THE FOLDING OF PROTEIN CHAINS. PREDICTION OF TOBACCO MOSAIC VIRUS PROTEIN TERTIARY STRUCTURE

V. I. LIM and A. V. EFIMOV

Institute of Protein Research, Academy of Sciences of the USSR, 142292 Pushchino, Moscow Region, USSR

Received 10 July 1976

1. Introduction

A mechanism of self-organization of a protein chain into a native globule was recently suggested by one of us [1,2]. The main assumption of this mechanism is that the interaction of fluctuating α-helices is the physical principle which dictates a three-dimensional distribution of the polypeptide chain material in the protein globule. According to this mechanism, the formation of the tertiary structure of protein passes through three stages. Firstly local interactions form fluctuating α-helices on the greater part of the chain length. At the second stage long-range interactions fold the protein chain into a highly helical intermediate globule (HIG) in which the three-dimensional distribution of the chain material is close to the native distribution. Finally, the HIG is transformed into a native structure by unwinding of the excess part of α-helices.

Using this mechanism we predicted the tertiary structure of the protein of strain *vulgare* tobacco mosaic virus. The structure was searched in two stages: (1) the formation of the HIG; (2) the transformation of the HIG into the native structure by unwinding the excess α-helices and the retention of the α-helices and β-structures which are predicted by the stereochemical theory of the secondary structure of globular proteins [3,4].

2. Method and results

Consider an α-helix in which a greater part of hydrophobic side groups form a hydrophobic cluster extending from one terminus of the helix to the other. We shall call this α-helix and s-helix (see fig.1a). Proline is allowed to be in any position of the s-helix, if this will lead to an elongation of the hydrophobic cluster. An analysis on CPK models showed that the presence of proline in the internal turns of the α-helix does not result in a sharp distortion of its geometry. One border of the hydrophobic cluster will be called the right border (RB) and the other the left border (LB) (see fig.1a). It follows from the determination of the s-helix that neighbouring hydrophobic side groups along the RB and LB must be located in positions i, i ± 3 or i, i ± 4, where i is the amino acid residue number along the chain beginning from the N-terminus. We shall designate the pair of hydrophobic residues located in positions i, i ± 3 and in positions i, i ± 4 by the numerals 4 and 5, respectively. Then the geometry of the RB and LB of the s-helix can be simply described by a definite combination of these numerals. For example, $5_1-4_2-5_3-5_4$ (the indices designate the ordinal number of the pair of residues) means that the border of a hydrophobic cluster is formed from five hydrophobic residues located in positions i, i + 4, i + 7, i + 11, i + 15. Let us draw together the right borders of two s-helices (not necessarily identical) so that the angle between vectors $\vec{NC}$ will be obtuse and the side groups located along the right borders of both s-helices will come into contact (hydrophobic side groups with the hydrophobic and hydrophilic side groups with the hydrophilic) (see fig.1b). The dimer obtained in this way from the two s-helices will be called an F-structure. The stereochemical analysis leads to the conclusion that in an aqueous medium the F-structure is more advantageous energetically than any other di-s-helical structure [2]. In the case of interaction of several different s-helices, the F-structures with similar right

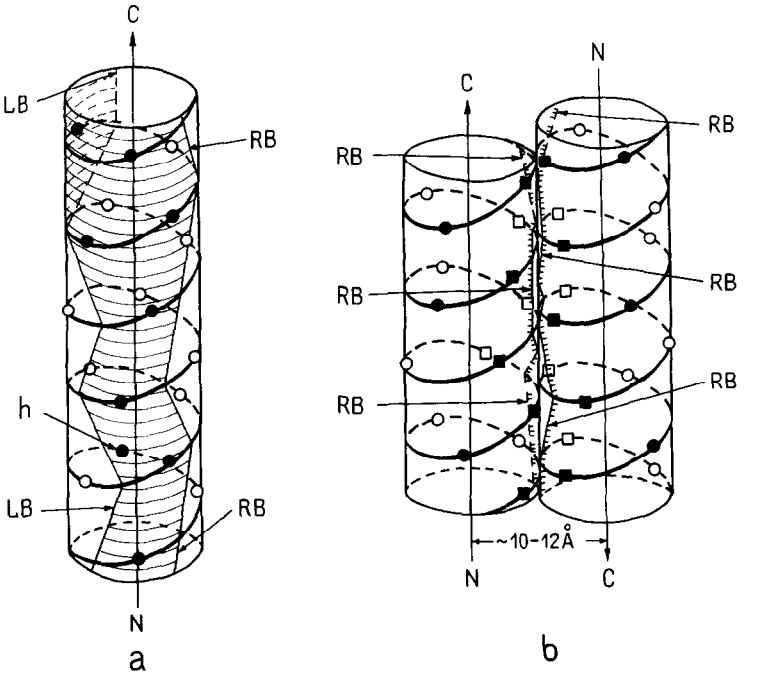

Fig.1. Schematic representation of the s-helix and the F-structure. (a) The s-helix. Solid circles denote hydrophobic residues, open circles – hydrophilic residues; hatched and unhatched regions are hydrophobic and hydrophilic clusters of the s-helix, respectively. h is a hydrophobic residue outside the hydrophobic cluster. RB and LB are the right and left borders of the hydrophobic cluster. The vector $\vec{NC}$ lies on the axis of the s-helix and is directed from the N- to the C-terminus of the s-helix. (b) The F-structure with an angle between the vectors $\vec{NC}$ close to 180°. Solid circles and squares denote hydrophobic residues, open circles and squares – hydrophilic residues. Squares show hydrophobic and hydrophilic residues which are located along the right borders of both s-helices and form interhelical hydrophobic and hydrophilic contacts.

borders or parts of the right borders will be the optimal di-s-helical structures. For example, there are four s-helices with the following right borders: 5_1–4_2–5_3–5_4, 4_1–5_2–4_3–5_4, 4_1–4_2–5_3, 5_1–4_2–4_3–5_4. In this case the two F-structures will be the optimal di-s-helical structures. One of them is the F-structure with the approached regions 5_1–4_2–5_3 and 5_2–4_3–5_4 and the other is the F-structure with the approached regions 4_1–4_2–5_3 and 4_2–4_3–5_4. It should be noted that the right border or its part described by the numeral 4 three or more times in succession or by the numeral 5 four or more times in succession cannot come into contact along the whole length in the F-structure. This is because such a fragment of the RB will be wound to 160° and more.

The formation of the HIG of tobacco mosaic virus protein was made by uniting F-structures. For the primary structure of strain *vulgare* tobacco mosaic virus protein see [5]. The polypeptide chain of tobacco mosaic virus protein has four regions (10–35, 40–60, 72–87, 110–133) which form s-helices if they are wound into an α-helical conformation (see fig.2). There may be three variants of forming a highly helical intermediate globule from these four s-helices:

(a) I + II + III + IV → (I·II) + (III·IV) → [(I·II)·(III·IV)];
(b) I + II + III + IV → (I·III) + (II·IV) → [(I·III)·(II·IV)];
(c) I + II + III + IV → (I·IV) + (II·III) → [(I·IV)·(II·III)].

Parentheses with numbers of s-helices designate the F-structure formed by corresponding s-helices. Square brackets mean the dimer formed by two F-structures. Topologically, variant (b) is impossible without a layering of hydrophilic fragments of interhelical regions on the hydrophobic or hydrophilic surface of F-structures. Such a layering will create steric hindrance

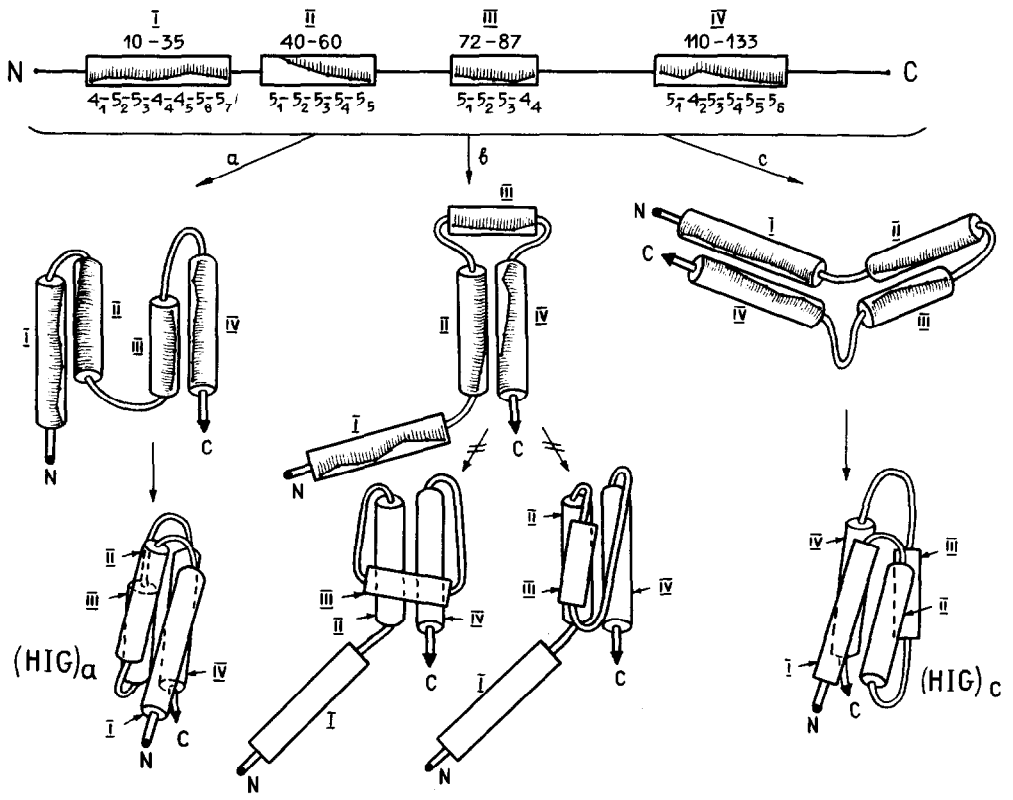

Fig.2. Variants a, b, and c of self-organization of a highly helical intermediate globule (HIG) of tobacco mosaic virus protein. ( ) – schematic representation of the geometry of the right borders of s-helices. The hatching along the RB is directed towards the hydrophobic cluster. In the case of variant b the crossed arrows (⊣⊢) denote that passing into the highly helical intermediate globule is impossible.

to the approach of hydrophobic surfaces of F-structures or will lead to a shielding of polar side groups from water molecules (see fig.2). Thus, our attention is attracted to only two other variants of selforganization of the highly helical intermediate globule, i.e. variants (a) and (c). The stereochemical analysis of the right border of every s-helix (for methods of analysis see in [3,4]) showed that in variant (a) regions 5_6–5_7, 5_2–5_3 and 5_1–5_2–5_3, 5_4–5_5–5_6 of the right border must be drawn together in the F-structure (I·II) and the F-structure (III·IV), respectively. In variant (c) regions 4_5–5_6–5_7 and 4_2–5_3–5_4 in the F-structure (I·IV) and regions 5_1–5_2–5_3 of both s-helices in the F-structure (II·III) must be drawn together (see fig.2). If we impose conditions of maximal adhesion of hydrophobic surfaces of F-structures and the absence of a layering of interhelical regions on the hydrophilic surface of F-structures, we shall obtain two intermediate highly helical globules shown in fig.2.

The $(HIG)_a$ and $(HIG)_c$ represent structures which are 'mirror-symmetrical' in the packing of s-helices, in the localization of the ends of the polypeptide chain and interhelical regions. The F-structures of the $(HIG)_c$ have more inter-s-helical contacts than the F-structures of the $(HIG)_a$. This difference in contacts is caused by regions 4_5 and 4_2 of the right border in the F-structure (I·IV). The region of the adhesion of hydrophobic surfaces of the F-structures (I·IV) and (II·III) in the $(HIG)_c$ is greater than that of the F-structures (I·II) and (III·IV) in the $(HIG)_a$. This circumstance signifies that the number of hydrophobic side groups localized on the surface of the $(HIG)_a$ is greater than that on the surface of the $(HIG)_c$. All

this leads to the conclusion that the $(HIG)_c$ is more favourable than the $(HIG)_a$ in intramolecular interactions. At the same time the $(HIG)_c$ is inferior to the $(HIG)_a$ in intermolecular interactions (due to the larger number of hydrophobic groups on the $(HIG)_a$ surface). The fact that tobacco mosaic virus protein forms a quaternary structure did not allow us to completely eliminate variant (a), i.e. the possibility of the formation of the $(HIG)_a$. The $(HIG)_c$ is in good agreement with the structure of TMV protein suggested recently by Champness et al. [6] from X-ray data (5 Å resolution).

As the final structure of tobacco mosaic virus protein in both variants, (a) and (c), we propose structures which are obtained if the HIG is modified in the following way. Let us unwind the HIG helices not predicted by the theory of secondary structure of globular proteins and introduce the β-structures predicted by it [3,4]. For predictions of the secondary structure see [4]. The introduction of the predicted secondary structure (helices: 22–39, 44–53, 79–90, 121–134; β-structure: 9–13, 67–72, 92–96, 150–152) into the HIG does not lead to a change in its general architecture.

3. Discussion

The results give evidence of a fruitful approach to the prediction of the protein tertiary structure using the HIG formed from s-helices and F-structures. Apparently the formation of a protein hydrophobic core by adhesion of F-structures and s-helices should be considered as a key moment in the self-organization of a protein molecule. As is shown in the example of TMV protein, this process determines the main parameters of the globule such as its dimensions, shape and the distribution of the polypeptide chain material within the globule. The approach developed by us can be especially successful for predicting the tertiary structure of highly helical globular proteins. Subsequent publications in this direction will follow.

References

[1] Lim, V. I. (1975) Dokl. Akad. Nauk SSSR 222, 1467–1469.
[2] Lim, V. I. (1976) J. Mol. Biol., submitted for publication.
[3] Lim, V. I. (1974) J. Mol. Biol. 88, 857–872.
[4] Lim, V. I. (1974) J. Mol. Biol. 88, 873–894.
[5] Dayhoff, M. O. (1972) Atlas of Protein Sequence and Structure, Vol. 5, D-285, National Biomedical Research Foundation, Silver Spring.
[6] Champness, J. N., Bloomer, A. C., Bricogne, G., Butler, P. J. C. and Klug, A. (1976) Nature 259, 20–24.