Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 342 (2006) 441–451
BBRC
www.elsevier.com/locate/ybbrc

Amino acid propensities for secondary structures are influenced by the protein structural class

Susan Costantini [a,b,c], Giovanni Colonna [b,c], Angelo M. Facchiano [a,b,*]

[a] *Laboratorio di Bioinformatica e Biologia Computazionale, Istituto di Scienze dell'Alimentazione, CNR, Avellino, Italy*
[b] *Centro di Ricerca Interdipartimentale di Scienze Computazionali e Biotecnologiche, Seconda Università di Napoli, Italy*
[c] *Dipartimento di Biochimica e Biofisica, Seconda Università di Napoli, Italy*

Received 19 January 2006
Available online 8 February 2006

Abstract

Amino acid propensities for secondary structures were used since the 1970s, when Chou and Fasman evaluated them within datasets of few tens of proteins and developed a method to predict secondary structure of proteins, still in use despite prediction methods having evolved to very different approaches and higher reliability. Propensity for secondary structures represents an intrinsic property of amino acid, and it is used for generating new algorithms and prediction methods, therefore our work has been aimed to investigate what is the best protein dataset to evaluate the amino acid propensities, either larger but not homogeneous or smaller but homogeneous sets, i.e., all-α, all-β, α–β proteins. As a first analysis, we evaluated amino acid propensities for helix, β-strand, and coil in more than 2000 proteins from the PDBselect dataset. With these propensities, secondary structure predictions performed with a method very similar to that of Chou and Fasman gave us results better than the original one, based on propensities derived from the few tens of X-ray protein structures available in the 1970s. In a refined analysis, we subdivided the PDBselect dataset of proteins in three secondary structural classes, i.e., all-α, all-β, and α–β proteins. For each class, the amino acid propensities for helix, β-strand, and coil have been calculated and used to predict secondary structure elements for proteins belonging to the same class by using resubstitution and jackknife tests. This second round of predictions further improved the results of the first round. Therefore, amino acid propensities for secondary structures became more reliable depending on the degree of homogeneity of the protein dataset used to evaluate them. Indeed, our results indicate also that all algorithms using propensities for secondary structure can be still improved to obtain better predictive results.

Keywords: Amino acid propensities; Structural class of proteins; Protein structure; Secondary structure prediction; Statistical methods

The Anfinsen's experiments in the 1950s suggested that the primary amino acid sequence contains the information that specifies the folded native protein structure [1]. On the basis of this principle, in the 1970s some researchers developed methods to predict the native conformation of proteins from their amino acid sequences, one of the most challenging problems in molecular biology. The Chou and Fasman method [2,3], one of the early prediction methods, was based on a statistical procedure based on assigning conformation potentials, or propensities, to all amino acid residues. Conformation potentials, one for each type of secondary structure, are obtained from statistical analysis of proteins of known secondary structure, as ratio of the fractional occurrence of the residue in secondary structure elements of a given type to the fractional occurrence in all structures. For α-helix and β-strand propensities, each amino acid was classified as former, breaker or indifferent. These properties were used to identify potential α- and β-forming sites, which were then extended along the protein chain as long as the average propensity values calculated over a window of 5 or 6 residues were above a threshold value.

Several other prediction methods were developed over the years, based on different algorithms such as information theory [4], neural networks [5–8], nearest neighbour methods [9], multiple alignments [10–12], combination of

* Corresponding author.
E-mail address: angelo.facchiano@isa.cnr.it (A.M. Facchiano).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.01.159

multiple alignment and neural network [13], and hydrophobicity profiles [14]. These methods have reached relevant improvement in the accuracy of prediction, in comparison to the original Chou and Fasman method. Moreover, the reliability of the Chou and Fasman method has been criticized since the 1980s [15,16]. Nevertheless, many authors still use amino acid propensities or the Chou and Fasman method, for structure predictions [17–24] as well as for evolution studies [25] and in developing or evaluating new prediction methods [26–39]. The large use of this method may be due to its approach, simple but clear when compared to the most recent and accurate, but more sophisticated [40]. In fact, even if the original method has low accuracy in comparison to the most recent approaches, anyway the propensities for the different secondary structures represent intrinsic properties of amino acids and their use in developing new methods became successful [41,42].

The original dataset from which the Chou–Fasman parameters were calculated was quite small: it contained only 15 proteins, consisting of 2473 amino acid residues [2,43]. In 1989, the dataset was extended to include 64 proteins with 11,445 amino acid residues [3]. Later, in the 1998 the size of the dataset was expanded to include 144 proteins in order to analyse the reliability of the Chou–Fasman parameters [16].

The concept of protein structural classes was introduced by Levitt and Chothia [44] based on a visual inspection of polypeptide chain topologies in a dataset of 31 globular proteins. According to this concept, protein folds can be classified into one of four classes: all-α, all-β, α/β, and α + β. Since then, various quantitative classification rules have been proposed based on the percentages of α-helices and β-sheets in a protein [6,45–47].

In our work, we examined the problem to verify how the protein dataset used to compute amino acid propensities can affect the results, in particular, by using either large but not homogeneous datasets or smaller but homogeneous datasets consisting of only all-α, only all-β, or only α–β proteins. We have calculated the amino acid propensities for three types of secondary structures for 2168 proteins, i.e., the whole PDBselect dataset. Then, we subdivided these proteins in three secondary structural classes and calculated the amino acid propensities for each class. The prediction of secondary structure has been made using the different amino acid propensities calculated. Results are compared and discussed to evaluate the better criteria to choose protein dataset for computing amino acid propensities.

Methods

Database and definition of protein secondary structure. All analyses were performed using PDBselect [48] as a set of experimentally determined, non-redundant protein structures in the Protein Data Bank (see http://homepages.fh-giessen.de/~hg12640/pdbselect). We used the PDBselect list with <25% sequence homology, released in December 2003, which contained 2216 protein chains.

The secondary structure for every PDB entry was assigned by the DSSP algorithm [49] based on the analysis of backbone dihedral angles and hydrogen bonds. DSSP assigns seven different secondary structures, i.e., H: α-helix, G: 3_{10} helix, I: π-helix, E: extended strand, B: residue in isolated β-bridge, S: bend, and T: H-bonded turn. In addition, a "coil" state is assigned when no secondary structure is recognized. We applied the convention to define H, G, I as helix, E and B as strand, and others as coil [50,51]. 2168 of 2216 PDBselect proteins were accepted by DSSP for the analysis and constituted the total PDB subset for our work.

Assignment of structural class. The secondary structural content has been used to assign the protein secondary structural class, according to two different definitions of structural classifications. Nakashima et al. [45] consider proteins with >15% α-helical content and <10% β-strand content as all-α proteins, with <15% α-content and >10% β-content as all-β proteins, with >15% α-content and >10% β-content as mixed proteins, and the remaining as irregular.

According to the criterion of Chou [46], all-α proteins have at least 40% α-helical content and <5% β-strand content; all-β proteins have at least 40% β-strand content and <5% α-helical content; mixed proteins (considering the combination of α + β and α/β classes) contain more than 15% α-helical and 15% β-strand contents; irregular proteins have <10% α-helical and β-strand contents.

Propensity of amino acids in different secondary structural types. The residue propensity values in different secondary structural types (P_{ij}) were determined from the ratio of the residue's frequency of occurrence in helices, β-strand, and coil versus its frequency of occurrence in the protein subset:

$$P_{ij} = \frac{(n_{ij}/n_i)}{(N_j/N_T)},$$

where n_{ij} is the number of residues of type i in structure of type j, n_i is the total number of residues of type i, N_j is the total number of residues in structure of type j, and N_T is the total number of residues in the subset of PDB used in this analysis. Then, these values of propensities have been normalized as follows:

$$P_{ik}^{norm} = \frac{(P_{ik} - P_k^{min})}{(P_k^{max} - P_k^{min})},$$

where P_{ik} is the propensity of each amino acid in secondary structure element of type k (α, β or coil), P_k^{min} and P_k^{max} are the minimum and maximum values between the propensities P_{ik}.

Prediction of secondary structure. Starting from the N-terminal of each protein sequence, a running window of n amino acids is taken. The average value of α-helical propensities $\langle P_\alpha \rangle$, β-strand propensities $\langle P_\beta \rangle$, and coil propensities $\langle P_c \rangle$ has been determined for the n amino acids of each segment. These propensities have been determined by using windows of different lengths for three secondary structure elements (w_H, w_E, and w_c) and multiplied by different coefficients ($coeff_H$, $coeff_E$, and $coeff_c$). An exaustive scan for different windows and coefficients was made in order to find the values giving the better results.

The predicted secondary structure for the middle amino acid in the examined segment was assigned by choosing the higher value between the three average propensities of the segment. In this manner, if $\langle P_\alpha \rangle$ for one segment of length 7 is higher than $\langle P_\beta \rangle$ and $\langle P_c \rangle$, it has been assigned the secondary structure of type "H" to the 4th amino acid of that sequence. This procedure has been repeated for all proteins collected in the PDBselect database.

The prediction quality was examined by both resubstitution and jackknife tests.

Resubstitution test. The so-called resubstitution test is an examination for the self-consistency of a prediction algorithm. When the resubstitution test is performed for the current study, the secondary structure elements of each protein in a given dataset are predicted by the propensities derived from the same dataset, the so-called training dataset.

As a consequence, the propensities derived from the training dataset include the information of the protein used in the test. This will certainly give a somewhat optimistic error estimate because the same proteins are