COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

Volume No: 7, Issue: 9, May 2013, e201305001, http://dx.doi.org/10.5936/csbj.201305001

A method to predict edge strands in beta-sheets from protein sequences

Antonin Guilloux [a], Bernard Caudron [b], Jean-Luc Jestin [c,*]

Abstract: There is a need for rules allowing three-dimensional structure information to be derived from protein sequences. In this work, consideration of an elementary protein folding step allows protein sub-sequences which optimize folding to be derived for any given protein sequence. Classical mechanics applied to this system and the energy conservation law during the elementary folding step yields an equation whose solutions are taken over the field of rational numbers. This formalism is applied to beta-sheets containing two edge strands and at least two central strands. The number of protein sub-sequences optimized for folding per amino acid in beta-strands is shown in particular to predict edge strands from protein sequences. Topological information on beta-strands and loops connecting them is derived for protein sequences with a prediction accuracy of 75%. The statistical significance of the finding is given. Applications in protein structure prediction are envisioned such as for the quality assessment of protein structure models.

Research Article

Introduction

Rules relating protein sequence and its three-dimensional structure are of special interest for protein structure prediction. Protein structures are mainly composed of beta-strands arranged in sheets, of helices and of loops and turns connecting them [1-3].

Beta-strands composing protein beta-sheets are bound either in parallel or in anti-parallel in particular by hydrogen bonds between amino acids' main chain chemical groups [4-6]. Each beta-strand is bound to another two strands, except for the edge strands [7, 8]. Hydrophobic ordering plays an important role in the arrangement of amino acids and of beta-strands within beta-sheets. Hydrophobic side chains tend to be located centrally in the beta-sheet [9]. The more hydrophobic the beta-strand, the more centrally located is the beta-strand within the sheet [10]. The observation was found to be sufficient to account for beta-strand ordering in half of the beta-sheets and evidence for hydrophobic ordering was found in three-quarters of the beta-sheets [10, 11]. The length of beta-strands was also observed to be often smaller for edge strands [10]. Another rule was noted for four amino acids' long strands: such beta-strands are central only if their hydrophilicity is smaller than 35% [12]. The last beta-strand in the protein sequence which is the closest to the protein C-terminus, was also found to be generally located at an edge for beta-sheets containing three to six strands [13]. Most three-stranded beta-sheets were found to be arranged in a sequential and anti-parallel order [14]. It was further reported that introduction of the positively charged amino acid lysine is sufficient to convert aggregating beta-strands within multimers into edge strands of monomers [15, 16]. Between two beta-sheets, interlocked pairs of beta-strands were identified as a common motif of protein structures [17, 18].

Protein structures were classified according to their fold [19-23]. The protein fold is straightforwardly derived from tertiary structures. While tertiary structure prediction from protein sequences remains a challenge for most proteins, their secondary structure is generally well predicted from their sequence [24-35]. Protein folding from a one-dimensional polypeptide chain into a three-dimensional compact protein globule was widely analyzed experimentally and theoretically [36-44].

An elementary protein folding step is defined here as the formation of a non-covalent bond between two atoms of the protein chain, such as a hydrogen bond. In this work, consideration of an elementary step of protein folding is shown to provide information on the three-dimensional structure from sequences.

Experimental procedures

The programs pdb2 and pdb23 are written in perl. Their entry files are single PDB references of protein structures or lists of them [45, 46]. The program output files are tables (.xls files). The program removes DNA and RNA structures as well as those of peptides and proteins of less than 50 amino acids and analyzes only the first protein chain given in the DBREF key of the PDB file.

The program pdb2 uses the protein sequence in the three-letter amino acid code as found within the SEQRES key of .ent PDB files. From each .ent PDB file, a text .txt file contains the values of DBREF, SEQRES characterizing the protein sequences and the number of alpha carbon atoms (CA) within the PDB file so as to identify missing atoms within the structure. The mass of each atom is taken as the number of its nucleons, except for the selenium atom which was given the mass of a sulfur atom for the calculations, so as to avoid the bias due to selenomethionines deriving from methionine substitutions engineered for crystal diffraction studies. The protein sub-sequences were noted if their length does not exceed 20 amino acids (cf. results). L is the number of amino acids in the protein chain. For integers i within the 1 to L range, and j within the i to i+20 range, each sequence S(i,j) corresponding to the peptide from amino i to amino acid j is taken into account. If its mass M is not a square, the sequence

[a] *Analyse algébrique, Institut de Mathématiques de Jussieu, Université Pierre et Marie Curie, Paris VI, France*
[b] *Centre d'Informatique pour la Biologie, Institut Pasteur, Paris, France*
[c] *Département de Virologie, Institut Pasteur, Paris, France*

* Corresponding author. Tel.: +33 144389496
E-mail address: jjestin@pasteur.fr (Jean-Luc Jestin)

Supporting information for this article is available at http://www.csbj.org

S(i,j) is rejected. If its mass is a square, that is if the value $M^{1/2}$ equals its integer part $I(M^{1/2})$, then the sequence S is said to be optimized for folding (cf. results): S(i,j) = SOF(i,j). For all values of i and j associated to a protein, the set of all SOF(i,j) is drawn within a graph in red: Figure 3 shows the case of the human transthyretin protein of PDB reference 1eta.

Using the program pdb23 for any beta-sheet named (sheetID), the number (V) of SOF of the protein chain is given for each amino acid (AA) in the three-letter code in the downloadable output file together with the mean number (V_m) of SOF per strand which is averaged over all amino acids of the beta-strand. To eliminate SOF sequences of length 1 corresponding to the unique amino acid cysteine, the SOF length was taken as (j-i) with its values in the range i+1 to j. Only beta-sheets with more than three beta-strands are taken into account (cf. discussion); beta-strands that are three or less amino acids long, are not considered within this analysis. Beta-sheets which do not contain edge strands such as those in beta-barrel structures have been excluded from this study.

Both programs can be used at the addresses (http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::pdb2) and (http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::pdb23).

The first training set of 29 structures (cf. supplementary material) was constituted by choosing one protein structure per fold in the SCOP database [22]. The two non-redundant test lists consisting of 83 protein structures from the PDB containing at least one open beta-sheet with more than three strands (cf. supplementary material) were established using the program check.pl by removal of proteins containing engineered substitutions within protein domains (except for the engineering of methionine to selenomethionine mutations whose impact for the calculation is described above). Proteins with similar functions and from similar organisms were also removed from the test sets. The all-alpha proteins found were further eliminated as they did not contain an edge strand within a beta-sheet. Protein homology within the test set was evaluated using the program Pisces [47]. The protein structures were visualized from pdbxxxx.ent PDB files using the software Pymol by highlighting their ribbon characterized by the amino acids' alpha carbons.

Results

A mechanical system consisting of a folding entity is modelled as a sphere (Figure 1). The reference frame is fixed with respect to the rotating folding entity so that its kinetic energy equals zero in this frame. A chemical group folding onto the folding entity is defined as the folding unit and is represented by a small sphere of mass m and velocity X. After folding, the folding entity is a larger sphere of mass M and velocity Y.

The kinetic energy of the folding unit is noted $mX^2/2$. After folding, the kinetic energy of the larger folding entity is $MY^2/2$. The internal energy released during folding is noted Ui. The difference in energy due to the breaking and the formation of bonds such as hydrogen bonds during the folding step is noted Ep. Energy conservation during the folding step can then be written as in equation (1):

$$\frac{mX^2}{2} = \frac{MY^2}{2} + E \qquad (1)$$

$$\text{with } E = Ui + Ep$$

Equation (1) is of special interest when considered over the field of rational numbers **Q**: for any given value of E, equation (1) has an infinite number of solutions in X and Y if (m/M) is a square (cf. Appendix). The folding of a mass m which is a square is further considered: for energy conservation to be ensured during the elementary folding step while having an infinite number of solutions in X and Y, it is sufficient for M to be a square. This condition prompted us to investigate the corresponding peptide sequences which are thereby optimized for folding. According to this model, if equation (1) has no solution in X and Y, then energy is not conserved during the elementary folding step and folding cannot proceed. Sets of protein sub-sequences with optimal folding properties (SOF) can be defined for any protein sequence. According to the elementary protein folding step (Figure 1), symmetry is gained during folding, as the small sphere of mass m on the surface of the folding entity yields a sphere after folding: the inequivalent group of mass m becomes equivalent to the other parts of the entity after folding. This formalism is used in this study to predict edge strands in protein beta-sheets (Figure 2).

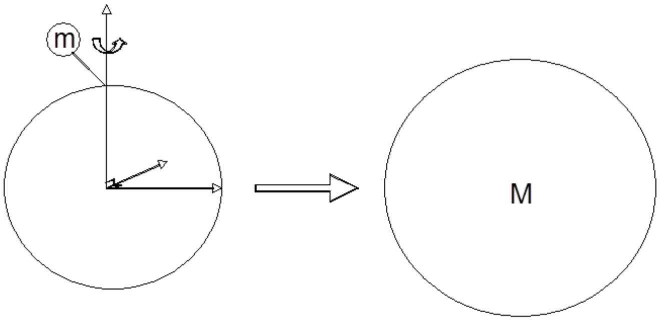

Figure 1. Elementary step for the folding of a small group of mass m onto the folding entity to yield a larger folding entity modelled by a sphere of mass M. Symmetry is gained during this elementary folding step.

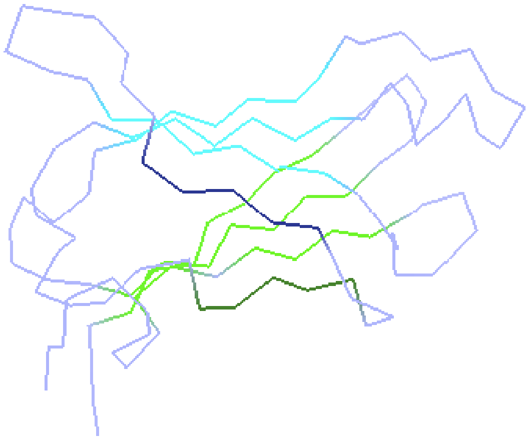

Figure 2. Representation of predicted edge strands in the structure of human transthyretin (PDB reference 1eta) [48]. Lines represent virtual bonds between the alpha carbons of adjacent amino acids in the protein. Two superimposed beta-sheets (blue and green) consisting of four beta-strands each contain two edge strands (dark blue and dark green) and predicted according to the rule.

The longer a sequence with optimal folding properties (SOF), the less stable it is upon amino acid substitution during evolution, given that the probability for an amino acid mutation to occur increases with the sequence length. Conversely, the shorter a SOF, the higher its robustness upon mutation. Accordingly, we did not consider SOF which are more than twenty consecutive amino acids long (Figure 3).

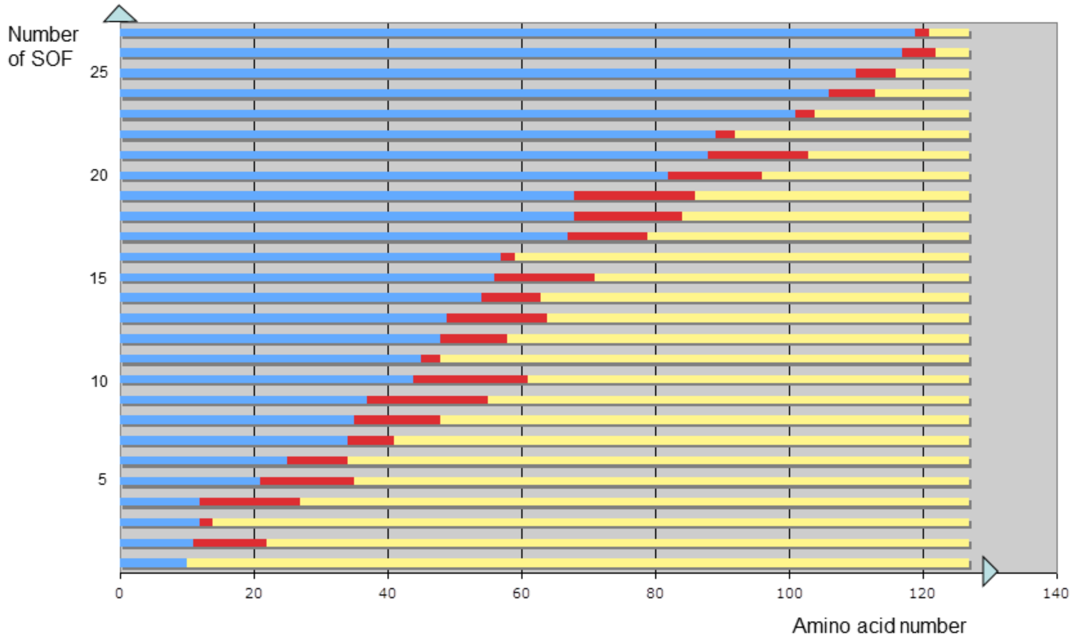

Figure 3. Set of sequences with optimal folding properties (SOF) highlighted in red for human transthyretin whose structure was described (PDB reference 1eta) [48]. The amino acid numbers are drawn on the horizontal axis. Each red segment corresponds to a peptide sequence with optimal folding properties (SOF).

Edge strands are bound to a unique other beta-strand within beta-sheets while central strands are bound to two other beta-strands, thereby highlighting distinct symmetry properties. As the elementary folding step changes the symmetry of the system (Figure 1), we reasoned that sequences with optimal folding properties (SOF) might be correlated with the position of beta-strands within sheets located either centrally or on the edge.

By using a first training set of 29 proteins, a correlation was noted between extreme values of the mean number of SOF for a strand (V_m) and its location on the edge of sheets of more than three strands. The following first rule was then established: the lowest value of V_m corresponds to an edge strand for ($0 \leq V_m < 0.34$). If a V_m value does not exist in this range for all strands of the beta-sheet, the maximal V_m value predicts an edge strand (Table 1).

Table 1. Edge strands and central strands of the human transthyretin structure. Extreme values of the number of SOF (V_m) highlighted in bold predict the edge strands noted in bold. The position of the strand in the sheet is central (C) or on the edge (E). Amino acids in the single-letter code are numbered according to the structure (PDB reference 1eta) [48].

Beta-strand	Position	V_m
P11-D18	C2	1.9
G53-H56	**E2**	**4.3**
R103-S112	C1	1.1
S115-T123	E1	1
V28-A36	C1	3
A45-T49	**E1**	**3.6**
E66-D74	C2	2.8
H90-A97	E2	2.3

A first test set of 83 protein domain structures was made of protein structures with at least one open beta-sheet of at least four strands of more than three consecutive amino acids. Out of 96 predictions, 59 edge strands (61.5%) were predicted correctly.

As all beta-sheets considered in this study are composed of two edge strands and of at least two central strands, there is a probability of one half or less for the random assignment of an edge strand. In order to compute the *p* value of the test, the probability of failing at most k times within n assays (one assay for each protein beta-sheet) when the probability of failure is taken as 0.5 was computed using the binomial law as in Equation (2):

$$p = \frac{\sum_{i=0}^{k} C_n^i}{2^n} \qquad (2)$$

where $C_n^i = \dfrac{n!}{i!(n-i)!}$

The severity of this statistical test is highlighted by the fact that the probability of 0.5 is only exact for four-stranded beta-sheets, while it is less for the other beta-sheets of five strands or more considered in this work. For the first test set, the *p* value obtained for n = 96 and k = 37 was less than 0.0158 and was therefore considered as significant as it is less than 5%.

To improve the rule, the first test set was then used as the second training set in which the 37 structures associated to incorrect predictions of edge strands were further analyzed. It was noted that the rule is not valid anymore in case a central strand's end is bound to a two-dimensional knot (2D knot, Figure 4); the extreme V_m value is then associated to this strand or these strands, but not to an edge strand. The two-dimensional knot is defined here as the crossing of the polypeptide's main chain on a two-dimensional representation of the protein's structure along two axes, either the beta-sheet's axis (which crosses two alpha carbons within the first and last strands and minimizes the sum for all the strands of the distances of an alpha carbon per strand to the axis; Figure 4B and 4D) or the axis which includes the alpha carbon at the strand's end and which is perpendicular to the sheet's plane defined at proximity of the strand's end by two alpha carbon atoms at positions m and m+2 in the strand and by one alpha carbon of the paired amino acid in an adjacent strand (Figure 4A and 4C). The 2D knot is within a loop between a beta-strand and a helix or between two strands of the sheet considered (Figure 4). A 2D knot is not a three-dimensional knot in the polypeptide chain.

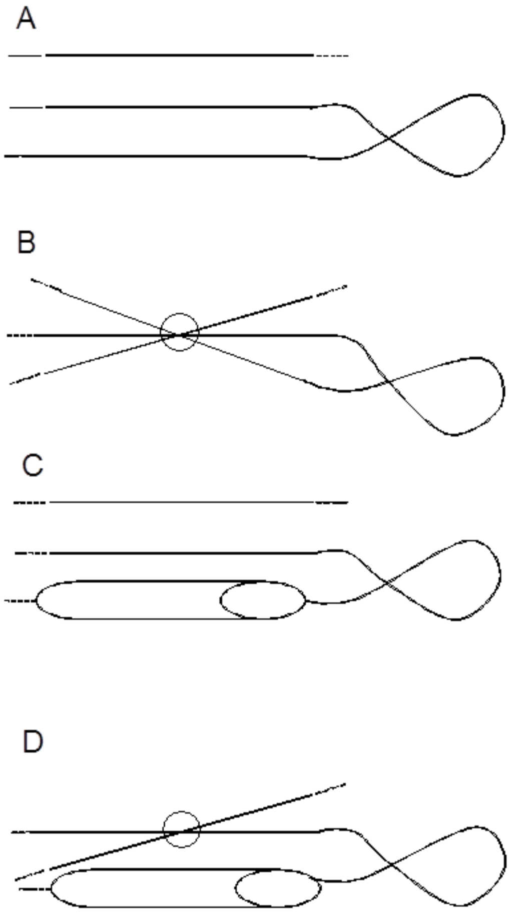

Figure 4. Bidimensional representations of polypeptide main chains containing 2D knots within loops between two beta-strands (A and B) or between a beta-strand and a helix (C and D). A beta-sheet's axis is represented by a small circle (B and D).

A second test set of 83 protein domain structures was then established to verify the improved rule (cf. supplementary material). 69 topological information predictions were found to be correct among the 92 predictions, thereby corresponding to a prediction accuracy of 75%. Use of equation (2) yielded a p value smaller than 8.4×10^{-7}. This upper limit of the probability for at most 23 prediction failures by random assignments among 92 assays indicates the finding's statistical significance, which is far below the commonly accepted standard threshold of 0.05.

As an amino substitution within an edge strand was shown to alter beta-sheet aggregation, a link between protein solubility and correct predictions of topological information was further investigated. The prediction of protein solubility from their sequence had been widely investigated [49-54]. Using the protein solubility prediction program Proso II (http://mips.helmholtz-muenchen.de/prosoII/prosoII.seam), the second test set was found to be composed of 44 soluble proteins and 39 insoluble ones. The correct topological prediction rate of 75% was not found to be significantly different for soluble proteins (78%) and for insoluble proteins (74%).

The protein domains of the second test set were found to be distributed over the three major domains which are Bacteria (41 chains), Eukarya (36 protein chains from Animalia (25), from Fungi (5), from Plantae (4) and from Protista (2)) and Archaea (3 chains) and include three viral proteins. In this test set, biases were finally noted towards human proteins (about one fifth of the protein chains), pathogenic micro-organisms (about one sixth of the chains from seven species), *Escherichia coli* proteins (one sixth of the chains) and proteins from thermophilic bacteria (one tenth of the chains). The second test set with proteins chosen according to different functions and organisms was then analyzed by looking for potential sequence homology using the culling server Pisces [47]. Accordingly, four pairs of sequences were found to have more than 40% identity, namely (1na7, 1zog), (1gav, 2ms2), (1nxw, 2pl1), (2boi, 2chh) which are still considered as different topological predictions; even though two sequences may be highly homologous, their sequence differences can yield two distinct and correct topological predictions by identification of the two edge strands for example. The topological information prediction using the improved rule has a statistical significance which remains unaltered.

Discussion

A major challenge in biological chemistry consists of the identification of relationships between protein sequences and their functions on genomic scales [55, 56]. While knowledge of a protein structure does not necessarily imply that a function can be identified for the protein, deciphering of protein domain structures remains of major interest and can provide clues for potential functions [57]. To circumvent expensive and time-consuming experimental techniques such as NMR or X-ray diffraction on protein crystals, promising approaches rely on computational biology, on the statistical analysis of known protein structures as well as on simulations of polypeptide chain dynamics [58]. Rules that relate the one-dimensional protein sequence and its three-dimensional structure properties were identified [9, 11, 16, 59-65]. The link between correlated mutations in multiple sequence alignments and interacting amino acids in the three-dimensional structure was extensively studied [66, 67]. Alignments of more than thousand well-chosen homologous protein sequences recently allowed the identification of a sufficiently large number of correlated mutations so as to decipher domain structures [68-70]. Three-stranded beta-sheets are generally arranged sequentially in anti-parallel [14]. These beta-sheets were not considered in this work which focussed on larger sheets of more than three beta-strands, first because of previously established rules [14], second because the statistical analysis carried out above would not be as straightforward as in Equation (2) (i.e. in the case of a three-stranded beta-sheet, the probability to identify an edge strand by random assignment is not one half or less) and third because the improved rule may not apply to three-stranded beta-sheets which were excluded during the analysis of the first training set.

In comparison to the topological information prediction accuracy of 75% described above, machine-learning approaches yielded edge strand prediction accuracies of 70% and 75.6% using support-vector machines [12, 71, 72]. Decision-tree algorithms allowed an 83% prediction accuracy to be obtained [12]. It should be of interest to combine different approaches to possibly improve further the edge strand prediction accuracy in protein beta-sheets. Interestingly, the notion of quasi-spherical random proteins was put in the context of natural proteins and introduced independently of this work [73]. The efficiency of the method used here shall be largely improved by applying it to several homologous sequences whose three-dimensional structures are expected to be similar.

Equations from classical mechanics are commonly treated over the field of real numbers. Using the field of rational numbers has been found of interest in different fields of the natural sciences connected to classical mechanics [74, 75]. It may constitute the basis for a new extension of theoretical chemistry [76]. In the field of genetic coding, substitution matrices made also use of discretized parameters such as p-adic integers or p-adic rational numbers [77-79]; it is of interest for the understanding of why the genetic code is the way it is. Importantly, the formalism described herein, i.e. the treatment of Eq.(1) was validated within the genetic code [74], providing thereby support for its application to proteins. In the field of biological chemistry, the genetic code is of special importance because of its quasi-universality within living organisms on earth for several billions of years [80, 81]. Experimentally, it has been the subject of numerous studies so as to develop applications in protein engineering [82-84]. Theoretically, Rumer noticed discrete symmetries linked to degeneracy in the genetic code [85, 86]. A rationale accounting for those discrete symmetries derived from the discrete nature of single-base mutations which have a major role in protein evolution [78, 79, 87, 88]. More recently, the codon arrangement in the genetic code was found to optimize kinetic energy conservation in polypeptide chains by considering the masses of the canonical amino acids: the formalism constituted by an equation from classical mechanics treated over the field of rational numbers was validated by the statistical significance of the codon arrangement within the genetic code [74].

The notion of protein sub-sequences with optimal folding properties (SOF) was introduced in this work. The elementary folding step allows the definition of criteria which are not necessary for folding, but which are sufficient to define protein sub-sequences with optimal folding properties. Edge strands are noticeable within beta-sheets as they are the only strands which pair with a unique other beta-strand; central strands generally pair indeed with two other beta-strands. Beta-barrel structures constitute an exception as they do not have edge strands, so that these structures were not considered here. The formalism suggested in this study allows the identification of sequences which optimize folding within proteins. Prediction of edge strands based on the consideration of the elementary folding step and of symmetry changes (Figure 1) is consistent with the fact that edge strands and central strands differ by their symmetrical properties with respect to neighbouring strands in protein beta-sheets. From our statistical analysis of hundreds of protein structures, we conclude that the formalism associated to the elementary folding step applied to given protein sequences allows information on the topology of their three-dimensional structure to be extracted.

It will be of interest to try and apply the same formalism to other secondary structure elements such as protein helices. The algorithm described herein to get topological information on beta-sheets from sequences for thousands of potential protein structure models provides a basis for a fast check of their quality. Applications within the critical assessment of techniques for protein structure prediction (CASP) are envisionned [89].

Acknowledgements

The authors thank A. Kempf for physical insights, B. Néron for interfacing the programs and reviewers for their criticisms.

Abbreviations

X and Y are velocities; m and M are respectively the masses of the folding unit and of the folding entity after the folding step; PDB is the Protein Data Bank; SOF is a Sequence Optimized for Folding; L is the number of amino acids of a protein chain (its length); p is the statistical p value as defined in equation (2).

Citation

Guilloux A, Caudron B, Jean-Luc Jestin JL (2013) A method to predict edge strands in beta-sheets from protein sequences. Computational and Structural Biotechnology Journal. 7 (9): e201305001. doi: http://dx.doi.org/10.5936/csbj.201305001

References

1. Efimov AV (1993) Standard structures in proteins. Prog Biophys Mol Biol 60: 201-239.
2. Banavar JR, Maritan A, Micheletti C, Trovato A (2002) Geometry and physics of proteins. Proteins 47: 315-322.
3. Rose GD, Fleming PJ, Banavar JR, Maritan A (2006) A backbone-based theory of protein folding. Proc Natl Acad Sci USA 103: 16623-16633.
4. Pauling L, Corey RB (1951) Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proc Natl Acad Sci USA 37: 729-740.
5. Salemme FR (1983) Structural properties of protein beta-sheets. Prog Biophys Mol Biol 42: 95-133.
6. Parisien M, Major F (2007) Ranking the factors that contribute to protein beta-sheet folding. Prot Struct Func Bioinf 68: 824-829.
7. Minor DL, Jr., Kim PS (1994) Context is a major determinant of beta-sheet propensity. Nature 371: 264-267.
8. Ruczinski I, Kooperberg C, Bonneau R, Baker D (2002) Distributions of beta-sheets in proteins with application to structure prediction. Prot Struct Func Genet 48: 85-97.
9. Von Heijne G, Blomberg C (1978) Some global beta-sheet characteristics. Biopolymers 17: 2033-2037.
10. Sternberg MJ, Thornton JM (1977) On the conformation of proteins: hydrophobic ordering of strands in beta-pleated sheets. J Mol Biol 115: 1-17.
11. King RD, Clark DA, Shirazi J, Sternberg MJ (1994) On the use of machine learning to identify topological rules in the packing of beta-strands. Protein Eng 7: 1295-1303.
12. Siepen JA, Radford SE, Westhead DR (2003) Beta edge strands in protein structure prediction and aggregation. Protein Sci 12: 2348-2359.
13. Sternberg MJ, Thornton JM (1977) On the conformation of proteins: towards the prediction of strand arrangements in beta-pleated sheets. J Mol Biol 113: 401-418.
14. Sternberg MJ, Thornton JM (1977) On the conformation of proteins: an analysis of beta-pleated sheets. J Mol Biol 110: 285-296.
15. Wang W, Hecht MH (2002) Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins. Proc Natl Acad Sci USA 99: 2760-2765.
16. Richardson JS, Richardson DC (2002) Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci USA 99: 2754-2759.

17. Kister AE, Finkelstein AV, Gelfand IM (2002) Common features in structures and sequences of sandwich-like proteins. Proc Natl Acad Sci USA 99: 14137-14141.
18. Papatheodorou TS, Fokas AS (2009) Systematic construction and prediction of the arrangement of the strands of sandwich proteins. J R Soc Interface 6: 63-73.
19. Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261: 552-558.
20. Orengo CA, Flores TP, Taylor WR, Thornton JM (1993) Identification and classification of protein fold families. Protein Eng 6: 485-500.
21. Chothia C, Hubbard T, Brenner S, Barns H, Murzin A (1997) Protein folds in the all-beta and all-alpha classes. Annu Rev Biophys Biomol Struct 26: 597-627.
22. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG et al. (2000) SCOP: a structural classification of proteins database. Nucleic Acids Res 28: 257-259.
23. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T et al. (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35: D291-297.
24. Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of accuracy and implications of simple methods for predicting secondary structure of globular proteins. J Mol Biol 120: 97-120.
25. Cohen FE, Sternberg MJ, Taylor WR (1980) Analysis and prediction of protein beta-sheet structures by a combinatorial approach. Nature 285: 378-382.
26. Zhu ZY, Blundell TL (1996) The use of amino acid patterns of classified helices and strands in secondary structure prediction. J Mol Biol 260: 261-276.
27. Rost B, Sander C (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci USA 90: 7558-7562.
28. Steward RE, Thornton JM (2002) Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. Proteins 48: 178-191.
29. Cheng J, Baldi P (2006) A machine learning information retrieval approach to protein fold recognition. Bioinformatics 22: 1456-1463.
30. Zimmermann O, Hansmann UH (2006) Support vector machines for prediction of dihedral angle regions. Bioinformatics 22: 3009-3015.
31. Kountouris P, Hirst JD (2009) Prediction of backbone dihedral angles and protein secondary structure using support vector machines. Bmc Bioinformatics 10: 437.
32. Zhang N, Duan G, Gao S, Ruan J, Zhang T (2010) Prediction of the parallel/antiparallel orientation of beta-strands using amino acid pairing preferences and support vector machines. J Theor Biol 263: 360-368.
33. Zafer A, Yucel A, Hakan E (2011) Bayesian models and algorithms for protein beta-sheet prediction. IEEE ACM Trans Comp Biol Bioinf 8: 395-409.
34. Subramani A, Floudas CA (2012) Beta-sheet topology prediction with high precision and recall for beta and mixed alpha/beta proteins. PLoS One 7: e32461.
35. Kountouris P, Agathocleous M, Promponas VJ, Christodoulou G, Hadjicostas S et al. (2012) A comparative study on filtering protein secondary structure prediction. IEEE-ACM Trans Comp Biol Bioinf 9: 731-739.
36. Anfinsen CB (1962) Some observations on the basic principles of design in protein molecules. Comp Biochem Physiol 4: 229-240.
37. Matouschek A, Kellis JT, Jr., Serrano L, Fersht AR (1989) Mapping the transition state and pathway of protein folding by protein engineering. Nature 340: 122-126.
38. Ptitsyn OB (1995) Molten globule and protein folding. Adv Protein Chem 47: 83-229.
39. Chaffotte AF, Guijarro JI, Guillou Y, Delepierre M, Goldberg ME (1997) The "pre-molten globule," a new intermediate in protein folding. J Protein Chem 16: 433-439.
40. Dobson CM, Sali A, Karplus M (1998) Protein folding: a perspective from theory and experiment. Angew Chem Int Ed 37: 868-893.
41. Dill KA (1999) Polymer principles and protein folding. Protein Sci 8: 1166-1180.
42. Mirny L, Shakhnovich E (2001) Protein folding theory: from lattice to all-atom models. Annu Rev Biophys Biomol Struct 30: 361-396.
43. Onuchic JN, Wolynes PG (2004) Theory of protein folding. Curr Opin Struct Biol 14: 70-75.
44. Thirumalai D, O'Brien EP, Morrison G, Hyeon C (2010) Theoretical perspectives on protein folding. Annu Rev Biophys 39: 159-183.
45. Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J et al. (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D 54: 1078-1084.
46. Rose PW, Beran B, Bi CX, Bluhm WF, Dimitropoulos D et al. (2011) The RCSB Protein Data Bank. Nucleic Acids Research 39: D392-D401.
47. Wang GL, Dunbrack RL (2003) Pisces: a protein sequence culling server. Bioinformatics 19: 1589-1591.
48. Hamilton JA, Steinrauf LK, Braden BC, Liepnieks J, Benson MD et al. (1993) The x-ray crystal structure refinements of normal human transthyretin and the amyloidogenic Val-30-->Met variant to 1.7-A resolution. J Biol Chem 268: 2416-2424.
49. Wilkinson DL, Harrison RG (1991) Predicting the solubility of recombinant proteins in Escherichia coli. Biotechnology 9: 443-448.
50. Idicula-Thomas S, Kulkarni AJ, Kulkarni BD, Jayaraman VK, Balaji PV (2006) A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. Bioinformatics 22: 278-284.
51. Magnan CN, Randall A, Baldi P (2009) SOLpro: accurate sequence-based prediction of protein solubility. Bioinformatics 25: 2200-2207.
52. Huang HL, Charoenkwan P, Kao TF, Lee HC, Chang FL et al. (2012) Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. BMC Bioinformatics 13 Suppl 17: S3.
53. Agostini F, Vendruscolo M, Tartaglia GG (2012) Sequence-based prediction of protein solubility. J Mol Biol 421: 237-241.
54. Smialowski P, Doose G, Torkler P, Kaufmann S, Frishman D (2012) PROSO II--a new method for protein solubility prediction. Febs J 279: 2192-2200.
55. Raes J, Harrington ED, Singh AH, Bork P (2007) Protein function space: viewing the limits or limited by our view? Curr Opin Struct Biol 17: 362-369.
56. Vernikos GS, Gkogkas CG, Promponas VJ, Hamodrakas SJ (2003) GeneViTo: Visualizing gene-product functional and structural features in genomic datasets. Bmc Bioinformatics 4: 53.
57. Juncker AS, Jensen LJ, Pierleoni A, Bernsel A, Tress ML et al. (2009) Sequence-based feature prediction and annotation of proteins. Genome Biology 10: 206.
58. Floudas CA (2007) Computational methods in protein structure prediction. Biotechnol Bioeng 97: 207-213.

59. De la Cruz X, Hutchinson EG, Shepherd A, Thornton JM (2002) Toward predicting protein topology: an approach to identifying beta hairpins. Proc Natl Acad Sci USA 99: 11157-11162.
60. Penel S, Morrison RG, Dobson PD, Mortishire-Smith RJ, Doig AJ (2003) Length preferences and periodicity in beta-strands. Antiparallel edge beta-sheets are more likely to finish in non-hydrogen bonded rings. Protein Eng 16: 957-961.
61. Zhu L, Yang J, Song JN, Chou KC, Shen HB (2010) Improving the accuracy of predicting disulfide connectivity by feature selection. J Comput Chem 31: 1478-1485.
62. Basu S, Bhattacharyya D, Banerjee R (2011) Mapping the distribution of packing topologies within protein interiors shows predominant preference for specific packing motifs. Bmc Bioinformatics 12: 195.
63. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB et al. (2012) Principles for designing ideal protein structures. Nature 491: 222-227.
64. Caudron B, Jestin JL (2012) Sequence criteria for the anti-parallel character of protein beta-strands. J Theor Biol 315: 146-149.
65. Bomar MG, Raghavender US, Spindel AWI, Kodukula K, Galande AK (2012) The ST pinch: A side chain-to-side chain hydrogen-bonded motif. Prot Struct Func Bioinf 80: 1259-1263.
66. Altschuh D, Lesk AM, Bloomer AC, Klug A (1987) Correlation of coordinated amino-acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol 193: 693-707.
67. Taylor WR, Jones DT, Sadowski MI (2012) Protein topology from predicted residue contacts. Protein Science 21: 299-305.
68. Seno F, Trovato A, Banavar JR, Maritan A (2008) Maximum entropy approach for deducing amino acid interactions in proteins. Phys Rev Lett 100: 078102.
69. Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. Proc Natl Acad Sci USA 109: 10340-10345.
70. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149: 1607-1621.
71. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20: 273-297.
72. Brown WM, Martin S, Chabarek JP, Strauss C, Faulon JL (2006) Prediction of beta-strand packing interactions using the signature product. J Mol Model 12: 355-361.
73. Brylinski M, Gao M, Skolnick J (2011) Why not consider a spherical protein? Implications of backbone hydrogen bonding for protein structure and function. Phys Chem Chem Phys 13: 17044-17055.
74. Guilloux A, Jestin JL (2012) The genetic code and its optimization for kinetic energy conservation in polypeptide chains. Biosystems 109: 141-144.
75. Madarasz JX, Szekely G (2013) Special relativity over the field of rational numbers. Int J Theor Phys 52: 1706-1718.
76. Thiel W (2011) Theoretical chemistry - Quo vadis? Angew Chem Int Ed 50: 9216-9217.
77. Khrennikov AY, Kozyrev SV (2009) 2-Adic clustering of the PAM matrix. J Theor Biol 261: 396-406.
78. Jestin JL (2010) A rationale for the symmetries by base substitutions of degeneracy in the genetic code. Biosystems 99: 1-5.
79. Jestin JL (2012) DNA mutations and genetic coding, In: DNA replication and mutation. Leitner RP (ed.) Nova. pp. 113-122.
80. Eigen M, Lindemann BF, Tietze M, Winkler-Oswatitsch R, Dress A et al. (1989) How old is the genetic code? Statistical geometry of tRNA provides an answer. Science 244: 673-679.
81. Di Giulio M (2005) The origin of the genetic code: theories and their relationships, a review. Biosystems 80: 175-184.
82. Döring V, Mootz HD, Nangle LA, Hendrickson TL, De Crécy-Lagard V et al. (2001) Enlarging the amino acid set of Escherichia coli by infiltration of the valine coding pathway. Science 292: 501-504.
83. Chin JW, Cropp TA, Anderson JC, Mukherji M, Zhang ZW et al. (2003) An expanded eukaryotic genetic code. Science 301: 964-967.
84. Hoesl MG, Budisa N (2011) In vivo incorporation of multiple noncanonical amino acids into proteins. Angew Chem Int Ed Engl 50: 2896-2902.
85. Rumer YB (1966) About the codon's systematization in the genetic code. Proc Acad Sci USSR 167: 1393-1394.
86. Shcherbak VI (1989) Rumer's rule and transformation in the context of the co-operative symmetry of the genetic code. J Theor Biol 139: 271-276.
87. Jestin JL (2006) Degeneracy in the genetic code and its symmetries by base substitutions. C R Biol 329: 168-171.
88. Jestin JL, Soulé C (2007) Symmetries by base substitutions in the genetic code predict 2' and 3' aminoacylation of tRNAs. J Theor Biol 247: 391-394.
89. Moult J, Fidelis K, Kryshtafovych A, Tramontano A (2011) Critical assessment of methods of protein structure prediction (CASP)--round IX. Proteins 79: 1-5.

Keywords:
Protein structure; structure prediction; topology; beta-strands; central strand; loop; turn

Competing Interests:
The authors have declared that no competing interests exist.