Amino acid distribution rules predict protein fold

Alexander E. Kister and Vladimir Potapov†

†Department of Mathematics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, U.S.A., and †Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, U.S.A.

Abstract

In the present article, we provide a brief overview of the main approaches to analysing the sequence-structure relationship of proteins and outline a novel method of structure prediction. The proposed method involves finding a set of rules that describes a correlation between the distribution of residues in a sequence and the essential structural characteristics of a protein structure. The residue distribution rules specify the ‘favourable’ residues that are required in certain positions of a polypeptide chain in order for it to assume a particular protein fold, and the ‘unfavourable’ residues incompatible with the given fold. Identification of amino acid distribution rules derives from examination of inter-residue contacts. We describe residue distribution rules for a large group of β-sandwich-like proteins characterized by a specific arrangement of strands in their two β-sheets. It was shown that this method has very high accuracy (approximately 85%). The advantage of the residue rule approach is that it makes possible prediction of protein folding even in polypeptide chains that have very low global sequence similarities, as low as 18%. Another potential benefit is that a better understanding of which residues play essential roles in a given protein fold may facilitate rational protein engineering design.

Computational approaches to structure prediction: strengths and weaknesses

Current understanding of how amino acid sequences dictate a protein’s three-dimensional structure is based on Anfinsen’s thermodynamic hypothesis [1], which states that protein folding depends solely on the physical and chemical properties of amino acids and the sequence of residues in a polypeptide chain. This discovery moves the biological phenomenon of protein folding out of the living cells and turns it into a problem analysable with the help of computer ‘cells’.

Currently, the two main computational approaches used to predict protein structure are the ab initio and the homology-based methods. The ab initio protein structure modelling methods are based on analysis of inter-atomic contacts. Ideally, these calculations yield the structure of the protein. The advantage of the ab initio approach is that it is based solely on the physicochemical properties of residues, requiring no prior knowledge of solved structures. The disadvantage is the necessity for energy calculations, to high degrees of accuracy, of the many different strong and weak interactions such as electrostatic contacts, hydrogen bonds, van der Waals interactions and hydrophobic interactions [2–4]. Special difficulties are presented by calculation of water-mediated hydrophobic interactions, a dominant component of the folding process [3,6].

The homology approach is based on statistical analysis of known protein structures. The advantage of this approach is the possibility of using well-developed methods of sequence alignment and machine learning to assign the protein in question to a three-dimensional structure [7–10]. The disadvantage is the impossibility of predicting protein structure if there is no similar protein sequence template whose protein structure has already been described.

Residue distribution rules determine protein fold

In the present article, we describe an alternative approach to structure prediction that seeks to circumvent problems of protein structure modelling posed by conventional computational methods. It is based on the rules of distribution of residues in sequences that are responsible for the SSS (supersecondary structure) of proteins. The first step in determining the rules of sequence-structure correlation is to define the structural features essential to a specific protein fold. For sandwich-like protein structures, the essential requirement is the rigorously defined ‘interlock’ between two pairs of neighbouring β-strands [11]. This invariant substructure forms a hydrophobic core in nearly all known sandwich proteins.

The next step in the analysis is the delineation of the sequence characteristics necessary for the formation of the structural ‘invariant’ of a particular fold. For example, specific arrangement of strands in a β-sheet is governed by certain rules of distribution of residues within these strands that form inter-strand contacts. Analysis of residue-residue contact maps of different proteins that have the same fold allows one to understand which pairs of residues form the most essential contacts for a given structure [12–14]. The result...
of this analysis articulates the set of rules that dictate the distribution of amino acids involved in the critical contacts responsible for a particular fold. In the following sections, the ‘rules of residue distribution’ approach to sequence–structure prediction are developed for sandwich-like proteins.

Supersecondary structures of \( \beta \)-sandwich-like proteins

Spatial structures of sandwich-like proteins are composed of \( \beta \)-strands, which form \( \beta \)-sheets that pack face-to-face. The number of strands and their arrangement varies widely [15]. Detailed structural classification of these proteins is presented in two protein structure databases, SCOP [16] and CATH [17]. The ‘spatial structure’ referred to is the SSS of proteins. This structural category, hierarchically intermediate between secondary and tertiary structures, was introduced by Rao and Rossman [18]. There are several reasons for using SSS rather than the atomic structures of proteins in studies of sequence–structure correlations. First, the SSS definition is strict and unambiguous: proteins with the same SSS have an identical number of strands and the same mutual arrangement of these secondary-structural elements in a protein domain. Secondly, proteins with an identical SSS may belong to different families, have diverse protein functions and hence, possibly, very little global sequence similarity. Sequences with similar SSS, but widely dissimilar primary sequences, are the most informative for discovering residue distribution rules. Such sequences permit one to detect peculiarities within amino acid sequences that are responsible for a particular fold, rather than overall similarities among sequences.

One more advantage of SSS is that it simplifies protein modelling analysis. Use of ‘protein skeletons’ (SSS), rather than whole atomic three-dimensional protein structures, allows one to uncover the common structural features specific and unique to a given fold. For sandwich-like SSS, the unifying structural feature is the two pairs of strands in two \( \beta \)-sheets with a specific ‘interlock’ arrangement [11] (Figure 1). In the next section, we focus on proteins with a particular interlock arrangement consisting of a pair of strands (strands 1 and 4) in one \( \beta \)-sheet and a pair of strands (strands 2 and 5) in the other \( \beta \)-sheet. Residues of these strands form a hydrophobic core of their respective structures, and their interactions play a crucial role in protein stability. Thus knowledge of residue distribution rules in the interlock-forming strands is essential for understanding sandwich formations.

Residue distribution rules for sandwich-like proteins

The rules are divided into two groups. Rules for the first group specify favourable and unfavourable residues in strands and loops. These rules mostly stem from typical features of strands and loops in \( \beta \)-sheet structures. Residue variability in strands and loops was investigated in different ways: analysis of amino acid composition of immunoglobulins [19], examination of structural determinants in sequences of immunoglobulin domains [20], and analysis of sequence patterns and residue frequencies in \( \beta \)-strand and loop structures [21–25]. Application of ‘group 1’ rules defines secondary structures to a first approximation, i.e. it yields the approximate localization of strands and loops in a sequence.

The second group of rules determines whether a given arrangement of strands is consistent with a sandwich SSS. This determination is based on analysis of residue distribution in their strands and loops and in mapping out the possible residue–residue contacts necessary for the formation of two specific \( \beta \)-sheets in sandwich-like proteins. To define all potential contacts between strands, we considered all pairs of ‘facing’ residues in neighbouring strands in a \( \beta \)-sheet. For example, in the SSS shown in Figure 1, an interaction between strands 1 and 4 is achieved through contacts between residues at positions 1, 2, 3, 4 and 5 in strand 1 and residues at positions 5, 4, 3, 2 and 1 respectively in strand 4.

When deciding whether two residues may be in contact with each other, three types of possible contacts were considered: (i) hydrophobic contacts (both residues are hydrophobic); (ii) aromatic contact (both residues are aromatic); and (iii) polar contact (in a pair of residues, one is negatively and the other is positively charged). If a pair of residues could form any one of the three types of contacts, then they are considered to be in contact.

The rules of the second group dictate which pairs of residues need to be in contact with each other in order to form a sandwich structure, and the minimal number of such contacts. This threshold of necessary contacts is determined on the basis of the analysis of contact maps of proteins.
Examples of residue distribution rules for sandwich proteins

Group 1 rules govern the residue content of strands and loops. Examples of rules that proscribe amino acids in certain positions within strands are as follows.

(i) Rules regarding proline and glycine residues prohibit these residues: (a) at any hydrophobic position in the strands; (b) at two or more positions in a strand; and (c) at three positions in total within all six strands of both β-sheets.

(ii) Rules regarding alanine residues prohibit: (a) three or more alanine residues in one strand; (b) a total of two or more alanine residues at hydrophobic positions in interlock strands 1, 2, 4 and 5; and (c) two or more alanine residues in combination with either proline or glycine in the same strand.

(iii) Rules regarding aspartic acid, asparagine and glutamine residues prohibit: (a) three or more of these residues in one strand; and (b) two or more of these residues in combination with proline or glycine in one strand.

(iv) Aspartic acid, glutamic acid, lysine, arginine and histidine residues prohibit three or more of these polar residues at hydrophobic positions in the six strands, in total.

Examples of rules that specify the ‘favourable residues’ in strands are as follows.

(i) Rules regarding aspartic acid, glutamic acid, lysine, arginine, histidine, glutamine, threonine, tyrosine, serine and glycine residues require that: (a) at least one of the three hydrophilic positions must be occupied by these hydrophilic residues in each strand; (b) that there be a total of no less than eight hydrophilic residues at hydrophilic positions in the four interlock strands 1, 2, 4 and 5; and (c) three or more of these hydrophilic residues be located in ‘edge’ strands of the β-sheets (strands 2, 3 and 6).

(ii) Rules regarding isoleucine, valine, leucine and phenylalanine residues require that at least one hydrophobic position be occupied by one of these residues in each interlock strand 1, 2, 4 and 5.

Rules that govern residue distribution of ‘favourable’ residues in loops are not as strictly defined as the rules for strands. These rules are mostly based on the statistical analysis of frequency of amino acids in loops. Some examples of such rules are as follows.

(i) Rules regarding proline or glycine residues require that: (a) one or more of these two residues be present in at least three or more loops; and (b) at least one of these two residues should be present in every short loop with fewer than five residues.

(ii) Rules requiring that there be more hydrophilic residues (lysine, arginine, histidine, aspartic acid, glutamic acid, asparagine, glutamine, threonine, serine, tyrosine, cysteine, proline and glycine) than hydrophobic residues (isoleucine, valine, leucine, phenylalanine, methionine, alanine and tryptophan) in every loop.

Group 2 rules concern inter-residue contacts. The possibility of sandwich SSS formation is determined in accordance with the following rules: minimum numbers of hydrophobic and/or aromatic and/or polar contacts for pairs of strands in sheets are (i) five or more contacts between strands 1 and 4; (ii) three or more contacts between strands 2 and 5; (iii) one or more contacts for pair of strands 3 and 4, and pair of strands 5 and 6; and (iv) the total number of contacts for β-sheets comprising strands 1, 4 and 3 and 5 and 2 is 14.

Rules for contacts between two β-sheets are based on the concept of the hydrophobic core of proteins. Numerous observations of protein structures have confirmed the assumption that the hydrophobic effect is essential for protein stability [26–28].

Conclusion

The present review demonstrates that, alongside with traditional methods of sequence and structural alignments, there exists a novel approach to protein structural prediction based on formulation of rules that capture sequence and structural features essential to a particular protein fold. Rules of residue distribution for this set of proteins can identify sandwich-like SSS with 85% accuracy. Moreover, nearly all ‘false positive’ proteins turned up by the rules were β-proteins with correctly determined localization of strands in their sequences.

An important advantage of the ‘rules of residue distribution’ approach is the possibility of predicting the SSS of non-similar sequences. For example, sequences of two protein domains, PDB codes 1PFC (chain B, residues 124–230, of immunoglobulin heavy chain γ constant domain) and 2CRY (chain A, residues 117–220, of Kin of IRRE-like protein 3), have low degrees of homology. The global pairwise alignment of these two sequences using the Needleman–Wünsch algorithm revealed 18% similarity. Commonality of their SSS despite a low overall degree of homology implies the existence of ‘hidden’ similarity that could not be discovered by traditional sequence alignment methods. This hidden similarity can be described formally by the rules of residue distribution for sandwich-like proteins.
Another benefit of the proposed approach is the possibility of determining the structural role of residues. For example, one of the rules of residue distribution requires the presence of hydrophobic residues at three hydrophobic positions in strands 1 and 2. If this rule is relaxed to allow one of the three positions to be occupied by any non-hydrophobic residue, then 36% of proteins which satisfied the new residue distribution rules were false positives. There are other tests that confirmed that residue distribution in strands 1 and 2, which are decisive for sandwich-like SSS formation. In contrast, the rules which describe strands 3 and 6, which are not involved in the conserved part of the structure, i.e. ‘the interlock’, are not as important. Even if the rules for strands 3 and 4 are disregarded, the accuracy of detection of sandwich-like proteins is not significantly decreased. Understanding the relative importance of the structural role of various residues in a sequence may prove very useful for rational protein engineering design.

References