COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

CSBJ

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).

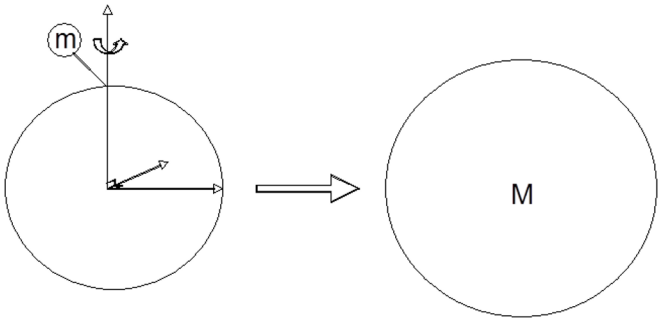

Figure 1. Elementary step for the folding of a small group of mass m onto the folding entity to yield a larger folding entity modelled by a sphere of mass M. Symmetry is gained during this elementary folding step.

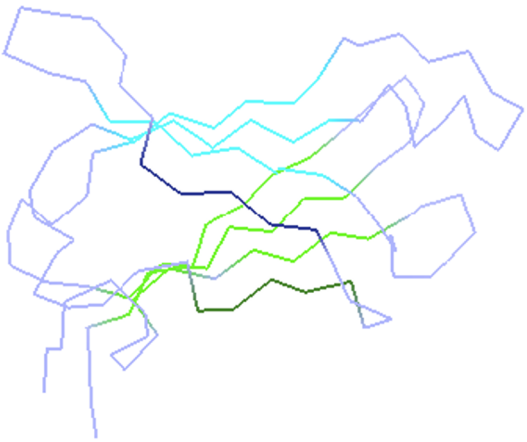

Figure 2. Representation of predicted edge strands in the structure of human transthyretin (PDB reference 1eta) [48]. Lines represent virtual bonds between the alpha carbons of adjacent amino acids in the protein. Two superimposed beta-sheets (blue and green) consisting of four beta-strands each contain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