Predicting the Tertiary Structure of a Lattice Designed Model Protein from its Primary Structure

R.A. BROGLIA¹ and G. TIANA²

¹*Dipartimento di Fisica, Università di Milano and INFN, sezione di Milano, via Celoria 16, 20133 Milano, Italy*

²*The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 16, 2100 Copenhagen, Denmark*

Abstract. Through systematic studies of lattice Monte Carlo simulations of the folding of designed heteropolymers, we have identified a hierarchy of specific elementary phenomena which control the way single domain protein fold: a) formation of few, local elementary structures, b) creation of the (post-critical) folding nucleus through the assemblage together of the local elementary structures, c) relaxation of the remaining amino acids to the native conformation. These results, which are consistent with a two-state kinetics of the folding of small, single domain proteins, where the local elementary structures and the folding nucleus can be viewed as hidden intermediates along the reaction pathway, provide the basis for a strategy to read the tertiary structure of a protein from its amino acid sequence.

Key words: model proteins, kinetics, hierarchical folding, prediction

An answer to the question of how to relate the protein primary structure to its full three-dimensional structure has long been awaited to come from a proper understanding of the mechanisms which are at the basis of the phenomenon of protein folding [1, 2, 3, 4]. The fulfillment of such expectation is likely to revolution the therapeutic drug industry, let alone the chemistry of enzymatic processes. It may also be within reach, in any case, to the extent that simple lattice models of protein folding do capture some of the essential properties of real proteins.

Although Anfinsen and collaborators had conclusively shown [5] that the 1D-structure of a protein determines its 3D-structure, researchers had long been stymied in their efforts to predict the latter from the knowledge of the former, because of Levinthal's paradox [6]: the number of all possible conformations of a polypeptide chain is too large to be sampled exhaustively. Nevertheless, proteins do fold into unique native states in seconds.

An important step forward in the study of protein folding was made by devising a simple (although not oversimplified) approach to the problem [7], in which the quest for the relation existing between the 1D- and the 3D-structure of a notional protein is formulated in such a way that Levinthal's paradox is circumvented. In essence, this approach turns the problem of protein folding inside out. First, a native conformation is chosen and then, the notional protein designed by minimizing the energy of the system with respect to the amino acid sequence (for fixed composi-

Figure 1. Snapshots of the folding of the designed sequence S_{36} (cf. Figure 2(a)), whose energy in the native conformation is $E_n = -16.50$. Starting from a random conformation (a), the system forms after $\approx 10^2$ MC steps local elementary structures (LES) (b), involving three sets of four amino acids (3–6, 11–14, 17–30), whose stability is provided by the bonding indicated by dotted lines. When the LES come together to form the folding core (indicate by dotted and dashed lines) after $7 \cdot 10^5$ MC steps (c), the system folds to the native conformation after only 10^3 MC steps (d). The amino acids participating in the bonding of the LES (dotted lines) are among some of the most strongly interacting amino acids, which occupy, in the native conformation (d), ‘hot’ and ‘warm’ sites [23] indicated by dark- and light-gray beads, respectively (red and yellow in Figure 2). The monomers numbers 1 and 36 of the sequence S_{36} are indicated for each conformation.

tion). To be noted that the method, although working well within the framework of a lattice model, is an approximation to the complete solution of the folding model problem [8, 9, 10].

Using a 20-letter three dimensional lattice model for heteropolymers [7, 11, 12, 13, 14, 15, 16, 17, 18], and contact energies obtained from a statistical analysis of real proteins [19], one finds that good-folder-sequences are characterized by a large gap $\delta = E_n - E_c$ (compared to the standard deviation σ of the contact energies) between the energy E_n of the designed sequence in the native conformation and the lowest energy (threshold E_c) of the compact conformations structurally dissimilar to the native conformation. In other words, Monte Carlo (MC) simulations testify to the fact that designed sequences displaying a large normalized gap ($\xi = \delta/\sigma \gg 1$, quantity closely related to the z-score [20]) in the native conformation fold on short call [21]. The success of the inverse folding model is connected with the fact that

good folders share a (small) number of conserved contacts (which eventually form the folding nucleus of the protein), contacts which act among the most strongly interacting and best conserved amino acids (so called ‘hot’ and ‘warm’ sites in Reference [23]).

We now know [24, 25] that this result is tantamount to saying that foldability is controlled by the presence of few, local elementary structures, stabilized by the conserved contacts and formed very early in the folding process, leading to the (post-critical) folding core (i.e., the minimum set of amino acids and associated native contacts needed to ensure foldability [26, 27], set whose existence has been experimentally confirmed [28]) when they assemble together. An example of this behaviour is provided in Figure 1 where snapshots of the time evolution of a designed 36mer (cf. Figure 2(a)) *en route* towards its native conformation are shown. Starting from a random coil (Figure 1(a)), the local elementary structures are formed only after $\approx 10^2$ steps of Monte Carlo (MC) simulations (Figure 1(b)), when some of the most strongly interacting amino acids establish their local contacts. They achieve $\approx 80 - 90\%$ stability after 10^5 MC steps, and when they assemble together after $\approx 10^6$ MC steps (Figure 1(c)) they form the (post-critical) folding nucleus of the notional protein, from which it reaches the native conformation (Figure 1(d)) in a short time ($\approx 10^3$ MC steps), provided the energy of the system is lower than E_c . Partially folded intermediates and not the individual monomers thus take care, through local guidance and non-local long range correlations (bonding between partially folded intermediates), of the process of protein folding, as testified by the disruptive effect mutations which affect the stability of these structures have on the folding ability of the designed sequence.

From this vantage point of view it is easy to see how to solve the protein folding problem, that is how to predict the 3D-structure of a protein, from its 1D-structure and the contact energies used in the design of the protein (for details we refer to the Appendix): 1) starting from a notional sequence and making use of the contact energies used in its design find the local elementary structures, 2) determine the possible folding cores by allowing the local elementary structures to interact among them, 3) relax the position of the remaining amino acids and determine the corresponding energy. The (single) compact structure which displays an energy smaller than E_c is the native conformation. This is in keeping with the fact that all sequences which in the compaction process display local elementary structures and thus a folding core will fold to a unique conformation, provided the associated total energy lies below E_c [25]. We have applied this strategy to representative members of essentially all the classes of lattice designed sequences available in the literature: 27mers [29, 30, 31], 36mers [23, 24, 32, 33, 34, 35, 36] 48mers [26] and 80mers [7]. In all cases the predicted native structure coincides with the compact structure used to design the protein by carrying out on it simulated annealing in sequence space.

An example of the results obtained by applying the strategy described above to a designed 36mer is shown in Figure 2. The results displayed in Figures 2(b) and

Figure 2.

←

Figure 2. Elements in the prediction of the 3D-structure of a notional protein starting from its 1D-structure: (a) designed 36mer sequence (1D structure) given as the input to solve the 1D $\rightarrow$ 3D puzzle, together with the 20×20 contact energies among the amino acids (Table 6 of Reference [19]), (b) local elementary structures obtained following protocol #1 of the Appendix. The amino acids participating in the contacts of these closed structures have been drawn in (a) in white colours, (c) only (post critical) folding core designed making use of the elementary structures and of the contact energies according to protocol #3 of the Appendix, to which is associated a compact conformation (shown in (d)) obtained by relaxing the amino acids not belonging to the core (according to protocol #4 of the Appendix), displaying an energy lower than E_c ($= -14.1$, cf. Reference [37]). Accordingly, the conformation (d) is the predicted wild type (3D) native conformation of the notional 1D-sequence shown in (a) (where the hot, warm and cold sites of the protein in its native state, calculated according to Reference [23] are displayed in terms of red, yellow and green beads). In fact, this is the native conformation used in the literature to design the sequence (a). Furthermore, Monte Carlo simulations testify to the correctness of the prediction. In fact, evolving the sequence (a) starting from random elongated conformations, it always folds into the conformation (d) in $\approx 0.701 \cdot 10^6$ MC steps. In (e) the time evolution of all the native contacts of one particular run (in which the chain folds in $0.65 \cdot 10^6$ MC steps) is shown. In particular, that of the contacts associated with the local elementary structures (dashed yellow lines) and with the contacts between the elementary structures in the folding nucleus shown in (c) (red dotted lines). To be noted that the sequence shown in (a) and called S_{36} in the literature [23, 24, 26], was designed following the inverse folding-model according to protocol #1 of the Appendix using the structure shown in (d) as native conformation (for fixed amino acid composition). It has an energy $E_n = -16.5$, and consequently an energy gap $\delta = 2.5$.

2(c) testify to the fact that the designed sequence shown in Figure 2(a) is a good candidate to code for a notional protein. In fact, the energy associated with the contacts stabilizing the three predicted local elementary structures (yellow dashed lines, Figure 2(b)) is -2.66 in the units we are using ($RT_{room} = 0.6$ kcal/mol), the interaction energy among these structures in the (post-critical) folding core (red dashed lines, Figure 2(c)) being equal to -5.15 , the total energy of this core thus being -7.81 . Relaxing the amino acids not contained in the core, we find for the sequence shown in Figure 2(a) that the system can lower its energy below E_c (to a value -16.5 , where E_c , calculated making use of the Random Energy Model [37] is -14.1) for the conformation shown in Figure 2(d), which is thus predicted to be the native conformation of the sequence shown in Figure 2(a). In fact, this conformation coincides with the one used in the literature to design the sequence under discussion. This sequence, known as S_{36} , folds into the native conformation shown in Figure 2(d) with an (average) first passage time of $0.71 \cdot 10^6$ MC steps, following the hierarchy of steps described above (formation of elementary structures in $\approx 10^2$ MC steps, formation of the (post-critical) folding core in $\approx 0.7 \cdot 10^6$ MC steps and folding in $\approx 0.71 \cdot 10^6$ MC steps). These steps can be also observed in Figure 2(e), where the time dependence of the native contacts for a particular run used to calculate the average folding time of the protein is shown as a function of the number of steps of the Monte Carlo simulation. We have found that the

1D→3D strategy discussed above works for all designed sequences displaying a large gap when measured in terms of the standard deviation of the contact energies.

Appendix

1. Selection of the sequence

Sequences are selected within the framework of a 20 letter lattice model of proteins, making use of Monte Carlo simulations to minimize the energy of the chain in the native conformation with respect to the amino acid sequence and for fixed composition [7].

2. Search of elementary structures

We distinguish two classes of (local) elementary structures which control, within lattice model calculations, the phenomenon of protein folding: those which isolated do not display any interaction (no internal interaction), and those which do display at least one (internal) interaction. We shall call them ‘open elementary structures’ and ‘closed elementary structures’, respectively.

Candidates to the role of open elementary structures are those displaying low values of the energy density

$$\epsilon_s = (j - i)^{-1} \sum_{i \leq l \leq j} \min_{k \in (i, j)} B_{\sigma(l)\sigma(k)},$$

where $B_{\sigma(l)\sigma(k)}$ are the contact energies between the l th and the k th amino acid of the chain (Table 6 of Reference [19]). The calculations have been carried out with the condition $(j - i) \leq 10$. Candidates to closed elementary structures are formed by maximizing the value of $p(i, j) = (j - i)^{-\gamma} \exp(-\beta B_{\sigma(i)\sigma(j)})$, where $(j - i)^{-\gamma}$ is the number of conformations that the segment $i - j$ of the chain can build and $\gamma = 1.68$ has been estimated by a complete enumeration of loops for small value of $j - i$ (cf. Reference [25]), $i + 2 \leq j \leq i + 8$ and $j - i$ is odd.

3. Search of the folding core.

The energy spectrum of the low-energy conformations which can be constructed making use of the elementary structures is calculated through a complete enumeration of the conformations having an energy smaller than a given threshold. Starting from an elementary structure a second one is placed in all possible conformations with respect to the first one which have at least one contact between the two, and its energy recorded. To each of these conformations displaying an energy lower than some chosen energy, a new elementary structure is added, and the process repeated until the composite system contains all elementary structures. The calculations are repeated altering the order in which the different elementary structures are placed together and allowed to interact with each other. Making use of the resulting energy

distribution of the conformations containing 2,3,..., all, of the local elementary structures, we select as potential (post-critical) folding cores of the notional protein, those conformations having an energy lying below a given threshold energy.

4. Relaxation of the monomers not belonging to the core

For each (potential) folding core, the variety of conformations of the remaining monomers are enumerated, and the total energy of the system calculated. The conformation which has energy lower than E_c (calculated following Reference [37]) is the native conformation of the notional sequence.

References

1. Creighton, T. E.: *Proteins*, W. H. Freeman and Co., New York (1993).
2. Branden, C. and Tooze, J.: *Introduction to Protein Structure*, Garland Publishing Inc., New York (1999).
3. Fersht, A.: *Structure and Mechanism in Protein Science*, W. H. Freeman and Co., New York (1992).
4. Creighton, T. E. (ed.): *Protein Folding*, W. H. Freeman and Co., New York (1992).
5. Anfinsen, C. B.: *Science* **181**, 223 (1973).
6. Levinthal, C.: in: P. Debrunner, J. C. M. Tsibris and E. Munck (eds.), *Mössbauer Spectroscopy in Biological Systems*, Univ. Illinois Press, Urbana, 1969, pp. 22–24.
7. Shakhnovich, E. I.: *Phys. Rev. Lett.* **72**, 495–522 (1994).
8. Deustch, J. M. and Kurosky, T.: *Phys. Rev. Lett.* **76**, 323 (1996).
9. Seno, F., Vendruscolo, M., Maritan, A. and Banavar, J. R.: *Phys. Rev. Lett.* **77**, 1901 (1996).
10. Morrissey, M. P. and Shakhnovich, E.: *Folding and Design* **1**, 391 (1996).
11. Goldstein, R., Luthey-Schulten, Z.A. and Wolynes, P.: *Proc. Natl. Acad. Sci. USA* **89**, 4918–4922 (1992).
12. Sali, A., Shakhnovich, E. I. and Karplus, M.: *J. Mol. Biol.* **235**, 1614–1636 (1994).
13. Shakhnovich, E. and Gutin, A.: *Proc. Natl. Acad. Sci. USA* **90**, 7195–7199 (1993).
14. Govindarajan, S. and Goldstein, R. A.: *Biopolymers* **36**, 43–51 (1995).
15. Gutin, A., Abkevich, V. and Shakhnovich, E. I.: *Phys. Rev. Lett.* **77**, 5433 (1996).
16. Bryngelson, J., Onuchic, J.N., Socci, N.D. and Wolynes, P.: *Proteins: Struct. Funct. and Genetics* **21**, 167–195 (1995).
17. Shakhnovich, E. I.: *Curr. Opin. Struct. Biol.* **7**, 29–40 (1997).
18. Pande, V.S., Grosberg, A. Ye., Tanaka, T., Rokhsar, D. S.: *Curr. Opin. Struct. Biol.* **8**, 68–79 (1998).
19. Miyazawa, S. and Jernigan, R. L.: *Macromolecules* **18**, 534 (1985).
20. Bowie, J., Luthy, R. and Eisenberg, D.: *Science* **253**, 164 (1991).
21. An equivalent version of the model [22] posits that good-folder sequences can be obtained by minimizing the time the heteropolymer employs to target onto the chosen native conformation with respect to amino acid sequence.
22. Mirny, L. A., Abkevich, V. I. and Shakhnovich, E. I.: *Proc. Natl. Acad. Sci. USA* **95**, 4976 (1998); *ibid* **92**, 1282 (1995).
23. Tiana, G., Broglia, R. A., Roman, H. E., Vigezzi, E. and Shakhnovich, E. I.: *J. Chem. Phys.* **108**, 757–761 (1998).
24. Broglia, R. A., Tiana, G., Pasquali, S., Roman, H. E. and Vigezzi, E.: *Proc. Natl. Acad. Sci. USA* **95**, 12930 (1998).
25. Tiana, G. and Broglia, R. A.: *J. Chem. Phys.* **114**, 2503 (2001).

26. Shakhnovich, E. I., Abkevich, V. A. and Ptitsyn, O.: *Nature* **379**, 96 (1996).
27. Abkevich, V. I., Gutin, A. M. and Shakhnovich, E. I.: *Biochemistry* **33**, 10026 (1994).
28. Fersht, A.: *Proc. Natl. Acad. Sci. USA* **92**, 10869 (1995).
29. Socci, N. D. and Onuchic, J. N.: *J. Chem. Phys.* **101**, 1519 (1994).
30. Li, H., Helling, R., Tang, C. and Wingreen, N.: *Science* **273**, 666 (1996).
31. Sali, A., Shakhnovich, E. and Karplus, M.: *Nature* **369**, 248 (1994).
32. Broglia, R. A., Tiana, G., Roman, H. E., Vigezzi, E. and Shakhnovich, E. I.: *Phys. Rev. Lett.* **82**, 4727 (1999).
33. Klimov, D. and Thirumalai, D.: *Phys. Rev. Lett.* **76**, 4070–4073 (1996).
34. Socci, N., Bialech, W. and Onuchic, J.: *Phys. Rev. E* **49**, 3440–4730 (1994).
35. Abkevich, V. I., Gutin, A. M. and Shakhnovich, E. I.: *J. Chem. Phys.* **101**, 6052 (1994).
36. Klimov, D. K. and Thirumalai, D.: *J. Mol. Biol.* **282**, 471 (1998).
37. Derrida, B.: *Phys. Rev. B* **24**, 2613 (1981).