Biophysical Chemistry, 31 (1988) 139–142
Elsevier

BPC 01293

The formation of protein secondary structure

Its connection with amino acid sequence

Piotr Zielenkiewicz, Danuta Płochocka and Andrzej Rabczenko

Institute of Biochemistry and Biophysics, Polish Academy of Sciences, 36 Rakowiecka, 02532 Warsaw, Poland

Received 5 September 1987
Accepted 1 February 1988

Protein folding; Tertiary structure; Sequence analysis

Statistical analysis of the occurrence of tetrapeptides in 35 globular proteins was performed. It was found that the amino acids along the polypeptide chain are close to being randomly distributed and that the same tetrapeptide segments exist in different types of secondary structure. Therefore, a new method was proposed for locating 'microdomains' in protein interiors. Amino acid replacements in the hydrophobic core of six proteins were analyzed. The results show that the locations of amino acids belonging to defined microdomains are extremely conserved. It is suggested that the structures found may play a role as nucleation centers in protein folding.

It is well known that the native, three-dimensional structure of proteins is, under physiological conditions, determined exclusively by the amino acid sequence [1]. It has been shown that protein folding cannot occur in a 'trial-and-error' manner [2], but rather via a unique folding pathway.

In most predictive methods, it has been assumed that the necessary intermediates of the folding pathway are protein secondary structures and that these are determined by the local sequence. Strong arguments against such thinking have recently been presented. The most spectacular among these are:

(1) The accuracy of the prediction of secondary structure by existing methods was found [3] to be less than 55% in a three-state assessment (α, β, coil) and less than 45% in a four-state assessment (α, β, turn, coil).

(2) The same five-residue sequence was found to exist in different secondary structures [4].

Correspondence address: P. Zielenkiewicz, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, 36 Rakowiecka, 02532 Warsaw, Poland.

(3) Indirect evidence shows that the polypeptide chain conformation is dependent on environmental factors. One can recall here the well-known example of the S-peptide of ribonuclease S (residues 1–20) which forms an α-helix in the presence of S-protein; however, in the absence of S-protein, this structure is lost [5].

Here, we should like to put another two additional bricks into this wall!

If the assertion that the local amino acid sequence of the polypeptide chain fragment determines its secondary structure is correct, it is reasonable to hypothesize that, conversely, the local amino acid distribution along the chain differs according to the types of structure. In other words, the set of sequences forming an α-helix should be different from those forming, say, a β-structure.

To test this hypothesis, a statistical analysis was performed for 35 different globular proteins of known structure, taken from the Protein Data Bank [6] (the list of proteins and particular details of the method including the derivation of mathematical expressions are presented elsewhere

[7]). Sequences of α-helical, β-structural and unstructured fragments were selected according to the Protein Data Bank.

Sets of tetrapeptide segments were constructed by commencing at the first amino acid of a given fragment and then moving along the chain by one amino acid per step for all fragments. This procedure gave 4704 tetrapeptide segments – 1301 α-helical, 686 β-structural and 2717 unstructured.

The number of times a given tetrapeptide sequence is repeated in each of the selected sets of sequences was investigated. The sum of repetitions of different tetrapeptides in a set was called the number of repetitions in that set.

The numbers of repetitions obtained were compared with results from a theoretical formula describing the expected number of repetitions, derived under the assumption that the polypeptide sequence is completely random:

$$s(r) = \sum_{i=1}^{r} \left(1 - \prod_{j=1}^{4} p(l_j)_i\right)^r - n + r$$

where $s(r)$ is the expected number of repetitions, r the number of tetrapeptides in the set, $n = 20^4$,

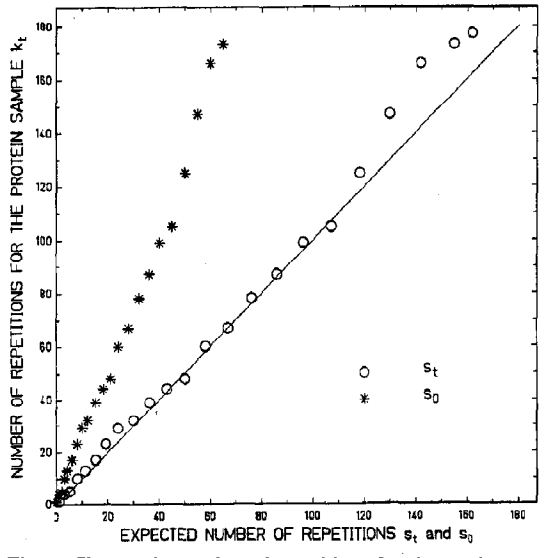

Fig. 1. Changes in number of repetitions for the total set vs. theoretical values assuming that the frequencies of occurrence of elements is equal to those of amino acids in the total set (○) and equal to 1/20 (△).

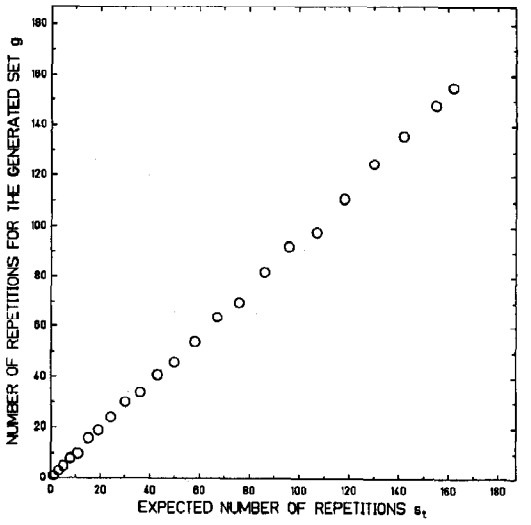

Fig. 2. Changes in number of repetitions for the generated set vs. values obtained by theoretical prediction.

$p(l_j)$, $j = 1, \ldots, 20$, the relative frequencies of occurrence of amino acids in the sample analyzed. The results are presented in fig. 1. Additionally, a random chain of letters was generated, taking into account the frequency of occurrence of each letter, and the number of repetitions was generated for the random chain. The numbers of repetitions for the sets generated are in agreement with values obtained by theoretical prediction (fig. 2).

It was found that the numbers of repetitions of tetrapeptides, obtained for α-helical, β-structural, unstructured and total sets (data shown for total set only), are in good agreement with theoretical predictions based exclusively on the knowledge of the frequencies of occurrence of individual elements in the sets considered. This means that, from the point of view of statistics, the arrangement of amino acids along the polypeptide chain is close to a random distribution. Also, we have found 74 tetrapeptide sequences (listed in table 1) occurring in at least two different structures.

We conclude that knowledge of the sequence of a short polypeptide segment is not sufficient to determine the conformation of that sequence in the protein. In other words, formation of secondary structure is the consequence of formation of the three-dimensional structure of the whole