, 167 (1995).
24. T. Veitshans, D.K. Klimov, D. Thirumalai, Fold. Des. **2**, 1 (1997).
25. S. Miyazawa, R. Jernigan, J. Mol. Biol. **256**, 623 (1996).
26. R.A. Goldstein, Z.A. Luthey-Shulten, P.G. Wolynes, Proc. Natl. Acad. Sci. USA **89**, 4918 (1992).
27. D.A. Hinds, M. Levitt, J. Mol. Biol. **243**, 668 (1994).
28. L. Mirny, E.I. Shakhnovich, J. Mol. Biol. **264**, 1164 (1996).
29. L. Mirny, E. Domany, Proteins **26**, 391 (1996).
30. B. Honig, F. Cohen, Fold. Des. **1**, R17 (1996).
31. E.I. Shakhnovich, Fold. Des. **1**, R50 (1996).
32. A.R. Fersht, Proc. Natl. Acad. Sci. USA **92**, 10869 (1995).
33. E.I. Shakhnovich, V.I. Abkevich, O. Ptitsyn, Nature **379**, 96 (1996).
34. M.H. Hao, H.A. Scheraga, Physica A **244**, 124 (1997).
35. H. Li, C. Tang, N. Wingreen, Proc. Natl. Acad. Sci. USA **95**, 4987 (1998)
36. M. Vendruscolo, E. Kussell, E. Domany, Fold. Des. **2**, 295 (1997).
37. J.N. Onuchic, N.D. Socci, Z. Luthey-Shulten, P.G. Wolynes, Proc. Natl. Acad. Sci. USA **92** (1995) 3626.
38. K. Yue, K. Fiebig, P. Thomas, H.S. Chan, E.I. Shakhnovich, K.A. Dill, Proc. Natl. Acad. Sci. USA **92**, 325 (1995).
39. N.D. Socci, J.N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).
40. E.I. Shakhnovich, A. Gutin, J. Chem. Phys. **93**, 5967 (1990).
41. H. Li, C. Tang, N. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).
42. P. Grassberger, Phys. Rev. E **56**, 3682 (1997).
43. M. Vendruscolo, E. Domany, J. Chem. Phys. **109**, 11101 (1998).
44. C. Micheletti, J.R. Banavar, A. Maritan, F. Seno, Phys. Rev. Lett. **80**, 5683 (1998).