Journal of Theoretical Biology 360 (2014) 109–116

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition

Loris Nanni [a,*], Sheryl Brahnam [c], Alessandra Lumini [b]

[a] DEI, University of Padua, Viale Gradenigo 6, Padua, Italy
[b] DISI, Università di Bologna, Via Venezia 52, 47521 Cesena, Italy
[c] Computer Information Systems, Missouri State University, 901 S. National, Springfield, MO 65804, USA

HIGHLIGHTS

- Protein structure identification
- Ensemble of several protein descriptors.
- Descriptors extracted from different protein representations.
- Support vector machines.

ARTICLE INFO

Article history:
Received 22 April 2014
Received in revised form
13 June 2014
Accepted 3 July 2014
Available online 12 July 2014

Keywords:
Protein structure class
Protein descriptors
Machine learning
Ensemble of classifiers
Support vector machines

ABSTRACT

Successful protein structure identification enables researchers to estimate the biological functions of proteins, yet it remains a challenging problem. The most common method for determining an unknown protein's structural class is to perform expensive and time-consuming manual experiments. Because of the availability of amino acid sequences generated in the post-genomic age, it is possible to predict an unknown protein's structural class using machine learning methods given a protein's amino-acid sequence and/or its secondary structural elements. Following recent research in this area, we propose a new machine learning system that is based on combining several protein descriptors extracted from different protein representations, such as position specific scoring matrix (*PSSM*), the amino-acid sequence, and secondary structural sequences. The prediction engine of our system is operated by an ensemble of support vector machines (SVMs), where each SVM is trained on a different descriptor. The results of each SVM are combined by sum rule. Our final ensemble produces a success rate that is substantially better than previously reported results on three well-established datasets. The MATLAB code and datasets used in our experiments are freely available for future comparison at http://www.dei.unipd.it/node/2357.

1. Introduction

The three-dimensional structure of a protein and its biological function are related: knowing a protein's structure is to know something about its biological function (Anfinsen, 1973). Since proper identification of a protein's structure assists researchers in predicting its function, it is no surprise that for the last twenty years the identification of protein structures has become one of the most active areas of research in computational biology, proteomics, and bioinformatics.

Structural features of proteins are typically described at four levels of complexity. The primary structure of a protein is essentially a polymer of 20 amino acids that constitutes the polypeptide chain. These amino acids are responsible for many functions in living organisms and provide clues for predicting secondary and tertiary structures and functions from polygenetic sequences. The secondary protein structure refers to the localized organization of parts of the polypeptide chain, which take on specific geometric arrangements – helices (α structures), strands (β sheets), and coils. The tertiary protein structure is the final specific geometric shape that a protein assumes, i.e., the complex and irregular folding of the peptide chain in three-dimensions. The quaternary structure describes the

* Corresponding author.
E-mail addresses: loris.nanni@unipd.it (L. Nanni), alessandra.lumini@unibo.it (S. Brahnam), sbrahnam@missouristate.edu (A. Lumini).

http://dx.doi.org/10.1016/j.jtbi.2014.07.003
0022-5193/

interactions between different peptide chains that make up the protein. Based on the work of Levitt and Chothia (1976), four structural classes of proteins are generally identified: all-α, which includes proteins with few strands; all-β, which includes proteins with few helices; α+β, which includes proteins with both helices and strands but with the strands segregated; and α/β, which includes proteins with both helices and strands but with the strands interspersed.

Protein structure prediction is the prediction of a protein's structure given its primary structure, or AA sequence. Predicting protein structure is an extremely difficult problem and is typically based on manually identifying the similarity of the folding patterns of already existing protein structures. Many large-scale sequencing projects have produced a tremendous amount of data on protein sequences, creating a huge gap between the number of identified sequences and the number of identified protein structures (Rost and Sander, 1996). Automated computational methods capable of fast and accurate prediction of protein structures would not only reduce this gap but also further our understanding of protein heterogeneity, protein–protein interactions, and protein–peptide interactions, which in turn would lead to better diagnostic tools and methods for predicting protein/drug interactions.

Key to the success in developing automated systems for protein structure prediction is the method of protein representation. Many methods of structural protein prediction are based on the simple amino acid composition (AAC), which represents a protein as a 20-dimensional vector corresponding to the frequencies of the 20 AAs in a given protein sequence (Chou, 1995; Nakashima et al., 1986). However, because this method ignores important sequential information and because similar AA sequences share similar folding patterns, AAC representations produce mediocre results. Chou's pseudo amino acid (PseAA) composition (Chou and Shen, 2007), one of the most studied methods of primary protein representation, overcomes some of the weaknesses of AAC by retaining additional information regarding a protein's sequential order along with the first 20 factors representing the components of the conventional ACC (Chou, 2001, 2009). Because the PseAAC approach (Chou, 2001, 2005) or Chou's PseAAC (Lin and Lapointe, 2013) has been widely and increasingly used, in addition to the web-server 'PseAAC' (Shen and Chou, 2008) built in 2008, recently three powerful open access softwares, called 'PseAAC-Builder' (Du et al. 2012), 'propy' (Cao et al., 2013), and 'PseAAC-General' (Du et al. 2014), were established: the former two are for generating various modes of Chou's special PseAAC; while the 3rd one for those of Chou's general PseAAC. PseAA has proven particularly effective in the prediction of protein structure on datasets with high-similarity but preforms poorly on datasets with low-similarity. To overcome this problem, Chou and Cai (2004) have proposed a representation called functional domain composition that is designed to handle low-similarity datasets.

Since proteins within the same class but with low sequence similarity show high similarity in their secondary structural elements, several methods have been proposed which utilize additional secondary structural prediction representations (Ding et al., 2012; Kong et al., 2014; Kurgan et al., 2008; Liu and Jia, 2010; Mizianty and Kurgan, 2009; Yang et al., 2010; Zhang et al., 2011). In (Jones, 1999), for instance, the authors use a protein's secondary structure representation for protein structural prediction that is based on the position specific scoring matrices (*PSSM*) generated by PSI-BLAST. Although such methods work relatively well for most classes, α/β and, especially, α+β classes have proven problematic. Moreover, most secondary structural representations focus on the content of secondary elements and ignore the useful information found in the position of secondary elements. The recent descriptors proposed by Kong et al. (2014) have produced promising results with α/β and α+β classes, and Dai et al. (2013) have proposed a powerful set of secondary features for protein structure classification, which they call position-based features of predicted secondary structural elements (PBF-PSSEs), that takes into consideration the position of secondary elements.

In addition to approaches based on ACC and *PSSM*, a large body of research has focused on the physicochemical and biochemical properties of individual amino acids, and representations of proteins have been proposed for predicting protein structural classes that are based on these properties (Bu et al., 1999). Using the physicochemical properties, a protein can be represented by a set of 20 numerical values taken from the amino acid index (AAindex) (Kawashima and Kanehisa, 2000).

Our goal in this work is to develop a system that performs better than previous predictors for protein structure classification. We accomplish this goal by building an ensemble of SVMs, where each SVM is trained using a different protein descriptor based on the following protein representations[1] (all of which are described in detail in Section 2):

- Position specific scoring matrix (*PSSM*) of proteins;
- Substitution matrix (SM) representation[2];
- Secondary structure elements (SSE) sequence.

In Section 2, we propose new descriptors based on SM and SSE. We test our ensemble of descriptors on three large datasets that are well-established in the literature. As reported in Section 3, our system significantly outperforms previous state-of-the-art approaches.

2. Pattern representation and feature extraction

In this study we apply a machine learning approach to the protein structure classification problem. The goal is to find an effective yet compact representation of proteins that is based on a fixed-length encoding scheme that can be coupled with a general purpose classifier, an approach that has been applied successfully to other biological problems, e.g., subcellular localization and protein–protein interactions (Chou and Shen, 2007; Nanni et al., 2010).

In this paper we investigate different protein representations (specifically, *PSSM*, SM, and SSE) and combine them with different machine learning methods for extracting descriptors from each representation. In other words, several descriptors are extracted from each representation, and then a set of fixed-length descriptors are extracted from these descriptors and used to train an SVM. Some descriptors are extracted multiple times, once for each physicochemical property considered in the extraction process. Representations that are based on physicochemical properties use the AAindex (Kawashima and Kanehisa, 2000) database[3], which currently contains 544 indices and 94 substitution matrices. An amino acid index is a set of 20 numerical values representing the different physicochemical properties of amino acids. Any given property of amino acids can be represented by a set of 20 numerical values, usually called a propensity scale, or amino acid index, where the *i*th numerical value is related to the *i*th amino acid of the alphabet.

Since many properties are highly correlated with each other, a property selection process is performed, as proposed in

[1] The representations are the ways used for storing a protein, i.e. the amino acid sequence or the PSSM matrix. They are not suited for training a machine learning system, e.g. large size, different size among different proteins. The descriptors are extracted from the representations and they are used to train a machine learning system.

[2] A substitution matrix describes the rate at which one character in a protein sequence changes to another character state over time. Substitution matrices are usually seen in the context of amino acid or DNA sequence alignments, where the similarity between sequences depends on their divergence time and the substitution rates as represented in the matrix (http://en.wikipedia.org/wiki/Substitution_matrix, accessed 07/14/2012).

[3] Available at http://www.genome.jp/dbget/aaindex.html. We have not considered the properties where the amino acids have a value 0 or 1.

Nanni et al. (2010). For all descriptors and databases, the same set of 25 properties is selected.

Because we adopt a fixed-length encoding scheme for our protein descriptors, we can use SVM[4] as the classifier. SVM (Cristianini and Shawe-Taylor, 2000) is a powerful binary-class prediction method that finds the equation of a hyperplane dividing a given training set into the two classes while simultaneously maximizing the distance between the two classes and the hyperplane. In the case where a linear decision boundary does not exist, kernel functions are used to project the data onto a higher-dimensional feature space that can be separated by a hyperplane. Typical kernels used in this case include polynomial and radial basis function kernels. We use radial basis function kernels in our system, with all features linearly normalized to [0 1] considering the training data. The radial basis function kernel have better classification performance compared with other kernels, as polynomial kernel or linear SVM, as mentioned in Yuan and Huang (2004).

Each descriptor tested in this work trains a different SVM. The scores of the SVMs are then normalized to mean 0 and standard deviation 1, considering the training data, and the final decision of the system is obtained by combining the scores by sum rule.

Below we describe in some detail each of the descriptors and feature extraction methods used in our system. Regardless of what descriptors is used, the final input must be a vector containing a set of discrete components. According to a comprehensive review (Chou, 2011), the general form of PseAAC for a protein sequence P is

$$P = [\Psi_1, \Psi_2, \ldots, \Psi_\Omega] \quad (1)$$

where Ω is the fixed length of the descriptor and depends on how to extract the information from the amino acid sequence.

2.1. Amino acid sequence (AAS)

As noted in the introduction, amino acids are the building blocks of proteins. All have common elements of an amine group, but the various functional groups give each amino acid distinct physical properties that influence protein formation and function.

Formally, AAS is a linear sequence of amino acids that is defined as follows:

$$P = [p_1, p_2, \ldots, p_N] \quad (2)$$

where $p_i \in \mathbf{A}$ and $\mathbf{A} = [A,C,D,\ldots,Y]$ is the set of 20 amino acids.

2.2. Position specific scoring matrix (PSSM) representation

Initially introduced by Gribskov et al. (1987) for the purpose of detecting distantly related proteins, *PSSM*[5] is generated from a group of sequences previously aligned by structural or sequence similarity and takes into consideration the following four parameters:

1. Position: the sequentially increased index of each amino acid residue in a sequence after multiple sequence alignment;
2. Probe: a group of typical sequences of functionally related proteins that have been aligned by sequence or structural similarity;
3. Profile: a matrix of 20 columns corresponding to the 20 amino acids;
4. Consensus: the sequence of amino acid residues that is most similar to all the alignment residues of probes at each position.

Formally, the *PSSM* representation for a given protein sequence of length N is an $N \times 20$ matrix, where each element $PSSM(i,j)$ is

$$PSSM(i,j) = \sum_{k=1}^{20} w(i,k) \times Y(j,k) \quad (3)$$

where $i=[1,\ldots,N]$, $j=[1,\ldots,20]$, $w(i,k)$ is the ratio of the frequency of appearance of the kth amino acid (considering the 20 amino acids) at position i of the probe and the total number of probes, and $Y(j,k)$ is the value of Dayhoff's mutation matrix between the jth and kth amino acids (i.e., $Y(j,k)$ is a substitution matrix).

2.3. Substitution matrix representation (SMR)

A variant of a representation method that was first proposed by Yu et al. (2011), an *SMR* for a given protein of length N is an $N \times 20$ matrix where each element $SMR(i,j)$ of the matrix is obtained using the following formula:

$$SMR(i,j) = M(P(i),j) \quad (4)$$

where $i=[1,\ldots,N]$, $j=[1,\ldots,20]$, $\mathbf{M}$ is a 20×20 substitution matrix (with element $\mathbf{M}_{i,j}$ representing the probability of amino acid i mutating to amino acid j during the evolution process), and $P=(p_1, p_2 \ldots, p_N)$ is a given protein. The representation matrix *SMR* of a given protein has the same dimension as its *PSSM* and can therefore be used to calculate the same set of fixed-length descriptors.

In this work, we consider representations obtained using 25 physicochemical properties, thereby obtaining 25 representations for training 25 separate SVMs.

2.4. Secondary structure elements (SSE) sequence representation

Protein secondary structure prediction takes an amino acid in a protein sequence and attempts to match it to one of the three secondary structural shapes: helix (H), strand (E), and coil (C). A number of computational approaches have been developed in the last few decades to predict the 3-state secondary structure from protein sequences.

In this study, PSIPRED (Jones, 1999) was chosen to predict protein secondary structure because it has been shown in Birzele and Kramer (2006) to outperform other competing prediction methods.

2.5. PSSM/SMR descriptors

2.5.1. Autocovariance matrix (AM)

Proposed in Yang et al. (2010), *AM* is designed to overcome the loss of local sequence-order information. This is accomplished by applying autocovariance variables to each column of the input matrix, thereby reducing each column to a fixed length.

In this work an autocovariance matrix (*AM*) is used to describe the average correlation between positions in a series of lags (i.e., the residue number when applied to protein sequences throughout the protein sequence).

Formally, *AM* is calculated, for its matrix representation *Mat*, as follows:

$$AM(k) = \frac{1}{N-lag} \sum_{i=1}^{N-lag} \left(Mat(i,j) - \frac{1}{N} \sum_{i=1}^{N} Mat(i,j) \right) \times \left(Mat(i+lag,j) - \frac{1}{N} \sum_{i=1}^{N} Mat(i,j) \right) \quad (5)$$

[4] SVM as implemented in the LibSVM toolbox available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[5] To extract the PSSM representation, first install PSI-BLAST and then issue the following MATLAB command: system('blastpgp.exe -i input.txt -d D:\PSI-BLAST\swissprot -Q PSSM.txt -j 3'), where input.txt is the protein sequence and PSSM.txt contains the PSSM matrix.

where N is the length of the sequence, $j=[1,\ldots,20]$, $lag=[1,\ldots, MaxLag]$ and $k=j+20\times(lag-1)$, $MaxLag=15$. Lag is the distance between one residue and its neighbors, and k is a linear index used to scan the cells of Mat and lag. In this approach the number of extracted features for a given protein is $\Omega=20+20\times(MaxLag-1)$.

So the descriptor of a protein sequence is: $[AM(1), AM(2), \ldots, AM(\Omega)]$.

2.5.2. 2D Discrete Cosine Transform (DCT)

DCT is a separable linear transformation that is widely used as an image compression technique for converting a signal into its elementary frequency components. The 2D transform is equivalent to performing *DCT* twice, first along one dimension and then along the second dimension.

The formal definition of the 2D *DCT* for a given input matrix A with M rows and N columns is as follows:

$$DCT(A)=a_i a_j \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} A_{m\times n}\cos\frac{\pi(2m+1)i}{2M}\cos\frac{\pi(2n+1)i}{2N},\quad 0\le i\le M, 0\le j\le N \tag{6}$$

where

$$a_i=\begin{cases}\frac{1}{\sqrt{M}} & i=0\\ \sqrt{\frac{2}{M}} & 1\le i\le M-1\end{cases}\quad\text{and}\quad a_j=\begin{cases}\frac{1}{\sqrt{N}} & j=0\\ \sqrt{\frac{2}{N}} & 1\le j\le N-1\end{cases} \tag{7}$$

In this work we use *DCT* to obtain compact descriptors from a matrix representation by retaining 400 discrete cosine coefficients. In this approach the number of extracted features for a given protein is $\Omega=400$, let us define DC the vector that contains the first 400 *DCT* coefficients:

$$DC((i-1)\times 20+j)=DCT(i,j);\qquad i=1:20,\quad j=1:20 \tag{8}$$

So the descriptor of a protein sequence is: $[DC(1), DC(2), \ldots, DC(\Omega)]$.

2.5.3. Ngram (NGR)

In 2-g descriptors are computed by counting the 2-g frequencies of occurrences from the amino acid sequence of a given protein, this approach is used in several papers, e.g. Ghanty and Pal (2009). Since there are 20 amino acids, there are 400 different combinations of amino acids and, therefore, 400 2-g features. Extracting 2-g features from the amino acid sequence is not an effective representation since the 2-g feature vector is comparatively much larger than the protein sequence and thus produces many zero values.

In Sharma et al. (2013) this problem is overcome by extracting 2-g descriptors directly from *PSSM* by counting the bi-gram frequencies of occurrences from *PSSM* linear probabilities. In this way, zero value components are avoided, producing a descriptor with more information.

Formally, the 2-g feature vector extracted directly from the *PSSM* can be described as follows. Let $\boldsymbol{PS}$ be the matrix representing the *PSSM* of a given protein (see Section 2.2). The matrix $\boldsymbol{PS}$ will have L rows, representing the amino acid sequence, and 20 columns. The frequency of occurrence of transition from the mth amino acid to the nth amino acid is computed as

$$B_{m,n}=\sum_{i=1}^{L-1}\boldsymbol{PS}_{i,m}\boldsymbol{PS}_{i+1,n} \tag{9}$$

where $1\le m\le 20$ and $1\le n\le 20$. This equation gives 400 frequencies of occurrences, $B_{m,n}$, producing the 400 element feature vector:

$$[B_{1,1}, B_{1,2}, \ldots, B_{1,20}, B_{2,1}, \ldots B_{2,20}, \ldots, B_{20,1}, \ldots, B_{20,20}]^T \tag{10}$$

where the superscript T is the transpose of the vector.

In this work we test this method considering both 2-g (NGR_2) and, by extension, 3-g (NGR_3) vectors (Paliwal et al., 2014).

2.5.4. A matrix-based descriptors: Pseudo PSSM (PP)

Given the *PSSM* of a protein, PP is widely used as a descriptor (Fan and Li, 2011; Jeong et al., 2011). In this work we extend this method to *SMR* to avoid a complete loss of sequence-order information.

Formally, given an input matrix Mat of dimension $N\times 20$, we define

$$E(i,j)=\frac{Mat(i,j)-\frac{1}{20}\sum_{v=1}^{20}Mat(i,v)}{\sqrt{\frac{1}{20}\sum_{u=1}^{20}\left(Mat(i,u)-\frac{1}{20}\sum_{v=1}^{20}Mat(i,v)\right)^2}} \tag{11}$$

where $i=[1,\ldots,20]$ and $j=[1,\ldots, 20]$.

$$PP(k)=\begin{cases}\frac{1}{N}\sum_{i=1}^{N}E(i,j) & k=1,\ldots,20\\ \frac{1}{N-lag}\sum_{i=1}^{N-lag}[E(i,j)-E(i+lag,j)]^2 & \mathrm{j}=1,\ \ldots,20,\ lag=1,\ldots,MaxLag\\ & k=20+j+20\bullet(\mathrm{lag}-1),\ MaxLag=15\end{cases} \tag{12}$$

The final descriptor is a vector PP of length 320. In this approach the number of extracted features for a given protein is $\Omega=40+\times(MaxLag-1)$. So the descriptor of a protein sequence is: $[PP(1), PP(2), \ldots, PP(\Omega)]$.

2.6. SSE descriptors

2.6.1. Set1

The first set of features is based on Kong et al. (2014), we changed some details of the extracted features due to the results obtained in the four datasets tested in this paper. It introduced a comprehensive set of features that reflect the general contents and special arrangements of the secondary structural elements of a given protein and that include four segment-level features specially designed for distinguishing proteins from the more difficult α+β and α/β classes.

Given the SSE of a protein, a simplified segment sequence (SS) is first created by removing all 'C' elements and converting every segment of the letter 'H' to 'α' and every segment of the letter 'E' to 'β' (see Kong et al., 2014 for details). The lengths of SSE and SS are denoted by N and N_1, respectively.

The set of features proposed by Kong et al. (2014) first include information about the contents of the secondary structural elements. The ratios of H and E elements in SSE are widely used and have been shown to improve prediction accuracy of protein structural classes (Kurgan et al., 2008). Thus, the first two elements are formulated as:

$$\begin{cases}\psi_1=N_{\mathrm{H}}/N\\ \psi_2=N_{\mathrm{E}}/N\end{cases} \tag{13}$$

where N_{H} and N_{E} are the number of H and E elements in SSE, respectively.

Although these two features reflect information about the contents of secondary structure, they ignore important sequence information. The next two elements are order-related features known as the secondary order composition moments of H and E, and they reflect the special arrangements of the secondary structural elements in SSE (Liu and Jia, 2010). Thus, the next two features are formulated as:

$$\begin{cases} \psi_3 = \sum_{j=1}^{N_H} p_{H_j}/N(N-1)) \\ \psi_4 = \sum_{j=1}^{N_E} p_{E_j}/N(N-1)) \end{cases} \quad (14)$$

where p_{H_j} and p_{E_j} are the jth order of H and E in SSE, respectively.

The length, or size, of secondary structural elements is important in deciding class. To utilize this information, the maximal lengths of α-helix and β-strand segments (denoted by Maxseg_H and Maxseg_E) are counted (Kurgan et al., 2008), while N_{LH} and N_{LE} are the number of occurrences of α segments with length >4 and β segments with length >2. This gives the following features:

$$\begin{cases} \psi_5 = \text{Maxseg}_H/N \\ \psi_6 = \text{Maxseg}_E/N \end{cases} \quad (15)$$

and

$$\begin{cases} \psi_7 = \text{N}_{LH}/N \\ \psi_8 = \text{N}_{LE}/N \end{cases} \quad (16)$$

Moreover, based on SS, a transition probability matrix TPM (Zhang et al., 2011) is constructed:

$$\text{TPM} = \begin{pmatrix} p_{\alpha\alpha} & p_{\alpha\beta} \\ p_{\beta\alpha} & p_{\beta\beta} \end{pmatrix} \quad (17)$$

where $p_{\alpha\alpha}=N_{\alpha\alpha}/(N_{\alpha\alpha}+N_{\alpha\alpha})$, $p_{\alpha\beta}=N_{\alpha\beta}/(N_{\alpha\beta}+N_{\alpha\beta})$, $p_{\beta\alpha}=N_{\beta\alpha}/(N_{\beta\alpha}+N_{\beta\alpha})$, $p_{\beta\beta}=N_{\beta\beta}/(N_{\beta\beta}+N_{\beta\beta})$ and where $N_{\alpha\alpha}$, $N_{\alpha\beta}$. $N_{\beta\alpha}$, and $N_{\beta\beta}$ are the number of substrings αα, αβ, βα, and ββ in SS.

Two features, which characterize the distributions of $\alpha-$ helix and $\beta-$ strand segments, are formally expressed as

$$\begin{cases} \psi_9 = (P_{\alpha\alpha}+P_{\beta\alpha})/2 \\ \psi_{10} = (P_{\alpha\beta}+P_{\beta\beta})/2 \end{cases} \quad (18)$$

Finally, the probability of α-helix and β-strand segments occurring in SS provides important information about the secondary structural segments. This information is extended to reflect the ability of a protein to form α-helix and β-strand segments in a given protein sequence. The final two features incorporating this information are based on both SSE and SS:

$$\begin{cases} \psi_{11} = N_\alpha/(N_1 N_H) \\ \psi_{12} = N_\beta/(N_1 N_E) \end{cases} \quad (19)$$

So the descriptor of a protein sequence is: $[\psi_1,\ldots,\psi_{12}]$.

2.6.2. Set2

Set 2 includes different content-based features of predicted secondary structure elements (CBF-PSSE) several of them proposed in this paper:

1. Predicted secondary structure elements (PSSE) content is one of the most widely used CBF-PSSEs (Kurgan et al., 2008; Liu and Jia, 2010; Mizianty and Kurgan, 2009; Zhang et al., 2011; Dai et al., 2013) and can be calculated by taking a sliding window and scanning through predicted secondary structure sequences as follows:

$$content_{SE} = \frac{Count_{SE}}{\sum_{x\in(C,H,E)} Count_x} \quad (20)$$

where $Count_{SE}$ is the total number of occurrences of the predicted secondary structure element SE, SE $\in\{C,H,E\}$, with H, E, and C denoting α-helix, β-strand, and coil, respectively. Let us define $\psi_1 = content_C$; $\psi_2 = content_H$; $\psi_3 = content_E$.

2. Another important feature is based on the composition moment vector (Dai et al., 2013), which is calculated as follows

$$CMV_{SE}^k = \frac{\sum_{j=1}^{Count_{SE}} PO_{SE_j}^k}{N^k} \quad (21)$$

where PO_{SEj} represents the jth position of the predicted secondary structure element SE, N is the length of the predicted secondary structure sequence, and k is the order of the composition moment vector (here $k\in\{1,2\}$). Let use store the features extracted in this step in $\psi_4,\ldots,\psi_9$.

3. Other four features are obtained considering the average length (normalized considering the length of the protein) of the segments of α and β (two features) and of α segments with length >4 and β segments with length >2 (other two features). Let use store the features extracted in this step in $\psi_{10},\ldots,\psi_{13}$.

4. Starting from transition probability matrix TPM (Zhang et al., 2011) constructed as in Section 2.6.1, i.e.

$$\text{TPM} = \begin{pmatrix} p_{\alpha\alpha} & p_{\alpha\beta} \\ p_{\beta\alpha} & p_{\beta\beta} \end{pmatrix} \quad (22)$$

We use as features its four components: $p_{\alpha\alpha}$; $p_{\alpha\beta}$; $p_{\alpha\beta}$; $p_{\beta\beta}$. Let use store the features extracted in this step in $\psi_{14},\ldots,\psi_{17}$.

5. We define SSc a variant of SS where the C elements are not removed. We use as features: the standard deviation of the length of each segment of the three elements (so three features are calculated, one for each of the three elements). Other two features are extracted as the standard deviation of the length of segments of α with length >4 and the standard deviation of the length of segments of β with length >2. Let use store the features extracted in this step in $\psi_{18},\ldots,\psi_{22}$.

6. Other three features are extracted as standard deviation (normalized considering the length of the protein) of the position of H, E, and C elements in SSE. Let use store the features extracted in this step in $\psi_{23},\ldots,\psi_{25}$.

So the descriptor of a protein sequence is: $[\psi_1,\ldots,\psi_{25}]$.

2.6.3. Set3

Here we proposed to extend the previous set of features with 75 features based on secondary structure shapes distance frequency extracted in the following way:

$$FP(k) = F(i,x,SSE); x=1,\ldots,3; i=1,\ldots,25; \text{k} = (x-1)\times 0+i \quad (23)$$

where $F(i,x,SSE)$ is the counts of the couple of the secondary structural element, helix (H) if $x=1$, strand (E) if $x=2$, and coil (C) if $x=3$, at a distance of i unit appeared in SSE.

For example, considering the SSE=HEEHHEEHHEHC we obtain

$$F(1,\text{H},\text{SSE})=2; \quad F(2,\text{H},\text{SSE})=1; \quad F(3,\text{H},\text{SSE})=3\ldots \quad (24)$$

So the descriptor of a protein sequence is $[FP(1),\ldots,FP(75)]$.

3. Experiments

3.1. Benchmark datasets

The present study has been evaluated on three widely used low-similarity benchmark datasets: FC699 (Kurgan et al., 2008), 1189 (Wang and Yuan, 2000), and 640 (Yang et al., 2010). Table 1

Table 1
The number of proteins belonging to the different structural classes in the four datasets.

Dataset	All-α	All-β	α/β	α+β	Total
FC699	130	269	377	82	858
1189	223	294	334	241	1092
640	138	154	177	171	640
25PDB	443	443	346	441	1673

presents for each database the number of samples in each of the four structural classes: all-α, all-β, α+β, and α/β.

As suggested in Kong et al. (2014) the 25PDB dataset (Kurgan and Homaeian, 2006) is used as training set, while the other three datasets FC699, 1189 and 640 are used as test sets (independent sets). Moreover, we have reported also the results obtained using the jackknife test for assessing the performance in the 25PDB dataset. In this way we can compare our approach with other papers where the 25PDB dataset is used for validating the proposed approaches.

3.2. Performance indicators

We examine the performance of our systems using the accuracy (or its complementary, the error rate) and the Mathew's correlation coefficient since this is the method most commonly employed in the literature.

Mathew's correlation coefficient (MCC) is regarded as a balanced measure of the quality of binary classifications even in the case where the two classes are of very different size. It returns a value proximal to 1 in the case of a nearly perfect prediction, 0 for a random prediction, and a negative value in the case of a disagreement between prediction and observation. For more details on MCC please read Xu et al. (2013, 2013a), Chen et al. (2013). MCC is defined as follows:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{25}$$

The aforementioned metrics are valid only for the single-label systems, as in the current case. For the multi-label systems whose emergence has become increasingly frequent in system biology and system medicine (Chen et al., 2012; Xiao et al., 2013), a completely different set of metrics as defined in a recent review (Chou, 2013) is needed.

3.3. Experimental results

The first experiment reported in Table 2 compares, in terms of accuracy, descriptors extracted from *PSSM* and *SMR*. To reduce the size of the table, we report only the most interesting results, along with the best performing weighted sum rule, *Fusion_Mat*, which is the fusion of $1 \times PSSM\text{-}DCT + 2 \times PSSM\text{-}AM + 2 \times PSSM\text{-}NGR\_3 + 1 \times PSSM\text{-}PP + 1 \times SMR\text{-}NGR\_2$. The weights and the methods chosen for this ensemble were based only on the training data (the test sets are never used for fixing parameters). All methods selected for building the ensemble are reported in Table 2. Inspecting Table 2 it is clear that *Fusion_Mat* outperforms its component methods.

In Table 3 we report the performance of the SSE descriptors, and some weighted sum rule:

- *Fusion_SSE*, which is the fusion of $1 \times Set1 + 2 \times Set2 + 1 \times Set3$;
- *Fusion_SSEa*, which is the fusion of $2 \times Set1 + 3 \times Set2 + 1 \times Set3$;
- *Fusion_SSEb*, which is the fusion of $2 \times Set1 + 4 \times Set2 + 1 \times Set3$;

Inspecting Table 3 it is clear that our proposed fusions increase the performance of *Set1* (based on the state-of-the-art approach (Kong et al., 2014) using our testing protocol).

Table 2
Comparison among approaches based on *PSSM* and *SMR*.

Accuracy rate	Dataset			
Descriptor Name	1189	640	FC699	25PDB
PSSM-DCT	76.7	75.2	80.4	55.5
PSSM-AM	85.3	84.2	85.9	70.4
PSSM-NGR_3	84.7	84.1	87.4	68.3
PSSM-PP	81.0	80.8	79.4	62.6
SMR-NGR_2	71.3	73.1	77.2	55.0
Fusion_Mat	**88.7**	**89.1**	**92.4**	**75.7**

Table 3
Comparison among approaches based on SSE.

Accuracy rate	Dataset			
Descriptor Name	1189	640	FC699	25PDB
Set1	81.0	82.0	87.3	79.8
Set2	81.5	81.7	90.0	80.0
Set3	72.6	73.1	88.8	73.5
Fusion_SSE	82.8	83.8	92.1	81.1
Fusion_SSEa	**83.2**	**83.9**	91.6	81.0
Fusion_SSEb	83.1	**83.9**	**92.2**	**81.3**

Table 4
Comparison among approaches based on SSE using jackknife testing protocol.

Accuracy rate	Dataset		
Descriptor Name	1189	640	FC699
Set1	80.7	81.4	92.5
Set2	83.3	81.0	92.2
Set3	72.2	69.1	88.9
Fusion_SSE	82.9	81.6	**93.4**
Fusion_SSEa	83.2	**83.6**	**93.4**
Fusion_SSEb	**83.3**	83.4	**93.4**

In Table 4 we report the performance of the approaches based on SSE using the jackknife as testing protocol, instead of the protocol proposed by Kong et al. (2014) that we have used in the other tables of this paper. It is clear that the performance is similar to that reported in Table 3.

In Table 5 we compare our best approach, the sum rule of both *Fusion_SSE* and *Fusion_Mat* (labelled *Here* in the following tables), with the state-of-the-art reported in the literature. In our tests the ensemble based on *PSSM* and *SMR* outperforms the ensemble based on SSE, anyway the fusion among *PSSM*, *SMR* and *SME* feature outperforms the stand-alone approaches and the state-of-the-art works. Only in the 25PDB dataset we do not outperform the state-of-the-art, the best performance in that dataset is obtained by Dai et al. (2013) (which instead works poorly in the 640 dataset).

Our proposed ensemble systems have been tested in different datasets without tuning the parameters of the approaches, so it could be considered quite robust system. Moreover, notice that in several papers, as Dai et al. (2013), the SVM parameters are chosen using the whole dataset, using a different grid search in each dataset. While here we have used always the same SVM parameters in all the datasets and for all the descriptors, for avoiding any overfitting in the datasets. If we use a grid search for choosing the SVM parameters, as Dai et al. (2013), our performance is boosted and we obtain an accuracy of 86% in the 25PDB dataset.

In Table 6 we compare class by class the MCC obtained by our best approach with the method proposed in Kong et al. (2014).

Table 5
Comparison with state-of-the-art.

DATASET	Method	References	Accuracy
FC699	Fusion_SSE	*Here*	92.1
	Fusion_Mat	*Here*	92.4
	Here	*Here*	94.5
	SCPRED	Kurgan et al. (2008)	87.5
	Liu and Jia	(Liu and Jia (2010)	89.6
	Dai et al.	Dai et al. (2013)	85.7
	Kong et al.	Kong et al. (2014)	92.0
1189	Fusion_SSE	*Here*	82.8
	Fusion_Mat	*Here*	88.7
	Here	*Here*	91.1
	SCPRED	Kurgan et al. (2008)	80.6
	MODAS	Mizianty and Kurgan (2009)	83.5
	RKS-PPSC	Yang et al. (2010)	81.3
	Zhang et al.	Zhang et al. (2011)	83.2
	Ding et al.	Ding et al. (2012)	82.0
	Dai et al.	Dai et al. (2013)	78.4
	Kong et al.	Kong et al. (2014)	83.5
640	Fusion_SSE	*Here*	83.8
	Fusion_Mat	*Here*	89.1
	Here	*Here*	91.4
	SCPRED	Kurgan et al. (2008)	80.8
	RKS-PPSC	Yang et al. (2010)	83.1
	Ding et al.	Ding et al. (2012)	83.4
	Dai et al.	Dai et al. (2013)	79.8
	Kong et al.	Kong et al. (2014)	84.5
25PDB	Fusion_SSE	*Here*	81.1
	Fusion_Mat	*Here*	80.4
	Here	*Here*	83.0
	SCPRED	Kurgan et al. (2008)	79.7
	MODAS	Mizianty and Kurgan (2009)	81.4
	RKS-PPSC	Yang et al. (2010)	82.9
	Zhang et al.	Zhang et al. (2011)	83.9
	Ding et al.	Ding et al. (2012)	84.3
	Dai et al.	Dai et al. (2013)	86.2

Table 6
Comparison by MCC.

DATASET	Method	All-α	All-β	α/β	α+β
FC699	Here	94.7	92.7	96.0	77.8
	Kong et al. (2014)	93.7	88.0	91.5	69.4
1189	Here	93.6	92.6	85.5	90.5
	Kong et al. (2014)	86.8	85.1	76.5	63.3
640	Here	94.4	90.4	88.4	82.1
	Kong et al. (2014)	92.6	83.3	80.0	64.3
25PDB	Here	90.2	79.3	69.1	59.6

Inspecting Tables 5 and 6 it is clear that our proposed fusion outperforms previous state-of-the-art approaches.

4. Conclusion

In this work we present a system for predicting protein structure classes using different protein representations and descriptors for training an ensemble of SVMs. The features that describe a given protein are obtained using representations based on *PSSM*, SM, and SSE. We present an empirical study where different feature extraction methods for representing proteins are compared and combined. Moreover, novel configurations are proposed and evaluated. The best performance is obtained when the different descriptors are combined by weighted sum rule. We also report a number of experimental results to assess the generality and robustness of our system, which was evaluated across three datasets and shown to outperform the state of the art.

In future studies we plan on examining different classification approaches and ensembles composed of other classifiers, such as AdaBoost and Rotation forest (Rodriguez et al., 2006). Unlike SVM, these classifiers require extra computational power when used in ensembles, which for the testing phase is not an issue but in the training phase this is a problem, especially when testing the robustness of our proposed systems by comparing several descriptors across several datasets.

The proposed system could be coupled with other systems for a further performance improvement. For example, Zhou et al. (2013) proposed a stringent DDI-based prediction approach for the host-pathogen protein–protein interaction, if Zhou et al. (2013) also include the approaches here proposed, Zhou et al. (2013) could have much better performance in their own prediction.

Because user-friendly and publicly accessible web-servers or code are essential for developing more useful predictors (Chou and Shen, 2009), the MATLAB code developed in this paper is freely accessible to the public. To maximize convenience of users as wished by many experimental scientist (Lin and Lapointe, 2013), we will make efforts to provide a web-server as well for the predictor presented in this paper.

References

Anfinsen, C., 1973. Principles that govern the folding of protein chains. Science 181, 223–230.

Birzele, F., Kramer, S., 2006. A new representation for protein secondary structure prediction based on frequent patterns. Bioinformatics 22, 2628–2634.

Bu, W.S., et al., 1999. Prediction of protein (domain) structural classes based on amino-acid index. Eur. J. Biochem. 266, 1043–1049.

Cao, D.S., Xu, Q.S, et al., 2013. Propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29, 960–962.

Chen, L, Zeng, W.M., et al., 2012. Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical–chemical interactions and similarities. PLoS One 7, e35254.

Chen W, Feng PM, et al. (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41: e69: open access at http://dx.doi.org/doi:10.1093/nar/gks1450.

Chou, K.-C., 1995. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 21, 319–344.

Chou, K.-C., 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Struct., Fucnt. Genet. 43, 246–255.

Chou, KC, 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.

Chou, K.-C., 2009. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr. Proteom. 6, 262–274.

Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). J. Theor. Biol. 273, 236–247.

Chou, KC, 2013. Some remarks on predicting multi-label attributes in molecular biosystems. Mol. Biosyst. 9, 1092–1100.

Chou, K.-C., Cai, Y.D., 2004. Predicting protein structural class by functional domain composition. Biochem. Biophys. Res. Commun. 321, 1007–1009.

Chou, K.-C., Shen, H.B., 2007. Review: recent progresses in protein subcellular location prediction. Anal. Biochem., 370.

Chou, K.-C. Shen, H.B. (2009) Review: recent advances in developing web-servers for predicting protein attributes, *Nat. Sci.*, **2**, 63–92 (openly accessible at ⟨http://www.scirp.org/journal/NS/⟩).

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press, Cambridge, UK.

Dai, Q., et al., 2013. Comparison study on statistical features of predicted secondary structures for protein structural class prediction: from content to position. BMC Bioinf. 14, 152.

Ding, S., et al., 2012. A novel protein structural classes prediction method based on predicted secondary structure. Biochimie 94, 1166–1171.

Du, P, Wang, X, et al., 2012. PseAAC-builder: a cross-platform stand-alone program for generating various special Chou's pseudo9amino acid compositions. Anal. Biochem. 425, 117–119.

Du, P, Gu, S, et al., 2014. PseAAC-general: fast building various modes of general form of Chou's pseudo9amino acid composition for large9scale protein datasets. Int. J. Mol. Sci. 15, 3495–3506.

Fan, G.-L., Li, Q.-Z., 2011. Predicting protein submitochondrion locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. Amino Acids 20, 1–11.

Ghanty, P., Pal, N.R., 2009. Prediction of protein folds: extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. IEEE Trans. Nanobiosci. 8, 100–110.
Gribskov, M., McLachlan, A.D., Eisenberg, D., 1987. Profile analysis: detection of distantly related proteins. Proc. Nat. Acad. Sci. (PNAS), 4355–4358.
Jeong, J.C., Lin, X., Chen, X.-W., 2011. On position-specific scoring matrix for protein function prediction. IEEE/ACM Trans. Comput. Biol. Bioinf. 8, 308–315.
Jones, D.T., 1999. Protein secondary structure prediction based on position specific scoring matrices. J. Mol. Biol. 292, 195–202.
Kawashima, S., Kanehisa, M., 2000. AAindex: amino acid index database. Nucleic Acids Res., 20.
Kong, L., Zhang, L., Lv, J., 2014. Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 344, 12–18.
Kurgan, L.A., Homaeian, L., 2006. Prediction of structural classes for protein sequences and domains-impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recognit 39, 2323–2343.
Kurgan, L.A., Cios, K., Chen, K., 2008. SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. BMC Bioinf. 9, 226.
Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–558.
Lin, SX, Lapointe, J, 2013. Theoretical and experimental biology in one. J. Biomed. Sci. Eng. (JBiSE) 6, 435–442.
Liu, T., Jia, C., 2010. A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. J. Theor. Biol. 267, 272–275.
Mizianty, M.J., Kurgan, L.A., 2009. Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. Bioinforma 10, 414.
Mizianty, MJ, Kurgan, L, 2009. Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. BMC Bioinf. 10, 414.
Nakashima, H., Nishikawa, K., Ooi, T., 1986. The folding type of a protein is relevant to the amino acid composition. J. Biochem. 99, 153–162.
Nanni, L., Brahnam, S., Lumini, A., 2010. A high performance set of PseAAC descriptors extracted from the amino acid sequence for protein classification. J. Theor. Biol. 266, 1–10.
Paliwal, K.K., Sharma, A., Lyons, J., Dehzangi, A., 2014. A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition. IEEE Trans. Nanobiosci. 44-50 (March).
Rodriguez, J.J., Kuncheva, L.I., Alonso, C.J., 2006. Rotation forest: a new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell. 28, 1619–1630.
Rost, B., Sander, C., 1996. Bridging the protein by structure predictions sequence-structure gap. Annu. Rev. Biophys. Biomol. Struct. 25, 113–136.
Shen, HB, Chou, KC, 2008. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem. 373, 386–388.
Sharma, A., et al., 2013. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. J. Theor. Biol. 320, 41–46.
Wang, Z.X., Yuan, Z., 2000. How good is prediction of protein structural class by the component-coupled method? Proteins 38, 165–175.
Xiao, X, Wang, P, et al., 2013. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal. Biochem. 436, 168–177.
Xu, Y, Ding, J, et al., 2013. iSNO-PseAAC: predict cysteine *S*-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS One 8, e55844.
Xu Y, Shao XJ, et al. (2013a) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine *S*-nitrosylation sites in proteins. PeerJ 1: e171: open access at ⟨https://peerj.com/articles/171.pdf⟩.
Yang, J., Peng, Z., Chen, X., 2010. Prediction of protein structural classes for low homology sequences based on predicted secondary structure. BMC Bioinf. 11, S9.
Yang, L., et al., 2010. Using auto covariance method for functional discrimination of membrane proteins based on evolution information. Amino Acids 38, 1497–1503.
Yu, X., et al., 2011. Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation. Amino Acids, http://dx.doi.org/10.1007/s00726-00011-00848-00728.
Yuan, Z, Huang, B, 2004. Prediction of protein accessible surface areas by support vector regression. Proteins 57, 558–564.
Zhang, S., Ding, S., Wang, T., 2011. High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure. Biochimie 93, 710–714.
Zhou, HF, Javad, R, Willy, H, Gao, S., Jin, J., Fan, M., Yong, CH, Wozniak, M, Wong., L, 2013. Stringent DDI-based prediction of H. sapiens-M. tuberculosis H37Rv protein–protein interactions. BMC Syst. Biol. 7 (no. 6), 1–15 (2013).