Hypothesis

PROTEIN FOLDING: GENERAL PHYSICAL MODEL

O. B. PTITSYN

Institute of Protein Research, USSR Academy of Sciences, 142292 Poustchino, Moscow Region, USSR

Received 12 June 1981

1. Introduction

The three-dimensional structure of a globular protein can be roughly described as a definite mode of packing of 'structural segments' (α-helices or β-strands) [1–3]. Prediction of the localization of these segments in a linear protein chain is now more or less possible [3–5]. However, the problem of self-assembly of these segments into the tertiary protein structure is impeded by the absence of visible criteria which are used by these segments to recognize each other. In fact, β-strands are connected by hydrogen bonds which do not depend on the amino acid sequences of these strands. α-Helices are connected with each other and with β-strands by hydrophobic interactions which depend not on the details of their amino acid sequences but on the total hydrophobic surface area screened from water [6].

It has been suggested that the mutual recognition of structural segments takes place not by the method of 'trial and error' but is a result of a directed process of protein folding proceeding through definite, most favourable intermediate states [7]. The possible concrete types of these directed processes have been discussed by us [5,8–12]. In this paper, which is a development of the references mentioned, I shall propose a general hypothesis on the directed mechanism of protein folding. According to this hypothesis, the mutual recognition of structural segments is determined not by the details of their amino acid sequences but simply by their mutual localization in a linear polypeptide chain.

2. Initiation of folding

Let us consider the simplest possible model of a polypeptide chain – a homopolymer of identical amino acids with hydrophobic side groups. This chain can possess in a water environment a secondary structure stabilized by hydrogen bonds between its peptide groups and a tertiary structure stabilized by hydrophobic interactions of its side groups. Let us start from the basic assumption that the folding of this chain into a three-dimensional structure begins with the formation of the most stable complex from two chain regions. (This means that the equilibrium between all these complexes is reached before the joining of other chain regions to these complexes and before the rearrangement of these complexes into a compact structure.)

The most stable complexes from two regions of a homogeneous chain can be either an antiparallel β-hairpin (fig.1) or a complex of two antiparallel α-helices (fig.2). The choice between these two possibilities is determined by the values of β-hairpin-coil (s_β) and α-helical hairpin-coil (s_α) equilibrium constants for this chain. In both cases the most stable complex will evidently include practically the whole chain bending in half near its middle point (fig.1,2). The chain regions equidistant from the middle will be near in space in this complex and will be connected by hydrogen bonds or by hydrophobic interactions. If the complex–coil equilibrium constant $s > 1$, the complex will be more stable than the coil, i.e. all chain regions equidistant from its middle point will 'recognize' each other already at the first stage of the folding. If the equilibrium constant $s \simeq 1$, the complex will fluctuate and the maximum probability of constants between chain regions will be shifted to its bend as the probability of the formation of a loop from n residues is inversely proportional to $n^{3/2}$. (Calculations show that this probability in typical cases will be maximal at distances about $N/4 - N/5$ from the middle of the chain where $2N$ is the total number of residues.)

Published by Elsevier/North-Holland Biomedical Press
00145793/81/0000–0000/$02.50

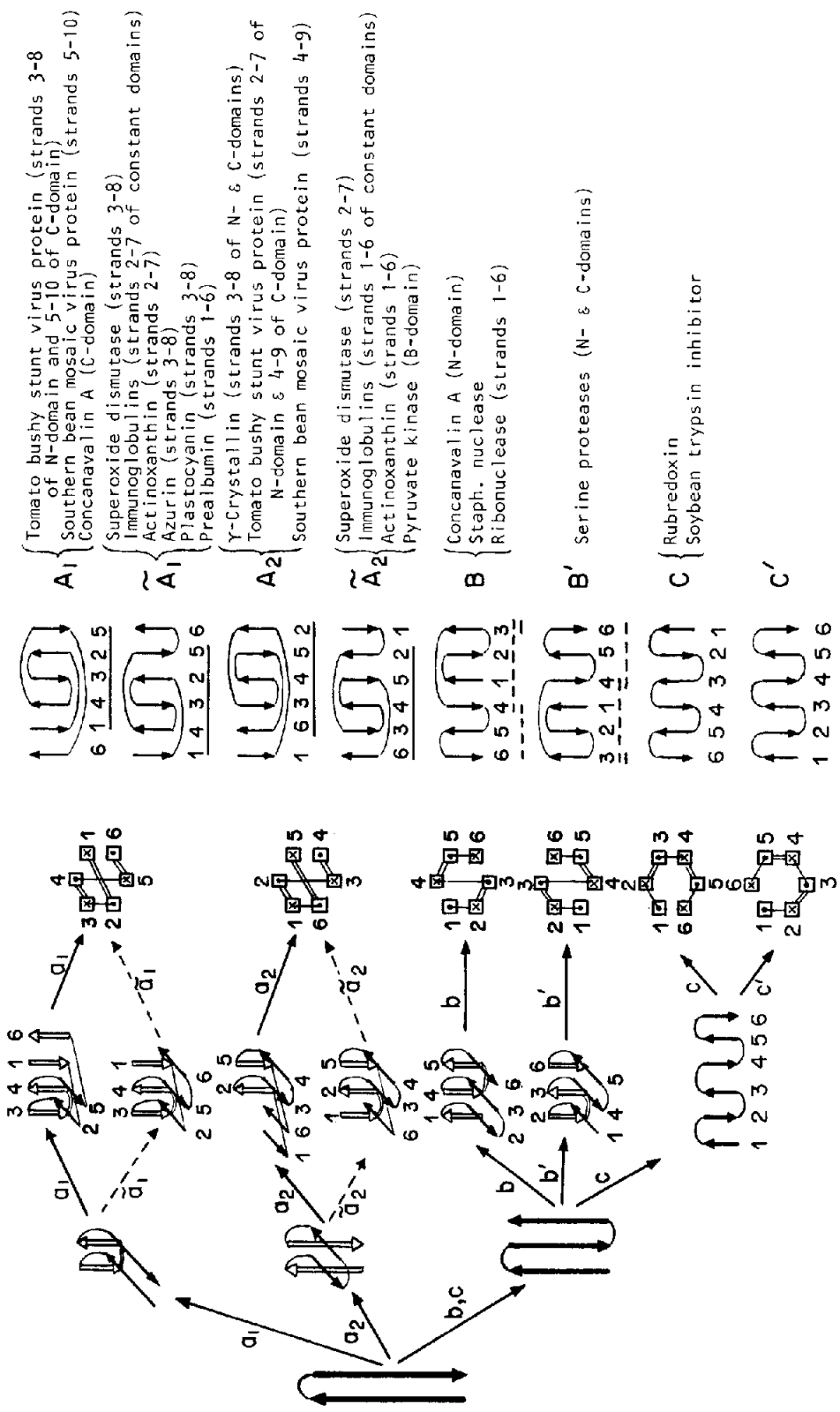

Fig. 1. The pathways of rearrangement of the initiating β-hairpin into a 6-stranded β-cylinder. Dotted lines show pathways a_1' and a_2' (see the text) which are possible only for fluctuating hairpins. The right twist of the β-structure is shown for double β-sheets. The closed cylinders are shown from the top (points and crosses mean β-strands directed up and down). The right part of the figure shows the topologies of the obtained structures and the proteins or domains which include these topologies. The β-strands involved into a double and into a simple Greek key are underlined by solid and dotted lines.

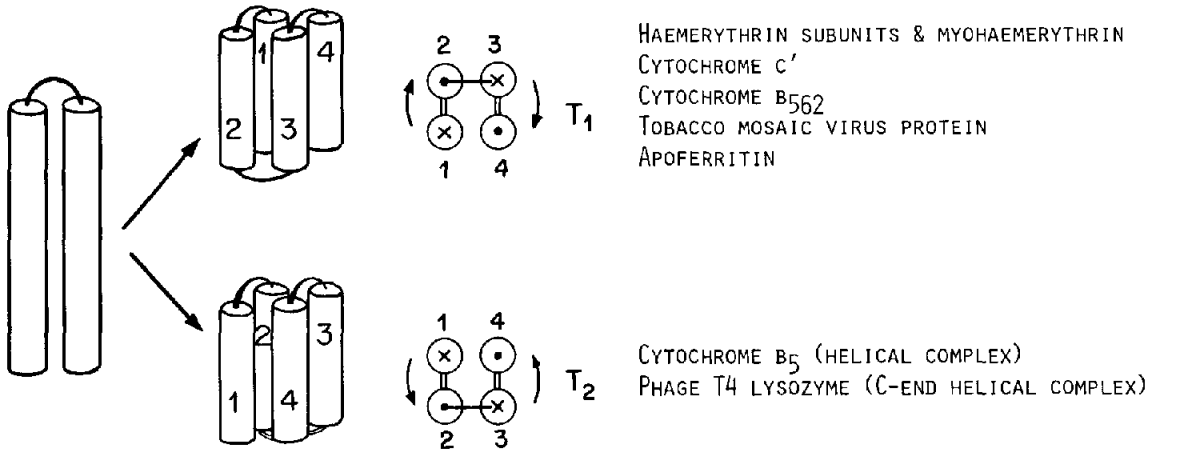

Fig.2. The pathways of rearrangement of the initiating α-helical hairpin into a tetrahelical complex. In the right part of the figure the top views of the tetrahelical complexes are presented (points and crosses mean α-helices directed up and down). Proteins which included these tetrahelical complexes are shown.

It follows that in both cases ($s > 1$ and $s \simeq 1$) the regions of a chemically homogeneous chain equidistant from its middle point will recognize each other. The only difference is that if $s > 1$ (the stable complex) practically all regions equidistant from the middle of the chain are stably connected with each other already at the stage of the initiating complex formation and if $s \simeq 1$ (the fluctuating complex) only the chain regions which are near its middle point are stably connected and other regions are fluctuating between the connected and the non-connected states.

3. Folding of β-structural chains

Let us consider the case when $s_\beta > s_\alpha$, i.e., the most stable initiating complex is an initiating β-hairpin. A half of hydrogen bonds of residues in this hairpin remains unsaturated and all hydrophobic groups remain unscreened from water. Therefore this hairpin must be rearranged into a compact structure which corresponds to the maximal saturation of hydrogen bonds and the maximal screening of hydrophobic groups. An antiparallel β-sheet rolled into a closed barrel is such a type of structure as in this barrel almost all hydrogen bonds are saturated and a half of hydrophobic groups is screened from water. The optimal number of β-strands in this β-sheet is equal to 6 because if this number is smaller than 6, a β-sheet cannot roll into a barrel and the increase of the number of β-strands above 6 leads to an additional loss of free energy at the formation of 'excessive' bends.

The initiating hairpin embracing the whole chain can be transformed into a β-barrel by 1 of the 3 following pathways:

(i) By its breaking with the formation of a double β-sheet (pathway *a* in fig.1);
(ii) By its rearrangement into a β-sheet from 3 β-strands with its subsequent breaking (pathway *b*);
(iii) By its rearrangement into a β-sheet from 6 β-strands (pathway *c*).

In all cases the double (pathways *a* and *b*) or single (pathway *c*) β-sheet is rolled afterwards into a closed cylinder. The 'memory' of the mutual recognition of chain regions in an initiating β-hairpin is conserved only on the first pathway while on the second and on the third pathways the recognition of chain regions will be determined by their mutual arrangement not in a hairpin but in a β-sheet into which this hairpin has been transformed. However, it is easy to see that just the first pathway is the most probable because double β-sheets formed as a result of the breaking of an initiating β-hairpin present more stable intermediate structures than single β-sheets formed as a result of a rearrangement of the β-hairpin. (In fact, in a single β-sheet the loss of free energy of residues in a β-hairpin and in a β-sheet are close to one another [12].) Moreover, the final structures formed on the pathway *a* are more favourable than the structures formed on the pathways *b* and especially *c* as in the first structures the hydrophobic interior of a cylinder

is better screened from water.

Various pathways of the type a differ only by the direction of the first breaking. This direction determines which direction of the second breaking (i.e., place of adjoining of the end chain regions) corresponds to the maximal screening of the hydrophobic interior of a double sheet [3,10]. In a stable hairpin ($s_\beta > 1$) the end chain regions remain connected even after the second breaking (pathways a_1 and a_2) while in a fluctuating hairpin they can also join independently (pathways $\tilde{a}'_1$ and $\tilde{a}'_2$). As a result, 4 structures are formed (A_1, $\tilde{A}'_1$, A_2 and $\tilde{A}'_2$) which coincide with each other in pairs for a closed cylinder ($A_1 = \tilde{A}'_1$, $A_2 = \tilde{A}'_2$). All these 4 structures have the topology of the 'double Greek key' [10] (chain regions arrangement 14325 or 63425) with the definite direction of the swirl of this key (counterclockwise if viewed from the external side of the cylinder).

A less probable pathway b of the rearrangement of a stable β-hairpin leads to two structures B and B'. Each of these structures has 2 overlapping topologies of the 'simple Greek key' [10] (chain region arrangements 4123 and 3654) and these structures differ only by the sign of the key swirl (clockwise or counterclockwise). Finally, the smallest probability has the pathway c which leads to 2 structures C and C' with the simplest 'up-and-down' topology of the meander ornament. Fluctuations of the end regions of β-sheets on pathways b and a can lead also to structures A_1 and A_2 as well as to some other structures (not shown).

4. Folding of α-helical chains

The case when $s_\alpha > s_\beta$, i.e., the most stable initiating complex is an α-helical hairpin (fig.2) can be treated in an analogous way. Almost all hydrogen bonds are saturated in this hairpin; however, only 1 hydrophobic side group/1 turn of the α-helical region is screened from water. Therefore this hairpin must be rearranged into a compact structure. It is possible to show that for chains with the dimensions usual for protein domains (~100–200 residues) the most stable compact structure is a complex of 4 α-helices. Only 2 of these complexes can be obtained from the initiating hairpin (fig.2); they differ only by the sign of their swirl (clockwise or counterclockwise as viewed from the N-end of the first helical region).

5. Comparison with the experiment for glubular proteins

A chemically homogeneous chain is, of course, a very crude model for the protein chain. Nevertheless, let us try to apply the proposed scheme to protein chains or to their parts which, being inhomogeneous by their chemical structure, are homogeneous by the preferable type of their secondary structure, i.e., to β-structural or α-helical proteins or domains. Of course, we must have in mind that the number of β-strands or α-helices in these proteins is determined mainly by the number of clusters of hydrophobic residues in their amino acid sequences [3,10] and therefore can be greater than 6 β-strands in β-proteins or not equal to 4 in α-proteins.

The majority of β-structural proteins or chains have the structures of closed β-cylinders or double β-sheets [1–3]. Thirteen of these proteins or domains have one of the topologies of the 'double Greek key' predicted by our scheme or even 2 such overlapping topologies (see fig.1). Three other domains (N- and C-domains of serine proteases and the N-domain of concanavalin A) have the structure of closed β-cylinders or a double β-sheet with the topologies of the 'simple Greek key', B and B', while rubredoxin and soybean trypsin inhibitor have the structures of closed cylinders with the 'up-and-down' topology C (fig.1). Moreover, the β-structural parts of α + β-proteins, Staphylococcal nuclease and pancreatic ribonuclease, have the structures of a double β-sheet and a V-shaped β-sheet with the topologies B (fig.1). (References to the X-ray papers are given in [2,3], see also [13,14] for the recently obtained structures of two proteins.) I do not know any purely β-structural protein or domain which would have a structure of a closed β-cylinder or a double β-sheet with some other topology not predicted by this scheme.

The tetrahelical complex predicted by our scheme is also the most popular structure of α-helical proteins or domains [15,16]. The structure T_1 (with the clockwise swirl) is present in at least 5 proteins with different primary structures and different functions [15,16] whereas the tetrahelical complex of cytochrome b_5 and the phage lysozyme have the structure T_2 (with the counterclockwise swirl). Finally, Ca^{2+}-binding domains of muscle proteins have the structure of the tetrahelical complex formed by approximately antiparallel α-helices 1 and 4 (as well as 2 and 3) with an approximately perpendicular arrangement of these

two pairs of α-helices. (References to the X-ray papers are given in [1,15], see also [16] for the recently obtained structure of one protein.)

6. Conclusions

The scheme of protein folding proposed in this paper is based on the simple principle of mututal recognition of the chain regions equidistant from the middle of the chain. This scheme does not at all take into account uniqueness of the amino acid sequences of the real proteins and therefore it can be formulated even for the simplest case of a chemically homogeneous chain. Nevertheless, the structures predicted by this scheme happen to be typical for the real β-structural and α-helical proteins with completely different amino acid sequences.

Moreover the structures predicted by this scheme can be distinguished from all possible structures of chemically homogeneous chains not only from the point of view of the folding but also from the point of view of the stabilities of these structures. Among 3840 possible topologies of the closed β-cylinder from 6 β-strands only 120 topologies satisfy the principle of the antiparallelity of structural segments neighbouring along the chain (see [3]) and only 8 from these 120 topologies (the topologies A_1, A_2, B, C and their mirror images) form purely antiparallel β-sheets with non-crossing connections. Among these topologies, topologies A_1 and A_2 are more favourable than B and especially than C as in topologies A_1 and A_2 the hydrophobic interior of the β-cylinder is screened from water from both sides, in topology B only from one side and in topology C it remains unscreened (see fig.1). In an analogous way from 48 possible topologies of tetrahelical complexes only two topologies (T_1 and T_2) satisfy all these criteria.

Thus, if we shall assume that connections between structural segments cannot cross either the surface of protein globule or each other and that a purely antiparallel arrangement of both β-strands and α-helices are more favourable than the mixture of their antiparallel and parallel arrangement we shall obtain that the above-mentioned topologies are more stable than all other possible topologies of a chemically homogeneous chain. We have shown above that the same topologies can be obtained as the result of the folding pathway of a chemically homogeneous chain, most natural from the physical point of view.

It permits us to suggest that the typical tertiary structures of at least α-helical and β-structural proteins, like their secondary structures, are determined mainly by the general properties of the polypeptide chain rather than by the unique amino acid sequences of the proteins selected in the course of biological evolution.

Amino acid protein sequence determines the choice between several possible types of tertiary structures (e.g., between structures A, B and C for β-proteins). It determines also the details of these structures (the exact number of structural segments, their lengths, the detail of their packing as well as conformations of loops). In some cases specific amino acid sequence can give rise to other structures which are unfavourable for homogeneous chains. And, last but not least, the unique amino acid sequence is very important for the formation of the active center of the protein. However, the set of typical tertiary structures of the globular proteins very probably is determined by general properties of polypeptide chains irrespective of their amino acid sequences.

Acknowledgements

The hypothesis presented in this paper is the result of numerous fruitful discussions with Dr A. V. Finkelstein to whom the author is very much indebted. The author also thanks Dr Yu. N. Chirgadze for reading the manuscript and O. A. Rakitina for improving the English.

References

[1] Levitt, H. M. and Chothia, C. (1976) Nature 261, 552–558.
[2] Richardson, J. S. (1977) Nature 268, 495–500.
[3] Ptitsyn, O. B. and Finkelstein, A. V. (1981) Quart. Rev. Biophys. 13, 339–386.
[4] Anfinsen, C. B. and Scheraga, H. A. (1975) Adv. Prot. Chem. 29, 205–300.
[5] Chou, P. Y. and Fasman, G. D. (1977) Trends Biochem. Sci. 2, 128–131.
[6] Clothia, C. (1974) Nature 248, 338–339.
[7] Ptitsyn, O. B. and Rashin, A. A. (1975) Biophys. Chem. 3, 1–20.
[8] Ptitsyn, O. B. and Finkelstein, A. V. (1978) Bioorgan. Khim. 4, 349–353.
[9] Ptitsyn, O. B. and Finkelstein, A. V. (1979) in: Protein: Structure, Function and Industrial Application, Proc. 12th FEBS Meet., Dresden, 1978, vol. 52, pp. 105–115, Pergamon, Oxford.

[10] Ptitsyn, O. B., Finkelstein, A. V. and Falk (Bendzko), P. (1979) FEBS Lett. 101, 1–5.
[11] Ptitsyn, O. B. and Finkelstein, A. V. (1980) in: Biomolecular Structure, Conformation and Evolution (Srinivasan, R. ed) vol. 1, pp. 119–132, Pergamon, Oxford.
[12] Ptitsyn, O. B. and Finkelstein, A. V. (1980) in: Protein Folding (Jaenicke, R. ed) pp. 101–115, Elsevier/North-Holland, Amsterdam, New York.
[13] Abad-Zapatero, C., Absel-Megnid, S. S., Johnson, J. E., Leslie, A. G. W., Rayment, I., Rossman, M. G., Such, D. and Tsukihara, T. (1980) Nature 286, 33–39.
[14] Pletnev, V. Z., Kuzin, A. P. and Trakhanov, S. D. (1980) Bioorgan. Khim. 6, 1420–1422.
[15] Weber, P. C. and Salemme, F. R. (1980) Nature 287, 82–84.
[16] Clegg, G. A., Stansfield, R. F. D., Bourne, P. E. and Harrison, P. M. (1980) Nature 288, 298–300.