

Folding and aggregation of designed proteins

R. A. BROGLIA*^{†‡}, G. TIANA[§], S. PASQUALI*, H. E. ROMAN*, AND E. VIGEZI*

*Dipartimento di Fisica Università di Milano and Istituto Nazionale di Fisica Nucleare, I-20133 Milan, Italy; [†]The Niels Bohr Institute, University of Copenhagen, 2100 Copenhagen, Denmark; and [§]Department of Physics, DTU Building 307, 2800 Lyngby, Denmark

Edited by Peter G. Wolynes, University of Illinois, Urbana, IL, and approved September 1, 1998 (received for review May 15, 1998)

ABSTRACT Protein aggregation is studied by following the simultaneous folding of two designed identical 20-letter amino acid chains within the framework of a lattice model and using Monte Carlo simulations. It is found that protein aggregation is determined by elementary structures (partially folded intermediates) controlled by local contacts among some of the most strongly interacting amino acids and formed at an early stage in the folding process.

Studies of how proteins fold have shown that the way protein clumps form in the test tube is similar to how proteins form the so-called “amyloid” deposits that are the pathological signal of a variety of diseases, among them the memory disorder Alzheimer’s (1–6). Protein aggregation traditionally has been connected to either unfolded or native states. Inclusion body formation has been assumed to arise from hydrophobic aggregation of the unfolded or denatured states, whereas the amyloid fibrils have been assumed to arise from native-like conformations in a process analogous to the polymerization of hemoglobin S.

By using lattice-model simulations (7–18), we find that aggregation arises from elementary structures that are controlled by local contacts, which eventually build the folding nucleus (16) of the heteropolymers where nonlocal contacts play an important role. Aggregation takes place when some of the most strongly interacting amino acids establish their local contacts, leading to the formation of a specific subset of the native structure. These elementary structures, which provide local guidance in both the folding and the aggregation process, can be viewed as the partially folded intermediates suggested to be involved in the aggregation of a number of proteins (6, 19–23).

We studied the simultaneous folding of two identical 20-letter amino acid chains, each composed of 36 monomers and designed to fold into their native conformation (Fig. 1*a*), within the framework of a simple lattice model of protein folding (14, 16, 18) and by using Monte Carlo (MC) simulations. Although the model does not treat side chains explicitly, the amino acids are chemically different. Their differences are manifested in pairwise interaction energies of different magnitude and sign, depending on the identity of the interacting amino acids. The configurational energy is

$$E = \frac{1}{2} \sum_{i,j}^N U_{m(i),m'(j)} \Delta(|\vec{r}_i - \vec{r}_j|), \quad [1]$$

$\{\vec{r}\}$ being the set of coordinates of all of the monomers describing a chain conformation. The quantity $\Delta(|\vec{r}_i - \vec{r}_j|)$ is a contact function. It is equal to one if sites i and j are at unit distance (lattice neighbors) not connected by a covalent bond, and zero otherwise. In addition, it is assumed that on-site repulsive forces prevent two amino acids from occupying the

same site simultaneously, so that $\Delta(0) = \infty$. There are 20 types of amino acids in the model. The quantities $U_{m(i),m'(j)}$ are the contact energies between amino acids of type m and m' and were taken from table 6 of ref. 24. The 36-mer chain denoted S_{36} and designed by minimizing, for fixed amino acid concentration, the energy of the native conformation with respect to the amino acids sequence is shown in Fig. 1*b*. At temperature $T = 0.20$ (in our temperature scale) it folds in 8×10^6 MC steps and at $T = 0.28$ the folding time is 8×10^5 MC steps. The fractional population of the native state corresponding to these two temperatures is 91% and 10%, respectively, to be compared with a population of 10^{-5} for the heteropolymer folding temperature of $T = 0.40$. All of the calculations presented below were carried out at the temperature of $T = 0.28$, optimal from the point of view of allowing for the accumulation of statistically representative samples of the different simulations, and at the same time leading to a consistent population of the native conformation. Pilot calculations carried out at $T = 0.24$ and $T = 0.26$, where the fractional population of the native state is 55% and 20%, respectively, leads to results that agree in detail with the results obtained at $T = 0.28$.

The characterization of the role played by the different amino acids in the folding process of S_{36} have been carried out in ref. 18, by using mutations (19 possible substitutions of monomers on each site). It was found that the 36 sites of the native conformation (see Fig. 1*a*) can be classified as “hot” (red beads, numbered 6, 27, and 30), “warm” (beads numbered 3, 5, 11, 14, 16, and 28), and “cold” (the rest of the beads) sites. On average, mutations on the 27 cold sites yield sequences that still fold to the native structure (neutral mutations), although the folding time is somewhat longer than for S_{36} . Sequences obtained from mutations on the six warm sites fold, as a rule, to a unique conformation, sometimes different but in any case very similar to the native one. Mutations on the three hot sites lead, in general, to complete misfolding (denaturation) of the protein.

Essentially two different outcomes of the simulation studies of the simultaneous folding of two S_{36} sequences have been observed: (i) both chains fold to their native conformation (Fig. 1*a*), and (ii) both chains get intertwined in conformations that are quite compact and display some amount of similarity to the native conformation (Fig. 2). In the first case, each chain targets into its minimum energy structure (native conformation) about which it fluctuates (25, 26). The second case is associated with an ensemble of compact low-energy conformations typical of those reached in the folding of a random chain, where the system spends little time in each conformation and displays conspicuous energy fluctuations.

At the basis of these phenomena are the elementary structures built out of the monomer sequences $S_4^1 \equiv (3,4,5,6)$, $S_4^2 \equiv (27,28,29,30)$, and $S_4^3 \equiv (11,12,13,14)$ (see Fig. 1*a*). They are controlled by the local contacts 3–6, 27–30, and 11–14, which are among some of the most strongly interacting amino acids. These structures aside from containing, at the local level,

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.

© 1998 by The National Academy of Sciences 0027-8424/98/9512930-4\$2.00/0
PNAS is available online at www.pnas.org.

This paper was submitted directly (Track II) to the *Proceedings* office. Abbreviations: MC, Monte Carlo; FPT, first passage time.

[†]To whom reprint requests should be addressed.

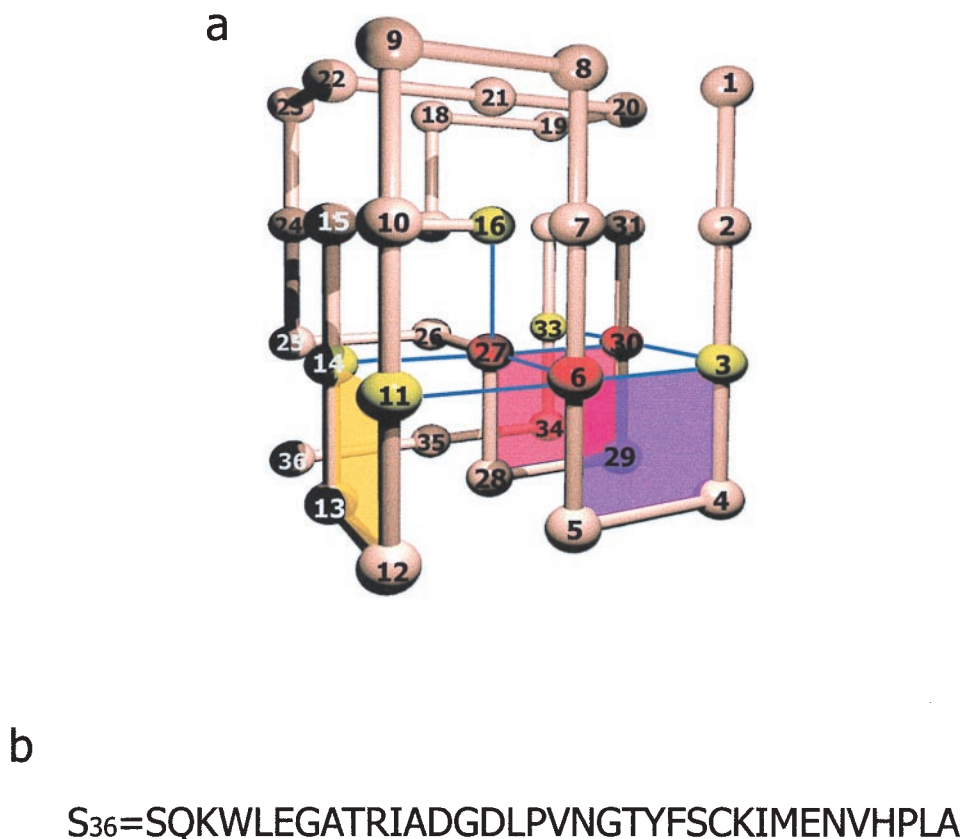


FIG. 1. (a) The conformation of the 36-mer chosen as the native state in the design procedure. Each amino acid residue is represented as a bead occupying a lattice site. The design tends to place the most strongly interacting amino acids in the interior of the protein where they can form most contacts. The strongest interactions are between groups D, E, and K (compare to *b*), the last one being buried deep in the protein (amino acid in site 27). The folding nucleus (16) (determined from the folding simulations) is formed by these amino acids (red beads) and by their nearest neighbors (yellow beads), and is shown by continuous blue lines. The structures formed by the amino acid sequences $S_4^1 \equiv (3,4,5,6)$, $S_4^2 \equiv (27,28,29,30)$, and $S_4^3 \equiv (11,12,13,14)$ are indicated by violet, red, and yellow, respectively. (b) Designed amino acid sequence S_{36} . The color plots in Figs. 1 and 2 were obtained by using the graphic program of ref. 31.

essentially all of the amino acids found in the folding nucleus (16) of the protein, provide the local guidance for its formation and thus are, to a large extent, responsible for the fast folding of the designed sequence S_{36} (Fig. 1*b*). In fact, the structures S_4^i ($i = 1,2,3$) can be viewed as the local “bricks” of a dynamical LEGO kit to model protein folding.

The pairs of strongly interacting monomers (3,6), (27,30), and (11,14) become nearest neighbors very early in the folding process, with the associated first passage time (FPT) being of the order of 10^2 MC steps in all three cases. The corresponding local contacts achieve 90–95% stability after 0.25×10^6 MC steps, a time to be compared with the FPT for the folding of both interacting chains (see *ii* above) and equal to 2×10^6 MC steps. The folding core is formed essentially when the three different bricks of the same chain assemble together, establishing the nonlocal contacts 6–27, 3–30, 6–11, and 27–14, at which time it becomes easy for nonlocal contacts 27–16 and 30–33 to fall in place. The speed with which this process is done crucially depends on the local contacts between monomers 3–6, 27–30, and 11–14, which control the stability of the local structures S_4^1 , S_4^2 , and S_4^3 .

Once the folding nucleus of both proteins is formed, it takes fewer than 3×10^4 MC steps for them to reach the native configuration. All of the contacts that maintain the bricks in place involve at least one amino acid occupying a hot site in the native conformation of the isolated protein (18), that is a strongly interacting amino acid (Fig. 1*a*). Once the hot site amino acids are in place, it takes 0.6×10^6 MC steps for both proteins to fold (FPT), in keeping with the fact that while the

FPT of the contact 6–27 is $\approx 0.4 \times 10^6$ MC steps, it takes $\approx 1.4 \times 10^6$ MC steps for it to become stable.

Aggregation results when one chain, in the process of establishing its nonlocal contacts, uses a hot site amino acid belonging to the other chain. The FPT associated with this phenomenon is typically 0.5×10^6 MC steps. In other words, aggregation occurs when the local structures (bricks) belonging to different chains attach to each other (see Fig. 2). Such a “mistake” can take place in a number of different ways, and not only in the ones that mimic the disposition of the bricks in the native core configuration, in keeping with the LEGO analogy. Because of the strongly interacting character of the amino acids occupying sites 27, 30, and 6, aggregation is, for all purposes, an irreversible process under native-like conditions, as seen by the results of simulations leading to aggregation that have been followed over 10^8 MC steps. We have repeated the calculations by using the contact energies $U_{m(i),m'(j)}$ (see Eq. 1) reported in table 5 of ref. 24 and obtained very similar results to the ones discussed above.

To see whether the local structures S_4^i ($i = 1,2,3$) are an artifact or not of the conspicuous dispersion displayed by the contact energies used in the calculations (24), we have repeated the simulations by using the Go model (8, 17). We found that the presence of the elementary structures S_4^i is, if anything, better defined in this case as compared to the case discussed previously, and that their role in the aggregation process is again essential. In particular, because all of these local structures now have equal energy content, in most of the events leading to aggregation, all three local structures of one

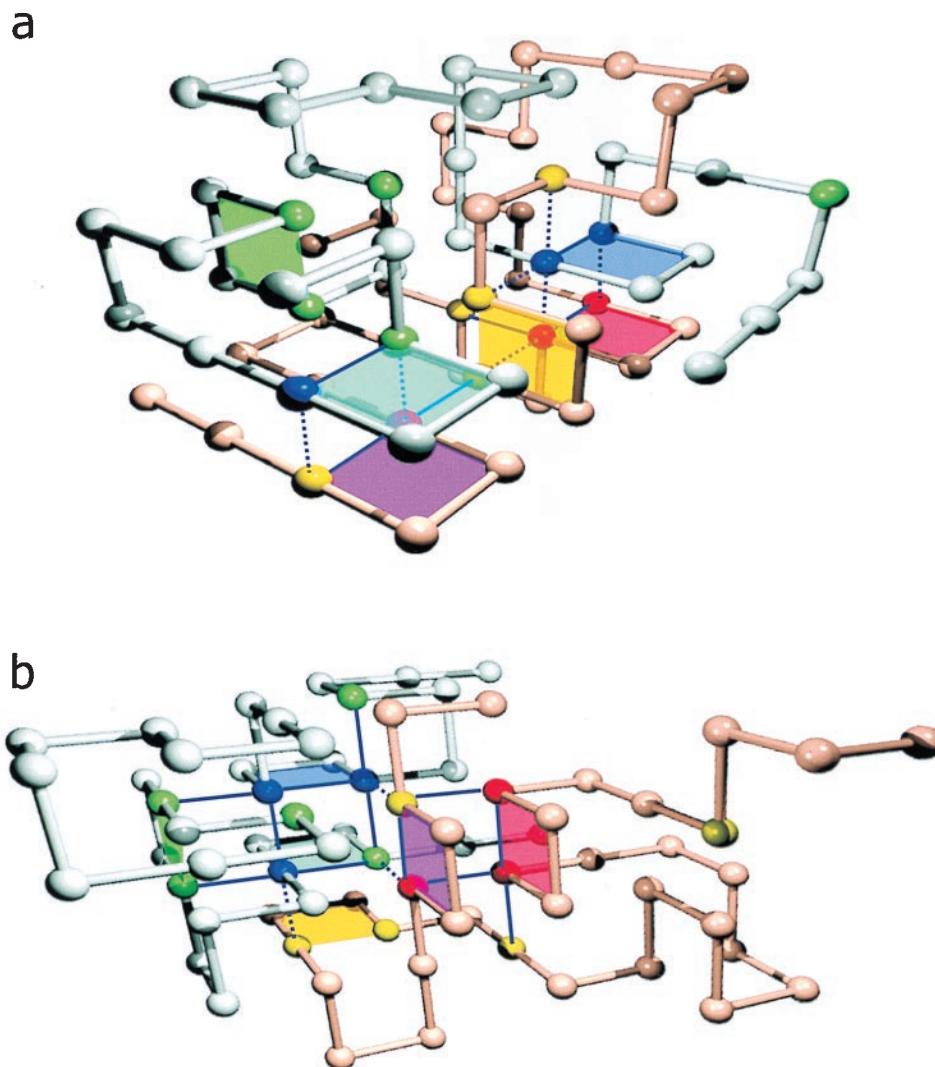


FIG. 2. Examples of aggregation. The hot sites of chain 1 and their nearest neighbors are shown as red and yellow beads, respectively, as in Fig. 1*a*. The hot sites of chain 2 are shown as blue beads, the corresponding nearest-neighbor amino acids in the native conformation by green beads. The three basic structures of chain 2 are shown in terms of light green, green, and light blue. The “correct” interactions of the folding nucleus of chain 1 are shown with a continuous line, the “wrong” ones by dashed lines. The presence of yellow beads close to the blue ones, and of green beads close to red ones, as well as of dashed lines connecting nearest neighbors, indicate that chain 1 has erroneously interpreted some of the elementary structures of chain 2 as belonging to itself and vice versa. (*a*) In this case, aggregation is controlled by the elementary structures made out of the sequences S_4^1 and S_4^2 . (*b*) Example of aggregation controlled by the elementary structures built out of sequences S_3^1 and S_3^2 .

chain find at least a local structure partner belonging to the other chain with which they interact.

By using again the contact energies of ref. 24, we have found that the rate of aggregation increases in a significant manner, by introducing cold (neutral) mutations. The chosen mutations are able to affect in a significant way the stability of one of the local structures, without much changing the ability the resulting isolated sequence S'_{36} has to fold on short call to the native conformation. In particular, substituting the amino acid R at position 11 of the designed sequence, by amino acid A, the rate with which aggregation takes place increases by 70% (i.e. from 22% to a 37% rate at a distance $d = 4$, where d represents the initial distance, in units of lattice spacing, between monomers number 18 of each of the chains). The reason for this increase is that it takes 0.6×10^6 MC steps for the pair of monomers 11–14 of the mutated sequence S'_{36} to establish a stable contact (as compared to 0.25×10^6 MC steps for S_{36}). Consequently, the other two local structures (associated with the monomer groups S_4^1 and S_4^2) have more time and thus a better chance to interact with the homologous structures of the other chain, than in the case of the simultaneous folding of two S_{36} sequences. Similar results have been obtained by perform-

ing single and multiple mutations in cold and warm sites of the native conformation. Because 75% of all sites are cold, and thus associated with neutral mutations (18), there is a large number of mutations that, while destabilizing the elementary structures and thus increasing the rate of aggregation, do not affect the stability of the protein in an important way. These results are consistent with a number of observations, in particular those carried out in the study of the amyloid-forming system transthyretin. When altered by any of 50 different mutations, this protein, which normally occurs in the blood plasma, deposits in the heart, lungs, and gut, causing a lethal disease called familial amyloidotic polyneuropathy (27, 28). These mutations do not alter normal folding of the protein but do destabilize the protein structure, facilitating the formation of partially folded intermediates that readily aggregate to one another (29, 30).

We conclude that a given protein will have a (small) number of local partially folded intermediates that control both protein folding and aggregation. Within the model of designed proteins these are the elementary structures that build the folding nucleus. Consequently, most of the aggregates of this protein, as well as of the sequences homologous to it, will display similar

native-like structures, independent of the nature of the effect triggering the aggregation.

Discussions with E. Shakhnovich are much appreciated. We thank the late Dr. N. D'Alessandro for help in modeling design. We gratefully acknowledge financial support by the North Atlantic Treaty Organization under Grant CRG 940231.

1. Fink, A. L. (1998) *Folding Design* **3**, R9–R23.
2. Silow, M. & Oliberg, M. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 6084–6086.
3. Mitraki, A. & King, J. (1989) *Biotechnology* **7**, 690–697.
4. Wetzel, R. (1998) *Cell* **86**, 699–702.
5. Janicke, R. (1995) *Philos. Trans. R. Soc. London B* **348**, 97–105.
6. Wetzel, R. (1994) *Trends Biotechnol.* **12**, 193–198.
7. Ueda, Y., Taketomi, H. & Go, N. (1975) *Int. J. Pept. Protein Res.* **7**, 445–449.
8. Go, N. & Abe, H. (1981) *Biopolymers* **20**, 1013–1031.
9. Lau, K. & Dill, K. (1989) *Macromolecules* **22**, 3986–3997.
10. Sklonik, J., Kolinski, A. & Sikorski, R. (1990) *Comm. Mol. Cell Biophys.* **6**, 223–247.
11. Covell, D. & Jernigan, R. (1990) *Biochemistry* **29**, 3287–3294.
12. Godzik, A., Kolinski, A. & Sklonik, J. (1994) *J. Comput. Chem.* **14**, 1194–1202.
13. Succi, N., Bialek, W. & Onuchic, J. (1994) *Phys. Rev. E* **49**, 3440–3443.
14. Shakhnovich, E. I. (1994) *Phys. Rev. Lett.* **72**, 3907–3910.
15. Klimov, D. & Thirumalai, D. (1996) *Phys. Rev. Lett.* **76**, 4070–4073.
16. Shakhnovich, E. I., Abkevich, V. & Ptitsyn, O. (1996) *Nature (London)* **379**, 96–98.
17. Pande, V. S., Grosberg, A. Y., Tanaka, T. & Rokhsar, D. S. (1998) *Curr. Opin. Struct. Biol.* **8**, 68–79.
18. Tiana, G., Broglia, R. A., Roman, H. E., Vigezzi, E. & Shakhnovich, E. I. (1998) *J. Chem. Phys.* **108**, 757–761.
19. King, J., Haase-Petingell, C., Robinson, A. S., Speed, M. & Mitraki, A. (1996) *FASEB J.* **10**, 57–66.
20. Speed, M. A., Morshead, T., Wang, D. I. O. & King, J. (1997) *Protein Sci.* **6**, 99–108.
21. Hurla, M. R., Helms, L. R., Li, L., Chan, W. & Wetzel, R. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 5446–5450.
22. Kim, D. & Yu, M. H. (1996) *Biochem. Biophys. Res. Commun.* **226**, 378–384.
23. Fink, A. L. (1995) *Annu. Rev. Biophys. Biomol. Struct.* **24**, 495–522.
24. Miyazawa, S. & Jernigan, R. L. (1985) *Macromolecules* **18**, 534–552.
25. Ptitsyn, O. B. (1992) in *Protein Folding*, ed. Creighton, T. E. (Freeman, New York), pp. 243–300.
26. Shakhnovich, E. I. & Finkelstein, A. V. (1989) *Biopolymers* **28**, 1667–1694.
27. McCutchen, S. L., Colon, W. & Kelley, J. W. (1993) *Biochemistry* **32**, 12119–12127.
28. McCutchen, S. L., Lai, Z., Mirov, G., Kelley, J. W. & Colon, W. (1995) *Biochemistry* **34**, 13527–13536.
29. Hamilton, J. A., Steinraut, L. K., Braden, B. C., Liepnieks, J., Benson, M. D., Holmgren, G., Sandgren, O. & Steen, L. (1993) *J. Biol. Chem.* **268**, 2425–2430.
30. Terry, C. J., Damas, A. M., Oliveira, P., Saraiva, M. J., Alves, I. L., Costa, P. P., Matias, P. M., Sakaki, Y. & Blake, C. C. F. (1993) *EMBO J.* **12**, 735–741.
31. Humphrey, W., Dalke, A. & Schulten, K. (1996) *J. Mol. Graphics* **14**, 33–38.