Artificial Intelligence in Medicine (2005) **xxx**, xxx–xxx

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences

Jishou Ruan [a], Kui Wang [a], Jie Yang [a], Lukasz A. Kurgan [b,*], Krzysztof Cios [c,d,e]

[a] *College of Mathematics and LPMC, Nankai University, Tianjin 300071, PR China*
[b] *Department of Electrical and Computer Engineering, University of Alberta, Alta., Canada*
[c] *University of Colorado at Denver and Health Sciences Center, Denver, CO, USA*
[d] *University of Colorado at Boulder, Boulder, CO, USA*
[e] *4cData, LLC, Golden, CO, USA*

Received 14 November 2004; received in revised form 22 January 2005; accepted 22 February 2005

KEYWORDS
Protein content prediction;
Composition vector;
Composition moment vector;
Primary protein sequence;
Secondary protein structure;
Proteomics;
Bioinformatics

Summary

Objective: One of interesting computational topics in bioinformatics is prediction of secondary structure of proteins. Over 30 years of research has been devoted to the topic but we are still far away from having reliable prediction methods. A critical piece of information for accurate prediction of secondary structure is the helix and strand content of a given protein sequence. Ability to accurately predict content of those two secondary structures has a good potential to improve accuracy of prediction of the secondary structure. Most of the existing methods use composition vector to predict the content. Their underlying assumption is that the vector can be used to provide functional mapping between primary sequence and helix/strand content. While this is true for small sets of proteins we show that for larger protein sets such mapping are inconsistent, i.e. the same composition vectors correspond to different contents. To this end, we propose a method for prediction of helix/strand content from primary protein sequences that is fundamentally different from currently available methods.
Methods and material: Our method is accurate and uses a novel approach to obtain information from primary sequence based on a composition moment vector, which is a measure that includes information about both composition of a given primary sequence and the position of amino acids in the sequence. In contrast to the composition vector, we show that it provides functional mapping between primary sequence and the helix/strand content.

* Corresponding author. Tel.: +1 780 492 5488; fax: +1 780 492 1811.
E-mail address: lkurgan@ece.ualberta.ca (L.A. Kurgan).

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2005.02.006

Results: A set of benchmarks involving a large protein dataset consisting of over 11,000 protein sequences from Protein Data Bank was performed to validate the method. Prediction done by a neural network had average accuracy of 91.5% for the helix and 94.5% for the strand contents. We also show that using the new measure results in about 40% reduction of error rates when compared with the composition vector results.
Conclusions: The developed method has much better accuracy when compared with other existing methods, as shown on a large body of proteins, in contrast to other reported results that often target small sets of specific protein types, such as globular proteins.

1. Introduction

Understanding protein functions is fundamental to understanding biological processes, and relies heavily on knowledge of the protein structure. Amino acid (AA) composition of thousands of proteins is widely available, e.g. in the Protein Data Bank (PDB) [1]. However, composition alone is not sufficient to determine protein function. Research in protein function is based on 3D shape information that, in turn, depends on its AA sequence. Thus, prediction of a protein structure from the AA sequence is an important computational and experimental goal. Experimental methods, such as X-ray crystallography and nuclear magnetic resonance spectroscopy, are used to determine protein structure. These methods, however, are time consuming, labor expensive and cannot be applied to some proteins [2]. Computational methods usually perform prediction of the 3D structure with an intermediate step of predicting secondary structure. The research mainly focused on development of advanced prediction methodologies, but still we are far away from providing a complete and accurate solution.

The 3D structure of proteins exhibits presence of several elements of the secondary structure. The dictionary of secondary structures of proteins [3] annotates each AA as belonging to one of eight secondary structure types: H (alpha-helix), G (3-helix or 3_{10} helix), I (5-helix or π-helix), B (residue in isolated beta-bridge), E (extended strand), T (hydrogen bond turn), S (bend) and "_" (any other structure). Typically, the above secondary structure types are reduced to three groups: helix (H, which includes types "H" and "G"), strand (E, which includes types "E" and "B") and coil (C, which includes all remaining types) [2]. Therefore, the secondary structures prediction aims to predict one of the three groups for each of the AA in the primary sequence. Using the coded representation for helix, strand and coil, the secondary structure of protein can be expressed as a sequence, called the secondary structure sequence, of the form: ...CCCEEECCHHHHCCCCCCEEEEEECCCCC... The secondary structure sequence has the same length as primary sequence.

The main thrust of computational methods for prediction of the secondary structure is to improve accuracy of the prediction [4]. Prediction of the secondary structure from AA sequence was initiated in late 1970s. The first approaches were based on information contained only in the primary AA sequence [5,6] and predicted three secondary structure types, with an accuracy of less than 60%. The next generation of methods considered information about 3–51 neighboring AAs through moving-window computations, and used pattern recognition and statistical methods [7], e.g. include methods based on Bayesian inference and decision rules [3,8,9]. Additional information, except the primary AA sequence, such as chemical properties of AAs based on polar–non-polar patterns and interactions [10,11], AA patterns in different types of helices [12], electric properties of AAs and their preferences in different structures [3] and structural features in side chain interactions [13,14] were also used, but the achieved accuracy was still less than 66%. In the 1990s, methods for prediction of secondary structure started to use information from alignments of sequences in protein sequence databases that match the query sequence. These methods achieved maximum accuracy of 78% [4]. However, since the alignment information cannot be found for a large number of proteins, the techniques using primary AAs sequence are still needed. Although a number of quality methods were developed and accuracy of the prediction continues to rise, protein structure prediction needs more work [4,7,15,16] and that was our motivation for the present work.

One of promising methods to improve accuracy of prediction of the secondary structure is to first learn the content of helix and strand structures in the sequence. This task is called gross-grain prediction of secondary structure content of proteins, or simply secondary structure content prediction. It is apparent that once the secondary structure content of a protein is known, the task of predicting sec-

ondary and tertiary structure becomes easier [17]. One of the reasons is that the secondary structure content can be predicted with much higher accuracy than the secondary structure itself. The secondary structure content can be learned by applying experimental or computational approaches. The experimental methods include spectroscopic methods, such as circular dichroism spectroscopy in the UV absorption range [18] and IR Raman spectroscopy [19]. Unsatisfactory accuracy and inconvenience of the experimental methods makes computational approaches worth pursuing [17]. The computational approaches have a long history and usually used statistical methods and information about AAs composition of proteins for prediction.

This paper introduces a new method for prediction of secondary structure content. Its characteristics are introduction of a new measure, called composition moment vector that is used to provide information necessary for prediction, with high accuracy, for both helix and strand content. We also note that our method was designed for prediction of content for large number of proteins, in contrast to other methods that predict content only for a specific type of proteins.

Below, we describe work relevant to the proposed method. Next, we introduce composition moment vector measure and compare it with the composition vector traditionally used for prediction. Finally, we describe neural network architecture for prediction of secondary structure and show results on a set of over 11,000 proteins.

2. Method

2.1. Background

Levitt and Chothia [11] observed that protein structures naturally group into four classes based upon the gross secondary structural content of their tertiary structures. Mitchie et al. [20] categorized proteins into three distinct structural classes based on the helix and strand content: mainly α-protein, mainly β-protein and $\alpha\beta$-protein. The task of prediction of secondary structure content aims to compute the amount of helix and strand structures in the secondary structure sequence. The amounts are described by counts of H and E structures divided by the length of the sequence. If the strand content is less than 5% then the protein is called as mainly α-protein. If the helix content is less than 5% then the protein is called as mainly β-protein. Otherwise, the protein is called as $\alpha\beta$-protein. Having this information before attempting to predict secondary structure would provide significant help in terms of improving prediction accuracy. Additionally, knowing the content information for some segments of primary AA sequences before predicting the secondary structure of the entire protein would provide even further improvement in the prediction accuracy. Prediction of content of helix and strand for the entire protein, as well as for protein fragments can be done using the described here method.

2.2. Relevant work

Traditional methods for the secondary structure content prediction use composition vector computed from the primary AA sequence. The vector consists of 20 elements, each being a normalized count of a specific AA in the primary sequence. The methods attempt to decompose the composition vector into three idealized component vectors, for helix, strand and coil, whose magnitudes are estimates of secondary structure content.

The prediction of secondary structure content started in 1973 by Krigbaum and Knutton who used multiple linear regression method to predict the content based on the composition vector [21]. Another approach, which used the composition vector, the molecular weight, and the presence or absence of the heme group in a protein as an input to a tandem neural network, was reported in Ref. [22]. In another work, the composition vectors and analytic vector decomposition technique were used in Ref. [23]. At the same time, an approach that used multiple linear regression method, the composition vector and structural class information was proposed [24,25]. A similar approach, using the composition vectors, auto correlation functions and multiple linear regression method was introduced in Ref. [26] and further improved by taking into account structural class information in Ref. [17].

The main difference between our method and the above methods lies in the measure used to perform prediction. The methods use the composition vector, supplemented by auxiliary information, such as structural class and molecular weight, to perform prediction. Although they achieved relatively high accuracy exceeding 90%, but it was done on small sets of less than 200 proteins, concerning specific class of proteins like globular proteins; they also assumed knowledge of structural classes [17,24–26]. Let us note that it is possible that the knowledge of structural classes and restriction on the number and type of used proteins resulted in good prediction results. As the class of considered proteins gets larger and the structural class is unknown the prediction accuracy usually degrades.

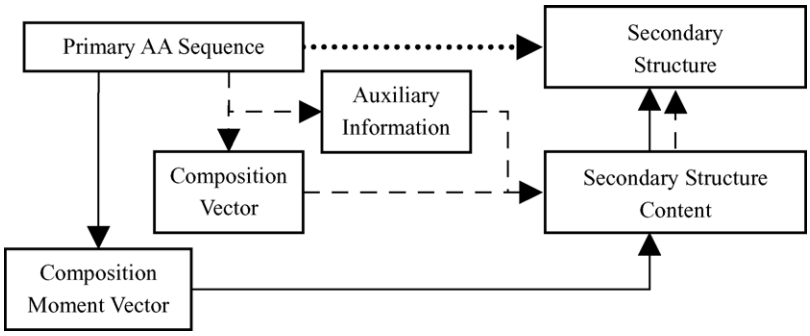

Figure 1 General framework for prediction of secondary structure in proteins.

In contrast, our method uses a novel measure, which we call composition moment vector. The composition moment vector is computed directly from the primary AA sequence. In the subsequent sections, we show theoretically and experimentally that the new measure performs better than the composition vector. The composition moment vector measure can also be used to characterize smaller protein segments. Such information can provide valuable help when performing prediction of secondary structure. Our method was tested on 11,000 proteins, not restricted to any particular protein family, and showed better results when compared with the use of the composition vector. It does so without using additional information about structural class, which is often difficult to obtain, especially for new proteins.

2.3. Architecture of the new method

A general framework for prediction of secondary structure of proteins is shown in Fig. 1. The dotted line shows an approach that does not utilize information about the secondary structure content. The dashed lines show an approach that utilizes information about the secondary structure content. Our method is illustrated by solid lines.

The introduced here method for prediction of secondary structure content consists of a two steps. A detailed architecture of the method is shown in Fig. 2. First, a composition moment vector, introduced in Section 2.4.2, is computed from a primary AA sequence. The vector provides complete numerical representation of the sequence. Next, the vector is input to a neural network that computes predicted amount of helix and strand contents. Each predicted value is computed by its own dedicated network. The predicted values constitute the secondary structure content values.

2.4. Composition moment vector and composition vector

Before we introduce the composition moment vector we provide motivation behind the new measure. The composition vector is a measure that is often used when performing the secondary structure content prediction [17,23–34]. Most often it is used to perform mapping between the primary AA sequence and the secondary structure content. In general, mapping between primary AA sequence and the secondary structure content is a function. This implies consistency, i.e. there are no two identical sequences with different secondary structure content. Therefore, the mapping between the composition vector, which is used to substitute the primary AA sequence, and the content, should also be a function. While this is true for work cited above, where only a small class of proteins, such a globular proteins, is considered, this mapping is inconsistent when considering larger protein body. The inconsistency means that several different AA sequences are represented by the same composition vector, but they have different secondary structure content. In

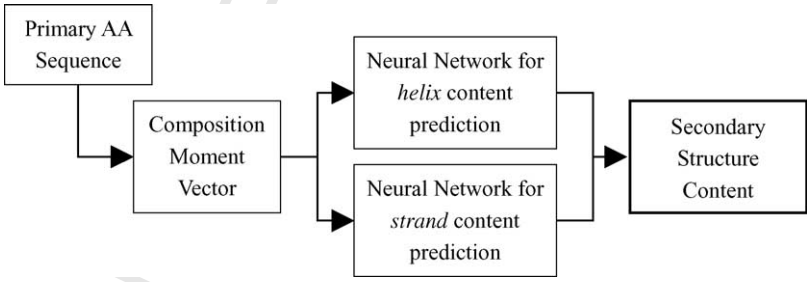

Figure 2 Detailed architecture of the method for prediction of secondary structure content.

Section 2.4.1, we discuss and show experimental results that support this statement. The inconsistency shows that, both in the sense of mathematics and biology, the composition vector cannot be used to perform reliable prediction of the secondary structure content for a large body of proteins. With the currently known secondary structure for only about 25,000 proteins, which is only a small fraction of the total number of proteins (NCBI protein database at ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/ contains approximately two million different proteins), a better measure is required.

2.4.1. Functional relation of the composition vector

Let $\xi_1, \xi_2, \ldots, \xi_N$ and $\eta_1, \eta_2, \ldots, \eta_N$ be the primary sequence and the secondary sequence, respectively. Vector consisting of counts of 20 AA in ξ_i, normalized by dividing by the length of the sequence is called the composition vector. The vector is used to derive a function between the primary AA sequence and the composition. The ratio of the total number of helices (denoted by H or α) in secondary structure sequence to the length of the protein is defined as $f_\alpha = \frac{n_\alpha}{N}$, and the corresponding ratio for strands (E or β) is defined as $f_\beta = \frac{n_\beta}{N}$. These ratios can be calculated directly from the primary AA sequence and define the helix and strand content. Mappings between secondary structure sequence and the helix content, and between secondary structure sequence and the strand content, are also functions. Based on the fact that all higher level protein structures are uniquely determined by their primary AA sequence, we conclude that mapping between the primary AA sequence and the secondary structure sequence should also be a function. If the primary AA sequence is represented by the composition vector, and the secondary structure sequence by the strand and helix content, the mapping between them also is regarded as a function.

Analysis of proteins stored in the PDB reveals that, in general, this mapping is not a function. To show it we performed an experiment using the PDB published in October 2000. Composition vector and the corresponding secondary structure content for about 6600 proteins were computed. Next, an exhaustive search was performed to find all pairs of proteins that have the same composition vector, but different structure content. The search returned 98 pairs of proteins, which have the same composition, but different helix or strand content. This shows that the mapping between the composition vector and the structure content is not a function when a large group of proteins is considered. This, in turn, means that the composition vector does not preserve all necessary information contained in the primary AA sequence, which is critical for consistent and accurate prediction of the secondary structure content.

2.4.2. Composition moment vector

A primary AA sequence contains information about the AA composition and their position. The composition vector, however, completely disregards the position information. Therefore, we propose a new measure, composition moment vector, which includes information about both composition and position of AA in the sequence. In contrast to the composition vector it also provides functional relation with the structure content, i.e. there must not be two or more primary AA sequences that would have different structure content but the same composition moment vector. The moment vector contains the same information as the composition vector plus the AA position information, and intuitively should give better description of the primary sequence. In addition, since it provides information about each AA in the primary sequence, it gives a more comprehensive description of the sequence than other measures, such as electronic or chemical groups [2], or general protein properties, such as average molecular weight and isoelectric point [22].

Let O be a protein in the PDB database, A_i be the ith AA, when the AA are ordered as follows: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y and (x_1, x_2, …, x_{20}) be the composition vector of O.

For an integer $k \geq 0$, we define *kth order moment vector* $(x_1^{(k)}, x_2^{(k)}, \ldots, x_{20}^{(k)})$ as

$$x_i^{(k)} = \frac{1}{N(N-1)\ldots(N-k)} \sum_{j=1}^{K_i} n_{ij}^k \quad \text{for } i = 1, 2, \ldots, 20 \tag{1}$$

where N is the length of the AA sequence, n_{ij} the jth position of the ith AA and K_i is the total number of the ith AA in the sequence.

Note that the zeroth order moment vector reduces to the composition vector.

To explain the new measure we show computation for the sequence AACDFFGGCKAWV. For the sequence $N = 13$. First, we compute K_i and n_{ij} as $K_1 = 3$ (count of A in the sequence), $n_{11} = 1$, $n_{12} = 2$ and $n_{13} = 11$ (positions of A in the sequence), $K_2 = 2$ (count of C in the sequence), $n_{21} = 3$ and $n_{22} = 9$ (positions of C in the sequence), etc. Next, we compute

$$x_1^{(1)} = \frac{1}{N(N-1)} \sum_{j=1}^{3} n_{1j} = \frac{1}{13 \times 12}(1 + 2 + 11) = \frac{14}{156} = 0.08974,$$

and all the remaining components of the first and higher moment vectors.

The *moment matrix* of protein O is defined as:

$$A_{K+1} = \begin{bmatrix} x_1 & x_2 & \cdots & x_{20} \\ x_1^{(1)} & x_2^{(1)} & \cdots & x_{20}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(K)} & x_2^{(K)} & \cdots & x_{20}^{(K)} \end{bmatrix} \quad (2)$$

where K is the maximal value among all K_i in O, for $i = 1, 2, \ldots, 20$.

It can be shown that two proteins are the same if and only if they have the same moment matrix, which assures functional relation between the matrix and the secondary structure sequence. The proof is shown in Appendix A.

Mappings between the moment matrix and f_α (helix content), and the moment matrix and f_β (strand content) are functions. However, computational complexity of computing the moment matrix prohibits using it for problems involving large number of long protein sequences. In order to cope with this high complexity problem the moment matrix is reduced to the zeroth and first moment vectors as $A_2 = \begin{bmatrix} x_1 & x_2 & \cdots & x_{20} \\ x_1^{(1)} & x_2^{(1)} & \cdots & x_{20}^{(1)} \end{bmatrix}$. The A_2 matrix, when represented as the $(x_1, x_2, \ldots, x_{20}, x_1^{(1)}, x_2^{(1)}, \ldots, x_{20}^{(1)})$ vector defines the *composition moment vector*.

The mapping between the composition moment vector and both helix and strand content in general is not a function, i.e. it is possible to construct two theoretical AA sequences that will have the same composition moment vector. For example, ACDEFGHIKLMNPQRSTVWYYWVTSRQPNMLKIHGFEDCA and YWVTSRQPNMLKIHGFEDCAACDEFGHIKLMNPQRSTVWY sequences have the same composition moment vector. On the other hand, such theoretical sequences do not constitute protein sequences that exist in nature. Therefore, when biological background is added to the underlying mathematical properties, the relation between the composition moment vector and structure content is a function. To prove this statement, an experiment using PDB (release #101) was performed. The database contains 18,604 protein files, and the total number of proteins with length greater than 3 is 34,218. For all protein sequences the following experiment was performed

1. For all sequences, $i = 1, 2, \ldots, 34{,}218$, compute the composition moment vector $A_2(i) = (x_1(i), x_2(i), \ldots, x_{20}(i), x_1^{(1)}(i), x_2^{(1)}(i), \ldots, x_{20}^{(1)}(i))$. If there is an incorrect symbol in sequence, which is not one of the AA symbols, then it is identified as ALA.
2. Exhaustively compare all $A_2(i)$ vectors. For each pair $A_2(i)$ and $A_2(j)$, where $A_2(i) = A_2(j)$ and $i \neq j$, and the corresponding two primary AA sequence are different, print the protein IDs of O_i and O_j.

In short, the experiment finds all protein pairs that are different but which are represented by identical composition moment vector. The experiment returned the following seven pairs of proteins: 1VANP and 1GO6D, 1SGPI and 1DS2I, 2RLNS and 1RBCS, 1NHP and 1JOA, 1PLMB and 1CJFC, 1LW8A and 1K3MA and 1JETB and 1B7HB. Table 1 shows the

Table 1 Protein pairs that are suspected to have the same composition moment vector (for long sequences "..." stands for identical fragments)

Returned pair	First protein sequence	Second protein sequence	Comments
1VANP and 1GO6D	LYS **ALA ALA**	LYS **DAL DAL**	Illegal symbol DAL
1SGPI and 1DS2I	VAL ASP CYS SER GLU TYR PRO LYS PRO ALA CYS THR **ALA** GLU ... GLY LYS CYS	VAL ASP CYS SER GLU TYR PRO LYS PRO ALA CYS THR **1LU** GLU ... GLY LYS CYS	Illegal symbol ILU
2RLNS and 1RBCS	LYS GLU THR ALA ALA ALA LYS PHE GLU ARG GLN HIS **NLE** ASP SER NH2	LYS GLU THR ALA ALA ALA LYS PHE GLU ARG GLN HIS **ALA** ASP SER NH2	Illegal symbol NLE
1NHP and 1JOA	MET LYS VAL ILE VAL LEU GLY SER SER ... PHE LEU SER **ALA** GLY MET GLN LEU ... LEU GLU ALA VAL LYS GLN GLU ARG	MET LYS VAL ILE VAL LEU GLY SER SER ... PHE LEU SER **CSO** GLY MET GLN LEU ... LEU GLU ALA VAL LYS GLN GLU ARG	Illegal symbol CSO
1PLMB and 1CJFC	PRO PRO PRO PRO PRO PRO PRO PRO	PRO PRO PRO PRO PRO PRO PRO PRO **PRO PRO PRO** **PRO PRO PRO PRO**	Incorrect sequence
1LW8A and 1K3MA	GLY **IIL** VAL GLU GLN CYS CYS ... GLU ASN TYR CYS ASN	GLY **ALA** VAL GLU GLN CYS CYS ... GLU ASN TYR CYS ASN	Illegal symbol IIL
1JETB and 1B7HB	LYS ALA LYS	LYS **NLE** LYS	Illegal symbol NLE

resulting pairs, and highlights in bold, fragments, which are different between the proteins in each pair.

Closer analysis of the results reveals that substituting illegal symbols by ALA results in identical primary sequences for the following pairs: 1VANP and 1GO6D, 1SGPI and 1DS2I, 2RLNS and 1RBCS, 1NHP and 1JOA, 1LW8A and 1K3MA and 1JETB and 1B7HB. The 1PLMB and 1CJFC contain incorrect sequences. The seven found protein pairs are in fact identical or incorrect, and because of that they have identical composition moment vectors. Therefore, if we assume that the used set of over 34,000 proteins provides representation of all proteins in nature, and then we can claim that two proteins are the same if and only if their composition moment vectors are the same. As the consequence, the mappings between the compositions moment vector and its helix and strand content, defined as

$$F : A_2 = \begin{bmatrix} x_1 & x_2 & \cdots & x_{20} \\ x_1^{(1)} & x_2^{(1)} & \cdots & x_{20}^{(1)} \end{bmatrix} \rightarrow f_\alpha \quad \text{and}$$

$$G : A_2 = \begin{bmatrix} x_1 & x_2 & \cdots & x_{20} \\ x_1^{(1)} & x_2^{(1)} & \cdots & x_{20}^{(1)} \end{bmatrix} \rightarrow f_\beta$$

are functions.

The composition moment vector is used to represent primary AA sequence when predicting the structure content. The vector is computed using only the primary AA sequence, and thus is easy to obtain in contrast to other approaches where, for example, information about the structural class is also required. Furthermore, the composition moment vector can be extended to predicting content of the fragments, which is a useful strategy for secondary structure prediction.

3. Prediction method

Since the mapping between the composition moment vector and structure content is a function we attempt to derive it computationally. Once found, the function can be used to predict the structural content for any given protein sequence. Several large protein datasets, described in the next section, were used to design a prediction method based on a neural network for proteins described using composition moment vector. The method is shown to significantly improve quality of prediction when compared with using composition vector.

3.1. Datasets

Two proteins sets were used to predict the structural content based on the composition moment vector, and the results compared with prediction when using the composition vector. The first was a general set of PDB proteins and was used to show numerical comparison. The second was a selected subset of PDB proteins that excluded proteins of low quality and homologous proteins of lower quality [35,36] to allow showing graphical comparison because of the smaller size.

The large dataset was extracted from the PDB (release #101). To assure good quality of data, the following was done:

- Peptides of length 3 or less were discarded.
- All proteins for which either the primary or secondary structure had any errors, such as illegal symbols, or inconsistencies in the secondary structure description were discarded.
- If more than one protein was recorded for a given protein file only the first protein was considered.

Among the available 18,604 protein files 11,206 sequences that satisfied the above rules were found. This dataset was further divided into these subsets:

- A_1 that consists of all sequences, for which secondary sequence contains both helices and strands; the sets consists of 9159 sequences.
- A_2 that consists of all sequences, for which secondary sequence contains helices, but does not contain any strands; the sets consists of 1642 sequences.
- A_3 that consists of all sequences, for which secondary sequence contains strands, but does not contain any helices, the set consists of 405 sequences.

For each of the above sets, the corresponding data that contain composition moment vector values were computed. The datasets are denoted as CMA_1, CMA_2 and CMA_3, respectively.

The second dataset includes proteins listed in the 25% PDB SELECT list released in October 2004 (the list of proteins can be obtained from http://homepages.fh-giessen.de/~hg12640/pdbselect/). It contains 2485 proteins whose secondary structure is either published in PDB database (release #103), or computed using DSSP procedure [37]. After removing nucleotide sequences and proteins with errors in the primary sequence 2439 proteins are left. Similarly to the large dataset it was divided into:

- B_1 that consists of 1707 protein sequences that have both helices and strands.
- B_2 that consists of 572 protein sequences that have only helices.

- B_3 that consists of 160 protein sequences that have only strands.

Analogically, CMB_1, CMB_2 and CMB_3 were also computed.

3.2. The neural network prediction

The prediction of secondary proteins content was performed using neural networks (NN). The universal approximation theorem for NNs states that every continuous function that maps intervals of real numbers to some output interval of real numbers can be approximated arbitrarily closely by a multi-layer feed forward network with one hidden layer [38]. Therefore, such network was used to approximate the F and G functions. Two NNs were designed for each prediction task, respectively. The networks topology consisted of 40 input nodes, 81 hidden nodes and 1 output node. Details are given in Appendix B. As shown in Fig. 2, one network was used to predict secondary structure content of helices, and another to predict secondary structure content of strands.

3.2.1. Prediction with large PDB dataset

The large PDB dataset was used for numerical comparison. After training one NN with the CMA_1 dataset, with back-propagation learning for 1000 epochs, and providing the helix content on the output, the following results were obtained:

- The average square error was 0.004197, and the average accuracy of the trained NN was 93.5% for the proteins in dataset A_1.
- On average, the number of samples for which the square error was less than 0.0025 was 5091, for less than 0.036 it was 6246 and for which variance was less than 0.0410 it was 6506.

The trained network was used to predict the helix content for the proteins from CMA_2 dataset:

- The average square error was 0.060115, and the average accuracy of the trained NN was 75.5% for the proteins in dataset A_2.
- On average, the number of samples for which square error was less than 0.0025 was 226, for less than 0.04 it was 874, and for less than 0.06 it was 1094.

The trained network was also used to predict the helix content for the CMA_3 dataset:

- The average square error was 0.000013 and the average accuracy of the trained NN was almost 100% for the proteins in dataset A_3.
- On average, the number of samples for which the square error was less than 0.00001 was 402.

Combining the results from all three datasets, the average accuracy for prediction of the helix content is not less than

$$0.935 \times \frac{9159}{11206} + 1.0 \times \frac{405}{11206} + 0.755 \times \frac{1642}{11206} = 0.91637 \approx 91.6\%$$

The second NNs were trained with the CMA_1 dataset, providing the strand content on the output. The following results were obtained:

- The average square error was 0.00355 and the average accuracy of the trained NN was 94.0% for the proteins in dataset A_1.
- On average, the number of samples for which square error was less than 0.0025 was 5909, for less than 0.035 it was 6706 and for which the variance was less than 0.0035 it was 6706.

The trained network was used to predict the strand content for the proteins from CMA_2 dataset:

- The average square error was 0.000013, and the average accuracy of the trained NN was almost 100% for the proteins in dataset A_2.
- On average, number of samples for which square error was less than 0.00001was 1625.

The trained network was used to predict the strand content for the proteins from CMA_3 dataset:

- The average square error was 0.02063 and the average accuracy of the trained NN was 86.6% for the proteins in dataset A_3.
- On average, the number of samples for which square error was less than 0.0025 was 116, for less than 0. 0049 it was 163 and for less than 0.01 it was 205.

The combined average accuracy for prediction of the strand content is not less than

$$0.94 \times \frac{9159}{11206} + 0.866 \times \frac{405}{11206} + 1.0 \times \frac{1642}{11206} = 0.94612 \approx 94.6\%$$

The results show that network trained to predict the helix content has the same accuracy for A_1 dataset compared with the network trained to predict the strand content. The network trained to predict the helix content performs almost perfectly on A_3 dataset, and similarly the network for the strand content for the A_2 dataset. Therefore, these

two networks provide very accurate results for computing structural class of proteins. When the network for helix content prediction shows value very close to zero, this means that the corresponding protein is mainly β-protein. Similarly, when the network for strand content prediction shows value close to zero, than corresponding protein is mainly α-protein. If neither of the above two happens, the protein is classified as αβ-protein.

3.2.2. Comparison of prediction using the composition vector versus the composition moment vector

Another study that uses the composition vector to predict the strand and helix content was performed. Results of this experiment are compared with the results discussed in the previous section. The analysis aims to confirm the benefits of introducing the composition moment vector.

Similarly, as for the experiments in the previous section, the large PDB set of 11,206 sequences was divided into the A_1, A_2 and A_3 sets. For each of these sets, a datasets that contains corresponding composition vector values was computed. The datasets are denoted by CA_1, CA_2 and CA_3, respectively. The training and testing procedures are identical to the procedures used above.

After training one of the NNs with the CA_1 dataset, and providing the helix content on the output, the results are:

- The average square error was 0.018397, and the average accuracy of the trained NN was 86.44% for the proteins in dataset A_1.
- The average square error was 0.039594, and the average accuracy of the trained NN was 80.1% for the proteins in dataset A_2.
- The average square error was 0.000452, and the average accuracy of the trained NN was 97.87% for the proteins in dataset A_3.

Combining the results achieved for the three datasets, the average accuracy for prediction of the helix content is not less than

$$0.8644 \times \frac{9159}{11206} + 0.9787 \times \frac{405}{11206} + 0.801 \times \frac{1642}{11206} = 0.8592 \approx 86\%$$

Similarly, the second NN was trained with the CA_1 dataset, and providing the strand content on the out, to predict content of the strand. The results are:

- The average square error was 0.010313, and the average accuracy of the trained NN was 89.85.0% for the proteins in dataset A_1.
- The average square error was 0.000034, and the average accuracy of the trained NN was almost 100% for the proteins in dataset A_2.
- The average square error was 0.02318, and the average accuracy of the trained NN was 84.88% for the proteins in dataset A_3.

Therefore, the average accuracy for prediction of the strand content is not less than

$$0.8985 \times \frac{9159}{11206} + 0.8488 \times \frac{405}{11206} + 0.994 \times \frac{1642}{11206} = 0.9168 \approx 91.7\%$$

Table 2 summarizes the comparison between results obtained by the composition vector and by the composition moment vector. The results show that the composition moment vector is better for prediction of both helix and strand content, when compared to using the composition vector. Using the new measure resulted in 40% reduction of error rates in case of helix content prediction and 35% reduction of error rates in case of strand content prediction.

The above results are compared graphically, computed based on the 25% PDB SELECT set. Similarly to the experiments performed with the large PDB, two NNs for prediction of helix and strand content were trained, with back-propagation learning method for 5000 epochs, on the CMB_1 dataset, and another two NNs were trained using identical set-up for the CB_1 dataset (the CB_1 contains composition vector values for proteins from B_1). The results of prediction using

Table 2 Comparison between prediction accuracy results by the composition vector and the composition moment vector

Predicted content	Dataset input data	A_1 (%)	A_2 (%)	A_3 (%)	Average prediction accuracy (%)
Helix	Composition vector	86.4	80.1	97.8	86.0
	Composition moment vector	93.5	75.5	100.0	91.6
Strand	Composition vector	89.9	100.0	84.9	91.7
	Composition moment vector	94.0	100.0	86.6	94.6

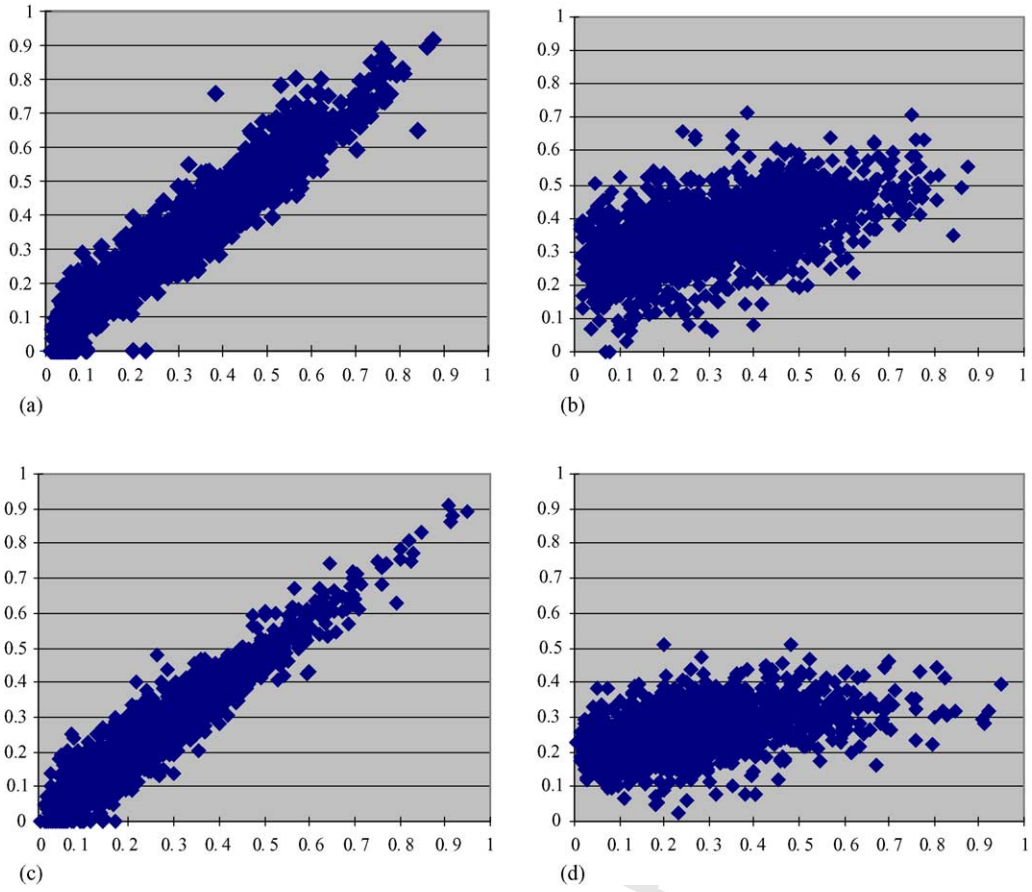

Figure 3 Comparison of results using composition moment vector vs. using composition vector on dataset B_1: (a) prediction of helix content using CMB_1; (b) prediction of helix content using CB_1; (c) prediction of strand content using CMB_1 and (d) prediction of strand content using CB_1.

the four NNs on the B_1 dataset are summarized in Fig. 3.

The plots show the actual content on the *x*-axis, while the predicted content is shown on the *y*-axis. The results on the diagonal indicate perfect predictions; the further from the diagonal the lower is the prediction quality. The results clearly show the prediction based on the composition moment vector yields better, more compact results, which are closer to the diagonal line. On the other hand, results of prediction with composition vector for both helix and strand contend are worse and align more horizontally. The quality of the results is described based on average error defined in Section 3.3.2. For composition moment vector based prediction it equals to 0.049 for helix prediction and 0.039 for strand prediction, while for composition vector it equals to 0.118 and 0.106 for helix and strand, respectively. Significant, about 60% error reductions for both helix and strand content, when using composition moment vector, was achieved.

3.3. Comparison with other prediction methods

Let us note that results achieved by other protein content prediction methods obtained on small, and unfortunately different protein data sets. Their accuracy ranged between 85 and 97% [17]. The methods usually uses training and testing protein sets that concern the same protein family, e.g. globular proteins, and do not provide results that can be generalized to an overall set of proteins. We also note that some of the existing methods assume knowledge of structural class, which is not available for majority of proteins and provides significant advantage to the methods that use them. In what follows, a side-by-side comparison of our method with other leading protein content prediction method is presented. The comparison is based on two experiments. The first involves test on a common small dataset to compare with another leading secondary content prediction method. The second test involves a jackknife procedure on larger and

general protein data sets, which is compared with results of several other prediction methods.

3.3.1. Comparison on a common dataset with leading multiple linear regression prediction method

The experiment is performed using a small test set of 143 proteins (28 α-proteins, 42 β-proteins and 73 $\alpha\beta$-proteins) based on the work SCOP [39]. Our method, where the two NNs are trained using the CMA_1 dataset is compared with recently proposed multiple linear regression (MLR) method by Zhang et al. [17]. The MLR method was shown superior to previous prediction methods. It was trained on a small set of 210 proteins (56 α-proteins, 75 β-proteins and 79 $\alpha\beta$-proteins) described in Ref. [20]. The regression-based method achieved 94.5% accuracy for the helix content and 94.9% accuracy for the strand content, on the test set of 143 proteins [17].

To follow the same procedure as described in Section 3.2, the test data was divided into three subsets: 83 protein sequences that have both helices and strands, see Fig. 4a, 16 protein sequences with no strands, see Fig. 4b, and 6 protein sequences with no helices, see Fig. 4c. The secondary structure information of the remaining 38 protein sequences, which are listed in Fig. 4d, could not be found in a recent release of the PDB database. We also note that the author of the MLR method was not able to provide us with the original test file. Therefore, a subset of 105 protein sequences was used to test our method.

The test was performed with the first NN used to predict content of the helix, and the second NN to predict content of the strand. The results for the helix content:

- For the 83 protein sequences having both helices and stands, the average square error is 0.008798 and the average accuracy is 90.62%.
- For the16 protein sequences with no strands, the average square error is 0.002965 and the average accuracy is 94.56%.
- For the 6 protein sequences with no helices, the average square error is 0.036108 and the average accuracy is 81%.

Therefore, the average accuracy for prediction of the helix content is not less than

$$0.9062 \times \frac{83}{105} + 0.9456 \times \frac{16}{105} + 0.81 \times \frac{6}{105} = 0.9067$$

The results for the strand content:

- For the 83 protein sequences having both helices and stands the average square error is 0.006859 and the average accuracy is 91.72%.
- For the16 protein sequences with no strands, the average square error is 0.021964 and the average accuracy is 85.18%.
- For the 6 protein sequences with no helices, the average square error is 0.005607 and the average accuracy is 92.51%.

1CP4, 1COT, 1BPQ, 1CSC, 1OMD, 2SCPA, 1ADL, 2RMCA, 2RSPB, 3RP2A, 1BCX, 1CELA, 1GCTA, 1ICM, 1IHSH, 1KNB, 1LEC, 1MPP, 1PPFE, 1RSY, 1SACA, 1SGC, 1SGT, 1SMRA, 1TTBA, 2APR, 2BBKH, 2HPEA, 8PAZ, 2PKAB, 1ACF, 1BSRA, 1MOLA, 1RIS, 3DFR, 1ALC, 1ALD, 1ALKA, 1AMP, 1APB, 1BAM, 1BSEA, 1BTC, 1CYO, 1DBS, 1DHIA, 1DRK, 1EDB, 1GMPA, 1GOB, 1GPB, 1HFC, 1HLEA, 1IAG, 1MRH, 1NHKL, 1OYA, 1PIPA, 1POH, 1SBP, 1TCA, 1THG, 1TPH1, 1TYS, 1WHTA, 1WHTB, 2ACU, 2AK3B, 2CHE, 2CHSA, 2CSTA, 2CUT, 2EBN, 2FCR, 2HPR, 2LAO, 2PKC, 3CPA, 3DNI, 5NLL, 4CLA, 6Q21A, 1UKZ
a) 83 protein sequences with both helices and strands
2ABK, 2CYP, 2C2C, 1YEA, 1ENJ, 1ARP, 1FLP, 1GDI, 1HBIA, 1LKI, 1LMB4, 1MNIA, 1TROA, 2LHB, 2MHR, 2SPCA
b) 16 protein sequences with no strands
1ACX, 1FNA, 2CLRB, 2FGF, 2MCM, 1BRSF
c) the 6 sequences with no helices
2BMHA, 1CDMA, 1HDSA, 1HDSB, 1REC, 2RAN, 1APTE, 1ARB, 2SIL, 1CBS, 1CONA, 1EAS, 1EPTB, 1HTRB, 1LOBA, 1MUA, 1SREA, 1THV, 2PLT, 1AAZA, 1DRF, 1FRD, 1AST, 1AYAA,

1DMB, 1FDD, 1GKY, 1GLG, 1NBAB, 1PBP, 1PHP, 1RVAA, 1TIB, 1XIB, 2EXO, 2IHL, 3PGA4, 3TGL
d) 38 protein sequences with no description in the PDB

Figure 4 Division of the test set for comparison between our and the MRL methods.

Table 3 Summary of the results obtained by the new method and the MLR method

Content prediction method Predicted content	This paper (%)	MLR method (%)
Helix	90.7	94.5
Strand	90.8	94.9

Therefore, the average accuracy for prediction of the strand content is not less than

$$0.9172 \times \frac{83}{105} + 0.8518 \times \frac{16}{105} + 0.9251 \times \frac{6}{105} = 0.9076$$

Comparison of results is shown in Table 3.

Although comparison of results shows that the MLR method has better results on the set of test proteins, we want to point out the following:

- The MLR method uses not only the composition vector but also structural class as its input. The structural class information contributes to the better accuracy, due to using different, customized regression models for each of the three classes, but it cannot be inferred directly from the primary sequence, and requires additional computations that are subject to errors. The structural class can be derived by applying computational prediction methods [29,40–42], or by assigning the class through sequence alignment and evolution relationship [20]. In case of prediction results for the MLR method, the structural class was predefined using knowledge about given proteins, and therefore it was 100% correct. We also note that recent results show that computational prediction of the structural class has limit of 60% accuracy [41], which would significantly decrease accuracy of content prediction. In contrast, our method is based solely on the composition moment vector, computed directly from the primary sequence.
- The test was performed on a well-prepared dataset concerning small family of globular proteins. The MLR method was trained using a very similar training set consisting of globular proteins, and therefore was better fitted to predict proteins from the test set. Our method was trained on a general set of 11,000 proteins, and although it achieved lower prediction rate, it provides a generalized solution for all protein types.

3.3.2. Comparison on the large protein set with other prediction methods

We use the CMB_1 dataset extracted from 25% PDB SELECT and describing 1707 proteins to perform the jackknife (leave-one-out) test. The results are compared with other commonly used secondary structure prediction methods based on two measures [17]:

$$e = \frac{\sum_{k=1}^{N} |F_k - D_K|}{N}, \quad \sigma = \sqrt{\frac{\sum_{k=1}^{N} (e - |F_k - D_K|)^2}{N-1}}$$

where e is the average error, σ the standard deviation, F_K the predicted helix or strand content, D_K the known content and N is the number of predicted proteins. We note that the lower the value of e the better the prediction quality.

The comparison uses additional dataset derived from 90% of PDB SELECT list released in October 2004. Analogically to the 25% PDB SELECT dataset, it contains 8595 proteins whose secondary structure is either published in PDB (release #103), or computed using DSSP procedure. After removing nucleotide sequences and proteins with errors in the primary sequence 8346 proteins are left and among them

Table 4 Comparison of secondary structure prediction results (MLR stands for multiple linear regression; AVD stands for analytic vector decomposition)

Reference	Prediction method	Dataset size	Structural class required	Jackknife $e(\sigma)$	
				Helix	Strand
[22]	MLR	104	No	0.129 (0.111)	0.120 (0.085)
[23]	AVD-1	262	No	0.145 (0.017)	0.120 (0.097)
	AVD-2	262	No	0.142 (0.115)	0.124 (0.105)
[17]	MLR	210	No	0.135 (0.103)	0.120 (0.097)
[25]	MLR-1	120	Yes	0.051 (0.055)	0.044 (0.052)
[24]	MLR-2	120	Yes	0.051 (0.053)	0.045 (0.053)
[17]	MLR	261	Yes	0.087 (0.067)	0.081 (0.065)
[17]	MLR-1	210	Yes	0.067 (0.060)	0.061 (0.057)
[17]	MLR-2	210	Yes	0.058 (0.057)	0.053 (0.053)
This paper	NN	1707	No	0.126 (0.096)	0.119 (0.099)
This paper	NN	395/514 (out of 6733)	No	0.115 (0.090)	0.103 (0.092)

6733 have both helixes and strands. The 6733 proteins were used in the jackknife fashion to predict helix content of randomly selected 395 proteins, and to predict strand content of randomly selected 514 proteins. For each jackknife test, 5000 epoch training of the NNs for strand and helix was done. The results include only a subset of 395 and 514 proteins, not the entire dataset, due to long training times; we note that for this experiment only over 10 days of 10 CPUs were needed for prediction of helix and another 10 days for prediction of strand content.

The results for our method and for the other methods are summarized in Table 4. We note that our results should be compared only with the methods that do not use structural class information (rows 1–3). All remaining methods assume knowledge of the structural class to derive specialized models for each of the classes to improve accuracy. At the same time computational prediction of the structural class has limit of 60% accuracy [41], which would significantly decrease their accuracy. We also note that method [22] is the only other method that uses NNs for the prediction; unfortunately, it did not show the jackknife test results but only the results on the training data. The results show that our method, which uses the composition moment vector only, achieves the highest prediction quality for both helix and strand content. Although the differences are not significant we note that the existing approaches use different prediction methods, and were tested on much smaller datasets. Therefore, it is impossible to evaluate how much our approach truly gains due to using the novel measure. We also note that better results were achieved for the 90% PDB SELECT when compared to the results for the 25% PDB SELECT. This could be attributed to higher homology of proteins in the 90% dataset. The prediction results obtained with composition moment vector are shown in Fig. 5. The plots show the actual content and predicted content on the *x*- and *y*-axes, respectively.

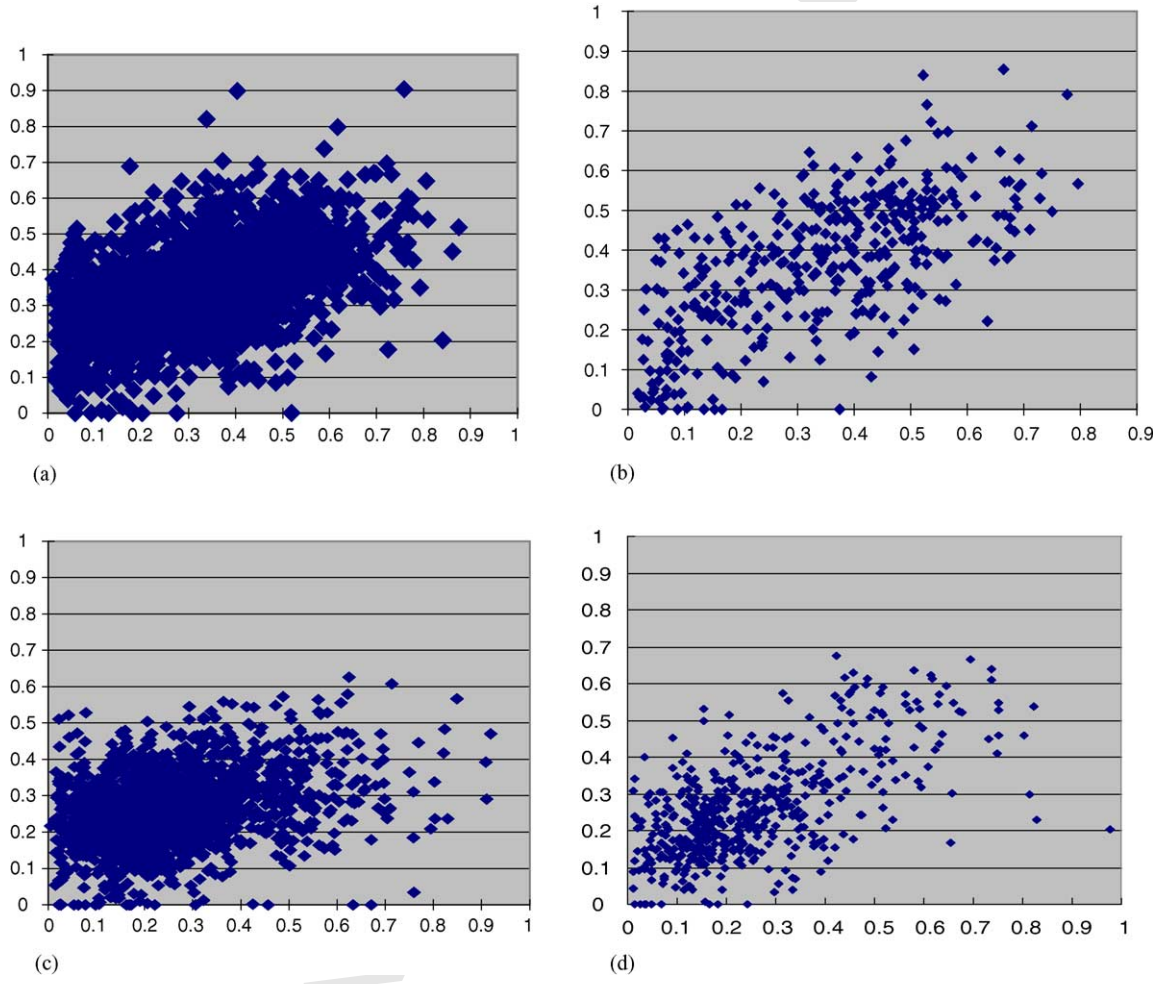

Figure 5 Results of jackknife test for strand and helix prediction for our method using set of 1707 proteins (a and c) and set of 395/514 proteins (b and d).

4. Conclusions and future work

The paper describes a novel approach for prediction of secondary structure content of proteins from their primary AA sequences. A new measure, called composition moment vector, to characterize primary sequences was proposed. The composition moment vector considers both the count and the position of an AA in the primary sequence. That results in achieving functional relation between the vector and the secondary structure content for a general set of proteins. In contrast, currently used composition vector considers only counts of AAs, and does not provide a functional relation with the secondary structure content for a general set of proteins. This impairs its usefulness for highly accurate prediction of the secondary structure content.

The new measure and neural networks were used to predict the secondary structure content. The method was tested on a large set of over 11,000 proteins extracted from the PDB and achieved high prediction accuracy. The results were compared with prediction that uses only the composition vector and confirmed superiority of our new measure. The new approach was also compared with other state-of-the-art prediction method on a small set of globular proteins. Since some other methods assume knowledge of structural class information (which can be predicted with only 60% accuracy limit), and were trained and tested using sets of similar globular proteins, they achieved slightly better predictive accuracy. However, comparison on larger protein sets show superiority of our method when compared with other methods that do not use the structural class information.

The new composition moment vector measure can be used not only to predict secondary structure content, but also to find protein structural class, secondary structure, and perform protein function prediction. It can also be used for analysis of relation between primary and secondary structure, which incorporates information about position of AA in the primary sequence. The introduced here measure could be used to perform computational analysis of primary protein structure that gives new insights into mechanisms of protein folding and function.

We note that our approach is limited by two main factors. First, prediction of content of strictly non-homologous proteins may results in lower accuracy of the current, NN-based, architecture because of high dissimilarity between proteins in the training and testing data sets. Second, computational complexity of computing high order moment vectors prohibits from using them for problems involving large number of long protein sequences. The solution is to use low order, including zero, first and second order moments, as shown in this paper, to perform the prediction.

Future work will involve prediction of helix and strand content for invariant fragments of proteins, i.e. fragments of proteins that have the same primary and secondary structure between different proteins. The mapping between the composition moment vector of invariant fragments and their secondary structure content is suitable for the NN-based prediction. Two possible issues need to be looked at. First, the invariant fragment must be correctly identified, for example along the lines proposed in Ref. [43]. They have shown that all primary sequences in the PDB are a covering set of all smaller proteins in 3D structures. Therefore, smaller protein sequences might be used to search for invariants in larger proteins. Second, the number of invariant fragments might be large in which case, a different NN might be used, such as a radial basis function that is characterized by much faster training time.

Acknowledgements

The authors would express their gratitude to their funding agencies. Ruan's work was supported in part by Liuhui Center for Applied Mathematics and the China—Canada exchange program administered by MITACS. Kurgan's work was supported in part by the Natural Sciences and Engineering Research Council of Canada (grant number G121210953).

Appendix A

An induction-based proof is given to show that two proteins are the same if and only if they have the same moment matrix defined by Eqs. (1) and (2) (Section 2.4.2).

A.1. Proof

Based on Eqs. (1) and (2) it is trivial to see that if two proteins are the same, i.e. they have the same primary sequence, their corresponding moment matrices are the same. Therefore, we need only to show that if two moment matrices are the same that implies that two primary sequences are the same.

Let's define two moment matrices, A_{K+1} and A'_{K+1}, which correspond to two primary sequences, as:

$$A_{K+1} = \begin{bmatrix} x_1 & x_2 & \cdots & x_{20} \\ x_1^{(1)} & x_2^{(1)} & \cdots & x_{20}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(K)} & x_2^{(K)} & \cdots & x_{20}^{(K)} \end{bmatrix} \quad \text{and}$$

$$A'_{K+1} = \begin{bmatrix} y_1 & y_2 & \cdots & y_{20} \\ y_1^{(1)} & y_2^{(1)} & \cdots & y_{20}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(K')} & y_2^{(K')} & \cdots & y_{20}^{(K')} \end{bmatrix}$$

We also note that n_{ij} and n'_{ij} are the jth position of the ith AA, K_i and K'_i the total number of the ith AA in the two sequences, respectively, K the maximal value among all K_i and K' is the maximal value among all K'_i, for $i = 1, 2, \ldots 20$.

If $A_{K+1} = A'_{K+1}$, then $K = K'$ and $K_i = K'_i$, for $i = 1, 2, \ldots, 20$.

For $K = 1$, which means that each AA appears at most once in each of the primary sequences, the A_{K+1} and A'_{K+1} matrices degenerate into

$$A_2 = \begin{bmatrix} x_1 & x_2 & \cdots & x_{20} \\ x_1^{(1)} & x_2^{(1)} & \cdots & x_{20}^{(1)} \end{bmatrix} \quad \text{and}$$

$$A'_2 = \begin{bmatrix} y_1 & y_2 & \cdots & y_{20} \\ y_1^{(1)} & y_2^{(1)} & \cdots & y_{20}^{(1)} \end{bmatrix}$$

If $A_2 = A'_2$, then using the definition of the first order moment vector (Eq. (1), Section 2.4.2) it can be shown that $n_{ij} = n'_{ij}$, for all i and j, which means that the corresponding two primary sequences are identical.

For $K = 2$, which means that each AA appears at most twice in each of the primary sequences, the A_{K+1} and A'_{K+1} matrices degenerate into

$$A_3 = \begin{bmatrix} x_1 & x_2 & \cdots & x_{20} \\ x_1^{(1)} & x_2^{(1)} & \cdots & x_{20}^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_{20}^{(2)} \end{bmatrix} \quad \text{and}$$

$$A'_3 = \begin{bmatrix} y_1 & y_2 & \cdots & y_{20} \\ y_1^{(1)} & y_2^{(1)} & \cdots & y_{20}^{(1)} \\ y_1^{(2)} & y_2^{(2)} & \cdots & y_{20}^{(2)} \end{bmatrix}$$

If $A_3 = A'_3$, then using the definition of the first and second order moment vectors (Eq. (1), Section 2.4.2) it can be shown that $n_{i1} + n_{i2} = n'_{i1} + n'_{i2}$ and $n_{i1}^2 + n_{i2}^2 = (n'_{i1})^2 + (n'_{i2})^2$ for all $K_i = K = 2$. This implies that $n_{i,1}n_{i,2} = n'_{i,1}n'_{i,2}$. Applying Viete's formula, $n_{i,1}$, $n_{i,2}$ can be represented as two positive roots of some polynomial $f(x) = x^2 - ax + b$. Similarly, $n'_{i,1}$, $n'_{i,2}$ can be represented as the two positive roots. This implies that $n_{i,1} = n_{i,2}$ and $n'_{i,1} = n'_{i,2}$.

For all $K_i < 2$, i.e. $K = 1$, $n_{ij} = n'_{ij}$ based on the argument for $K = 1$.

Therefore, the corresponding two primary sequences are identical.

For $K = 3$, which means that each AA appears at most three times in each of the primary sequences, the A_{K+1} and A'_{K+1} matrices degenerate into

$$A_4 = \begin{bmatrix} x_1 & x_2 & \cdots & x_{20} \\ x_1^{(1)} & x_2^{(1)} & \cdots & x_{20}^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_{20}^{(2)} \\ x_1^{(3)} & x_2^{(3)} & \cdots & x_{20}^{(3)} \end{bmatrix} \quad \text{and}$$

$$A'_4 = \begin{bmatrix} y_1 & y_2 & \cdots & y_{20} \\ y_1^{(1)} & y_2^{(1)} & \cdots & y_{20}^{(1)} \\ y_1^{(2)} & y_2^{(2)} & \cdots & y_{20}^{(2)} \\ y_1^{(3)} & y_2^{(3)} & \cdots & y_{20}^{(3)} \end{bmatrix}$$

If $A_4 = A'_4$, then using the definition of the first, second and third order moment vectors it can be shown that $n_{i1} + n_{i2} + n_{i3} = n'_{i1} + n'_{i2} + n'_{i3}$, $n_{i1}^2 + n_{i2}^2 + n_{i3}^2 = (n'_{i1})^2 + (n'_{i2})^2 + (n'_{i3})^2$ and $n_{i1}^3 + n_{i2}^3 + n_{i3}^3 = (n'_{i1})^3 + (n'_{i2})^3 + (n'_{i3})^3$ for all $K_i = K = 3$.

For a polynomial $f(x) = x^3 - ax^2 + bx - c$ we have $n_{i1} + n_{i2} + n_{i3} = n'_{i1} + n'_{i2} + n'_{i3} = a$. Using equality $(n_{i1} + n_{i2} + n_{i3})^2 - (n_{i1}^2 + n_{i2}^2 + n_{i3}^2) = (n'_{i1} + n'_{i2} + n'_{i3})^2 - [(n'_{i1})^2 + (n'_{i2})^2 + (n'_{i3})^2]$ we have $n_{i1}n_{i2} + n_{i3}n_{i1} + n_{i2}n_{i3} = n'_{i1}n'_{i2} + n'_{i3}n'_{i1} + n'_{i2}n'_{i3} = b$. Also, using equality $(n_{i1}n_{i2} + n_{i3}n_{i1} + n_{i2}n_{i3})(n_{i1} + n_{i2} + n_{i3}) = (n'_{i1}n'_{i2} + n'_{i3}n'_{i1} + n'_{i2}n'_{i3})(n'_{i1} + n'_{i2} + n'_{i3})$ we have $n_{i1}n_{i2}n_{i3} = n'_{i1}n'_{i2}n'_{i3} = c$.

Applying Viete's formula, we find that both $n_{i,1}$, $n_{i,2}$, $n_{i,3}$ and $n'_{i,1}$, $n'_{i,2}$, $n'_{i,3}$ are the three positive roots of the same polynomial $f(x) = x^3 - ax^2 + bx - c$. Therefore, $n_{ij} = n'_{ij}$, for $j = 1$, 2 and 3.

For all $K_i < 3$, i.e. $K = 1$ and $K = 2$, $n_{ij} = n'_{ij}$ based on the argument for $K = 1$ and 2.

Therefore, the corresponding two primary sequences are identical. Identical argument can be extended for all finite K. Therefore, by induction we have shown that if the two moment matrices are the same, the corresponding two primary sequences are identical. □

Appendix B

The NNs used in experiments described in this paper use a non-standard transfer function. The multi-layer perceptron NN traditionally use the s-

form transfer function $f(x) = \frac{e^x}{1+e^x}$. This paper uses another transfer function, denoted as $g(x)$, which is derived by integration of the uniform distribution $p(x) \sim U(-a, a)$. We note that $p(x) \to \delta(x)$, where $\delta(x)$ is the Dirac function, as $a \to 0$. We define $g(x) = \int_{-\infty}^{x} p(t)\,dt$, and based on experiments set $a = 6$.

Using $g(x)$ results in lowering computational complexity of NN training, and increases prediction accuracy, when compared with using $f(x)$ transfer function.

References

[1] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucl Acids Res 2000;28:235–42.
[2] Ganapathiraju MK, Klein-Seetharaman J, Balakrishnan N, Reddy R. Characterization of protein secondary structure. IEEE Signal Process Mag 2004;(May):78–87.
[3] Dwyer DS. Electronic properties of the amino acids side chains contribute to the structural preferences in protein folding. J Biomol Struct Dyn 2001;18(6):881–92.
[4] Rost B. Review: protein secondary structure prediction continues to rise. J Struct Biol 2001;134(2–3):204–18.
[5] Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol 1978;47:45–148.
[6] Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120(1):97–120.
[7] Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;19(1):55–72.
[8] Gibrat JF, Garnier J, Robson B. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J Mol Biol 1987;198(3):425–43.
[9] Schmidler SC, Liu JS, Brutlag DL. Bayesian segmentation of protein secondary structure. J Comput Biol 2000;7(1–2):233–48.
[10] Andrew CD, Penel S, Jones GR, Doig AJ. Stabilizing nonpolar/polar side-chain interactions in the alpha-helix. Proteins 2001;45(4):449–55.
[11] Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261:552–7.
[12] Zhang N. Markov models for classification of protein helices. Biochemistry 2001;218.
[13] Thomas A, Meurisse R, Charloteaux B, Brasseur R. Aromatic side-chain interactions in proteins. I. Main structural features. Proteins 2002;48(4):628–34.
[14] Thomas A, Meurisse R, Brasseur R. Aromatic side-chain interactions in proteins. II. Near- and far-sequence Phe-X pairs. Proteins 2002;48(4):635–44.
[15] Eisenhaber F, Person B, Argos P. Protein structure prediction: recognition of primary, secondary, and tetriary structural features from amino acid sequence. Crit Rev Biochem Mol Biol 1995;30:1–94.
[16] Rost B, Sander C. Third generation prediction of secondary structure. In: Webstar D, editor. Protein Structure Prediction: Methods and Protocols. Clifton, NJ: Humana Press; 2000. p. 71–95.
[17] Zhang ZD, Sun ZR, Zhang CT. A new approach to predict the helix/strand content of globular proteins. J Theor Biol 2001;208:65–78.
[18] Sreerama N, Woody RW. Protein secondary structure from circular dichroism spectroscopy. J Mol Biol 1994;242:497–507.
[19] Bussian BM, Sender C. How to determine protein secondary structure in solution by raman spectroscopy: practical guide and test case DNsae I. Biochemistry 1989;28:4271–7.
[20] Mitchie AD, Orengo CA, Thornton JM. Analysis of domain structural class using an automated class assignment protocol. J Mol Biol 1996;262:168–85.
[21] Krigbaum WR, Knutton SP. Prediction of the amount of secondary structure in a globular protein from its amino acid composition. Proc Natl Acad Sci 1973;70:2809–13.
[22] Muskal SM, Kim S-H. Predicting protein secondary structure content: a tandem neural network approach. J Mol Biol 1992;225:713–27.
[23] Eisenhaber F, Imperiale F, Argos P, Frommel C. Prediction of secondary structural contents of proteins from their amino acid composition alone. I. New analytic vector decomposition methods. Proteins 1996;25(2):157–68.
[24] Zhang CT, Zhang Z, He Z. Prediction of the secondary structure contents of globular proteins based on three structural classes. J Protein Chem 1998;17:261–72.
[25] Zhang CT, Zhang Z, He Z. Prediction of the secondary structure of globular proteins based on structural classes. J Protein Chem 1996;15:775–86.
[26] Zhang CT, Lin ZS, Zhang Z, Yan M. Prediction of helix/strand content of globular proteins based on their primary sequences. Protein Eng 1998;11(11):971–9.
[27] Cedano J, Aloy P, Peres-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. J Mol Biol 1997;266:594–600.
[28] Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 1995;21:319–44.
[29] Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
[30] Chou PY. Prediction of protein structural classes from amino acid composition. In: Fasman GD, editor. Prediction of Protein Structures and the Principles of Protein Conformation. New York: Plenum Press; 1989. p. 549–86.
[31] Eisenhaber F, Imperiale F, Argos P, Frommel C. Prediction of secondary structural contents of proteins from their amino acid composition alone. II. The paradox with secondary structural class. Proteins 1996;25(2):169–79.
[32] Klein P, Delist C. Prediction of protein structural classes from amino acids sequence. Biopolymers 1986;25:1659–72.
[33] Kneller DG, Cohen FE, Langridge R. Improvement in protein secondary structure prediction by an enhanced neural network. J Mol Biol 1990;214:171–82.
[34] Sheridan RP, Dixon JS, Venkataraghavan R. Generating plausible protein folds by secondary structure similarity. Int J Pept Protein Res 1985;25:132–43.
[35] Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci 1992;1:409–17.
[36] Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522.
[37] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–637.

[38] Hornik K, Stinchcombe M, White H. MLP's are universal approximators. Neural Netw 1989;2:359–66.
[39] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a natural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–40.
[40] Chou PY, Maggiora GM. Domain structural class prediction. Protein Eng 1998;11:523–38.
[41] Wang Z, Yuan Z. How good is prediction of protein structural class by the component coupled methods. Proteins 2000;38:165–75.
[42] Zhou GP. An intriguing controversy over protein structural class prediction. J Protein Chem 1998;17:729–38.
[43] Kihara D, Skolnick J. The PDB is a covering set of small protein structures. J Mol Biol 2003;223:793–802.