Tritium planigraphy: From the accessible surface to the spatial structure of a protein

(Received 27 January 1998; accepted 26 March 1998)

ELENA N. BOGACHEVA*, VITALII I. GOL’DANSKII†, ALEXANDER V. SHISHKOV*, ALEXANDER V. GALKIN‡, AND LUDMILA A. BARATOVA§

*Semenov Institute of Chemical Physics, Russian Academy of Sciences, Moscow, 117977; †Institute of Agricultural Biotechnology, Russian Academy of Agricultural Sciences, Moscow, 127550; and §Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, 119899, Russia

Contributed by Vitalii Gol’danskii, January 16, 1998

ABSTRACT The method of tritium planigraphy, which provides comprehensive information on the accessible surface of macromolecules, allows an attempt at reconstructing the three-dimensional structure of a protein from the experimental data on residue accessibility for labeling. The semiempirical algorithm proposed for globular proteins involves (i) predicting theoretically the secondary structure elements (SSEs), (ii) experimentally determining the residue-accessibility profile by bombarding the whole protein with a beam of hot tritium atoms, (iii) generating the residue-accessibility profiles for isolated SSEs by computer simulation, (iv) locating the contacts between SSEs by collating the experimental and simulated accessibility profiles, and (v) assembling the SSEs into a compact model via these contact regions in accordance with certain rules. For sperm whale myoglobin, carp and pike parvalbumins, the λ crh repressor, and hen egg lysozyme, this algorithm yields the most realistic models when SSEs are assembled sequentially from the amino to the carboxyl end of the protein chain.

Studies on the protein spatial structure use a panoply of physical, chemical, and biological methods, the foremost of which are certainly x-ray analysis and high-resolution NMR. Notwithstanding their merit, these methods are quite laborious and have a number of inherent limitations as applied to macromolecules. This makes topical a search for alternative means of obtaining structural information, including theoretical approaches (1–4).

There are ways to predict the tertiary structure of globular proteins by determining the hypothetically possible sites of contact between secondary structure elements (SSEs) and then arranging the latter into a three-dimensional (3D) complex that should reflect the spatial fold of the macromolecule (5–8). Their weak point is the multiplicity of the admissible models they produce. Thus even for a fairly simple protein, sperm whale myoglobin, there may be several hundred structures. Even if certain restrictions are imposed (6) (their number can be cut to 20), the choice of a single version remains largely arbitrary. Building a realistic 3D model would be greatly facilitated if data were available on the actually existing contacts. Such information can be obtained experimentally by tritium planigraphy. Our early work with sperm whale myoglobin, carp and pike parvalbumins, the λ crh repressor, and hen egg lysozyme, this algorithm yields the most realistic models when SSEs are assembled sequentially from the amino to the carboxyl end of the protein chain.

After labeling the protein, it is split into relatively short fragments so that each fragment has the least number of amino acid repeats; the cleavage pattern and means are chosen individually in accordance with the protein primary structure. After resolving the mixture of peptides, each is subjected to acid hydrolysis and amino acid analysis with simultaneous determination of radioactivity. As a result, we obtain the distribution of specific radioactivity along the protein sequence, i.e., the accessibility profile.

Residue Accessibility in Isolated Secondary Structure Elements and the Contact Regions

In the general case, the probability of residue labeling depends on the geometry and chemical properties of its side group, as well as on its shielding by other units close to it in the protein globule. Evaluation of the shielding is especially important for modeling the spatial structure, because this permits one to locate the regions of contact between separate parts of the polypeptide chain. The contribution of shielding can be esti-
mated by comparing the residue accessibility in the protein globule and in isolated elements of secondary structure. The former can be determined directly in the experiment (see above), whereas the latter has to be deduced by computer simulation of tritium bombardment.

Our simulation algorithm, which makes use of the Monte Carlo procedure, has been described in detail (19). Isolated SSEs were built with standard parameters: angles $\phi$, $\psi$, $\chi$, and bond lengths. Residue accessibility was calculated as $P_i = n_i/N_i$, where $n_i$ is the number of tritium atoms that reached an atom of the $i$th residue, and $N_i$ is the total of tritium atoms with trajectories passing through the residue volume. In the experiment, $P_i$ corresponds to the ratio of tritium atoms that labeled residue $i$ to their total flux incident on the target. The sites of potential contact were located by the decrease in residue accessibility in the whole molecule versus the isolated SSEs, taking 80% shielding as the threshold for assigning residues to the contact region.

3D Modeling of Globular Proteins

The approach outlined above has been used to model the 3D structure of four $\alpha$-helical proteins [sperm whale myoglobin (19), the $\lambda$ cro repressor, carp and pike parvalbumins (9, 20)] and one $\alpha/\beta$-protein [hen egg lysozyme (21)]. For all these proteins except pike parvalbumin III, x-ray data are available, which makes them convenient objects for testing and refining the technique. In the search for contacting SSEs, we took into account the length of the connector loop and its stereochemistry, following for the latter the general rules set by Efimov (22).

On the basis of the experience thus accumulated and with the wish to add practicality to the planigraphic approach, herein we focus on the most general patterns and the specific features that we encountered in modeling and formulate the ensuing rules that should be followed to build a realistic model.

Initial Data. The initial data for designing a 3D model are the accessibility profiles obtained experimentally for the whole protein and by computer simulation for isolated SSEs. The protein secondary structures were predicted from their amino acid sequences, by using the Finkelstein’s algorithm, analogous methods, or combinations thereof (23–25). The $\alpha$-helices were modeled as cylinders 5 Å in diameter with $C_{\alpha}$ atoms on their surface, and $\beta$-structures as 2.5-Å-thick flexuous slabs twisted clockwise (in the amino to carboxyl direction) by some $5^\circ$ per residue. The spatial model was constructed by combining the SSEs with an account of the contact regions located as above.

Fig. 1 exemplifies the accessibility profiles for the entire
molecule and isolated α-helices of the λ cro repressor, and the “spreadsheets” of the helices indicating the contact zones (at least 80% shielding). The choice of contacts between α-helices is governed by the length of loops connecting them, as distinguished below.

**Short Interhelix Loops.** Helices separated by one to three residues obviously cannot come in contact with each other, and the contact zones located on their surface must originate from association with other helices farther along the chain. Thus there is no contact between the first (amino proximal) three α-helices A, B, and C in myoglobin, which form a complex schematically shown in Fig. 2a. The same is observed in the λ cro repressor: helices A, B, and C with connectors of one and two residues (Fig. 2b). In myoglobin there is also no contact between helices D and E that have a two-residue connector. The first contact in this protein becomes possible between helices B and D, and the shape of the contact zone dictates their nearly perpendicular positioning (Fig. 3a). In the λ cro repressor, the first contacts are formed by helix D with the middle of helix B and the amino-proximal part of helix C (Fig. 3b).

**Long Interhelix Loops.** Longer connectors (four residues and more) allow a broader range of contacts between SSEs, as in parvalbumins and lysozyme. It turned out that the order of helix packing is basically important for obtaining a correct structure of the macromolecule. For this, one should distinguish below.

**Long Interhelix Loops.** Longer connectors (four residues and more) allow a broader range of contacts between SSEs, as in parvalbumins and lysozyme. It turned out that the order of helix packing is basically important for obtaining a correct structure of the macromolecule. For this, one should distinguish below.

**Prosthetic Groups.** Many proteins carry various “nonprotein” functional groups covalently attached to the polypeptide chain. This, as a rule, substantially influences the spatial structure of the macromolecule. For this, one should distinguish two cases: when such a group is attached in the course of protein synthesis and when such a group is attached after completion of the polypeptide chain. We encountered the former case with sperm whale myoglobin. The loop between helices E and F has eight residues and is long enough for packing F antiparallel to E and parallel to helix A, and their surfaces have appropriate vacant contact zones. On the other hand, it is known that helix F with helix B and partly with helix C form the “pocket” for the heme. If we assume an amino to carboxyl packing, the final complex of α-helices ABCDEFGH with the above F/E/A arrangement would then have to undergo profound changes to accommodate the heme. A correct configuration of the complex can be obtained only if helix F is taken integral with the heme. Indeed, heme incorporation into hemoglobin has been shown to take place cotranslationally (26). As the chemical bond with the heme is formed concurrently with helix F, it is not just admissible but necessary to consider their complex as a unity in modeling.

**β-Domains.** Compared with α-helices, β-strands are more labile fluctuating structural elements. In this respect, modeling of β-proteins as well as those containing both α-helices and β-domains is a special task. Let us consider this case with an α/β-protein, lysozyme (21) as an example. After the bihelical complex AB, the SSE next in sequence is a β-structure composed of three antiparallel β-strands: residues 42–46 (β1), residues 50–54 (β2), and residues 57–60 (β3). In principle, there are three ways of further modeling: (i) consecutive packing of only α-helices, whereby the entire region of residues 36–79 is deemed unordered, giving rise to a β-structure only upon completion of the helical complex; (ii) one-by-one attachment of β1, β2, and β3 to AB; or (iii) assembly of the three strands into a β-structure, which is then attached as a unity to the bihelical complex. These ways lead to dissimilar spatial models.

---

**Fig. 2.** Spatial model of the complex of nonintersecting α-helices connected with short loops (one to three residues). (a) Sperm whale myoglobin. (b) λ cro repressor.

**Fig. 3.** Complexes of α-helices in sperm whale myoglobin (a) and λ cro repressor (b) with the first interhelix contacts: helices B–D in a and helices C–D and B–D in b.

**Fig. 4.** Hexahelical complexes obtained with different modes of helix packing for carp parvalbumin. (a) Sequential amino to carboxyl packing (A + B + C + D + E + F). (b) Sequential carboxyl to amino packing (F + E + D + C + B + A). (c) Around the core pair of helices C and D most shielded in the macromolecule [(C + D) + E + F + B + A]. (d) Same as c but adding the AB complex as a unity [(C + D) + E + F + AB]. Only model a is consistent with the x-ray data.
According to the x-ray data, in lysozyme there are no contacts of helices A and B with helix C, which clearly rules out pathway i. Further, helix A has no contact with the β-strands at all, whereas helix B has closest contacts with β3 and β2 but only marginal contact with β1, which is just opposite to what would be produced via pathway ii; the latter also yields a set of interstrand contacts contradicting the actually existing one. Hence modeling must include a separate step of building the β-domain (Fig. 5a). When added to the AB complex, helix C, the next helix, comes in contact with the β-domain but not with helix A or B. The resulting model of the β-domain and the entire AB-β-CD “sandwich” shown in Fig. 5 agree nicely with the known x-ray structure of lysozyme. Such a situation, in which β-domains, which can be regarded as suprasubstructure elements, should be given “equal rights” with α-helices, may prove fairly common for multidomain proteins.

Disulfide Bridges. A prominent role in sustaining the native spatial structure is played by disulfide bridges between cysteines, which are often far away from each other in the protein sequence. A typical example is again lysozyme, whose four disulfide bridges (between residues 6–127, 30–115, 64–80, and 76–94) not only have the key part in forming the tertiary structure but also ensure the high conformational stability of the protein globule. Four of the eight cysteines are in α-helices and, by the extent of shielding, could have been classed with contact zones. It is noteworthy that pronounced shielding is found for the sulfur atoms of all lysozyme cysteines, and in all probability, this is what makes the protein quite tolerant of reducing agents.

If the tertiary structure is formed in the amino to carboxyl direction, one could have expected that disulfide bridges were formed between cysteines in neighboring SSEs, i.e., residues 30–60, 64–76, 80–94, and 115–127. However, as maximal matching of contact zones is sought, these residues prove to be located on the opposite sides of the complexes, and disulfide bridges are formed between residues remote in the sequence but close in the finally folded spatial model. Much the same perhaps takes place in reality.

Conclusions and Implications

Thus, our experience in modeling the protein spatial structure using the data of tritium planigraphy, albeit limited, allows some generalizations. The algorithm proposed herein yields correct results if the following rules are observed: (i) secondary structure elements are packed sequentially from the amino to the carboxyl end of the polypeptide chain; (ii) cofactors and other groups covalently attached to the protein cotranslationally are regarded as being integral with the corresponding SSE; (iii) in α/β-proteins, the β-domains and the α-helices are taken as integral units for assembly; (iv) disulfide bridges are formed at the final stage, within the established spatial structure composed of α-helices and β-structures, and “fasten” the cysteines remote in the sequence but close in the protein globule.

The fact that this modeling algorithm reflects the direction of polypeptide chain synthesis on the ribosome corroborates the idea that the protein spatial structure arises cotranslationally. From the physical standpoint, this appears quite plausible. Indeed, ribosomal synthesis of a protein of 150–300 residues takes 30–60 sec. This time is much shorter than that needed for protein renaturation (dozens of minutes or even hours) but exceeds by several orders of magnitude the characteristic times of secondary structure formation and intramolecular motions, including segmental movements. With the assertion that the spatial structure begins to be formed only after completion of synthesis, one has to further assume special factors that for a fairly long time prevent the folding of the polypeptide chain and protect it from various proteolytic agents. A number of recent works provide direct experimental evidence that the growing polypeptide on the ribosome acquires a certain spatial structure resembling that of the mature macromolecule (26–30); moreover, for firefly luciferase, the nascent protein exhibits enzymic activity even before its release from the ribosome (31). This brief discourse is only to underline the consistency of our modeling with the natural process and does not at all exclude the participation of other factors, such as molecular chaperones, in protein folding. Regardless of the particular sequence of events giving rise to the spatial structure of macromolecules, we now have grounds for saying that tritium planigraphy offers an independent experimental means for reconstructing the macromolecular organization—working from the surface to the interior—that may prove to be a valuable addition to the established approaches in this field of research.

We thank A. S. Spirin for his interest in this line of research and fruitful discussions. This study was partly supported by grants from the International Science Foundation (MML000) and the Russian Foundation for Basic Research (96–04-30692 and 96–03-34185).