Volume 248
Issue 2
21 September 2007
ISSN 0022-5193
ELSEVIER
Journal of Theoretical Biology

Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Journal of Theoretical Biology 248 (2007) 354–366
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Novel scales based on hydrophobicity indices for secondary protein structure

Lukasz A. Kurgan[a,*], Wojciech Stach[b], Jishou Ruan[c]

[a]*Electrical and Computer Engineering Department, University of Alberta, Edmonton, Canada, T6G 2V4*
[b]*Electrical and Computer Engineering Department, University of Alberta, Edmonton, Canada, T6G 2V4*
[c]*College of Mathematics and LPMC, Nankai University, Tianjin, PRC, 300071*

Received 22 February 2007; received in revised form 4 April 2007; accepted 14 May 2007
Available online 18 May 2007

Abstract

This paper is concerned with a branch of computational biology related to protein prediction and analysis of secondary structure of proteins. Although traditional methods use a simple amino acid composition to predict the secondary structure content, hydrophobicity has been recently found to improve the results in this and several related prediction tasks. To this end, we propose and analyze advantages of two new hydrophobicity index-based scales that incorporate information about long-range interactions along the protein sequence and contrast them with currently used raw hydrophobic index values. We also compare three leading hydrophobicity indices, i.e., Eisenberg's, Fauchere–Pliska's, and Cid's, using the proposed scales. The analysis is performed using fuzzy cognitive maps that quantify the strength of relation between the hydrophobicity scales/indices and the protein content values. A set of empirical tests that involve generation of fuzzy cognitive map models for a set of 200 low homology proteins have been performed. The results show that the secondary structure content along the protein sequence is characterized by about 2.5 times stronger relation with the two proposed hydrophobicity scales when compared with the currently used raw index values. The new scales exhibit stronger relation irrespective of the applied hydrobhobicity indices. Analysis of different scales shows superiority of the Eisenberg's hydrophobicity index, when used with the new scales. In contrast, the Fauchere–Pliska's index is found to perform better when compared with the two other indices when using raw hydrophobic index values that disregard the long-range interactions.

Keywords: Secondary protein structure; Hydrophobicity; Fuzzy cognitive maps

1. Introduction

One of the active research areas in computational biology is prediction and analysis of protein structure. Proteins are characterized by three structural levels: primary sequence of amino acid (AA), and secondary and tertiary structure. The secondary structures are usually grouped into four categories: α-helices, β-strands, tight-turns (Chou, 2000a), and coils. While the primary sequences are currently publicly known for over 3 millions of proteins, the secondary and tertiary structure is known for relatively small number of proteins, i.e., the Protein Data Bank (PDB) currently contains about 30 thousands proteins (Berman et al., 2000). At the same time, research in protein function and interactions requires knowledge of the tertiary structure. Experimental methods for discovery of tertiary (secondary) structure are relatively time consuming, labor expensive, and cannot be applied to some proteins. As a result, computational methods gain momentum (Jones, 2000). Computational methods are faster and cheaper and thus they can potentially close the existing gap between the number of known primary and unknown higher protein structures. Computational methods for secondary structure prediction are under development for over 3 decades, and although their accuracy is getting better, i.e., currently it reaches about 80% (Birzele and Kramer, 2006; Rost, 2001), much work still needs to be done. This is especially true when it comes to performing

[*]Corresponding author. Tel.: +780 492 5488; fax: +780 492 1811.
E-mail addresses: lkurgan@ece.ualberta.ca (L.A. Kurgan), wstach@ece.ualberta.ca (W. Stach), jruan@phys.ualberta.ca (J. Ruan).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.05.017

predictions for low homology proteins, i.e., the corresponding accuracy is about 65–68% (Lin et al., 2005). A-priori knowledge of protein content is an important piece of information for prediction of the secondary structure. While the secondary structure prediction aims to predict one of the three categories for each AA in the primary sequence, the secondary structure content prediction methods predict amounts of helix and strand structures in the protein sequence (the amount of coils and turns is obtained as a complement of the remaining two secondary structures). Meanwhile, significant efforts have been made to predict protein structural class (see Chou, 2000b; Luo et al., 2002; Chou and Cai, 2004a; Feng et al., 2005; Shen et al., 2005; Niu et al., 2006; Kurgan and Homaeian, 2006; Kedarisetti et al., 2006 as well as the references cited in a recent review Chou, 2005a) and to predict the protein fold patterns (Ding and Dubchak, 2001; Shen and Chou, 2006a).

The existing protein content prediction methods use a set of standard measures, which include AA composition, pair coupled composition and hydrophobicity, to perform prediction. Hydrophobicity is also used in a number of other protein prediction tasks, such as prediction of structural class, subcellular location, membrane protein type, etc. Hydrophobicity is usually expressed by an index that quantifies this property for each AA; these raw values are used by prediction methods. In contrast, this paper proposes hydrophobic scales that use the raw index values to compute new hydrophobicity values (scales). The main difference between the raw values and a scale is that the index value is specific to a given AA, i.e., there are 20 values for 20 AA, while the scale gives a potentially different hydrophobicity value for each AA along the protein sequence. In this paper we propose two new hydrophobic scales, investigate several hydrophobicity indices, and compare the quality of the two new scales with the currently used raw values. In contrast to the currently used raw values, the proposed scales incorporate information about long range (with respect to the protein sequence) interactions. We use fuzzy cognitive maps (FCMs) and their genetic algorithm-based learning method to perform comparisons. The FCMs are used to learn and quantify the relation between the two new hydrophobicity scales and the currently used raw index values, and the secondary structure content along protein sequences. The motivation for using FCMs comes from their ability to capture and quantify relations between multiple variables, which is critical for the considered biological system, not just relations between selected pairs of variables as in case of the commonly applied correlation-based analysis. Based on comprehensive experiments that use a set of 200 low homology proteins we show that the two new hydrophobicity scales are characterized by a much stronger functional relation with the secondary structure content than the currently used raw index values. We also show advantages and drawbacks of the considered three hydrophobicity indices with respect to the scales. Although this paper does not propose a new prediction method, it provides invaluable information that can be used to design new prediction methods, and which can be extended to the related protein prediction tasks.

2. Background and methods

2.1. Protein prediction methods and hydrophobicity

The secondary protein structure is defined using Dictionary of Secondary Structures of Proteins (DSSP) (Kabsch and Sander, 1983). DSSP annotates each AA as belonging to one of eight secondary structure types: H (alpha-helix), G (3-helix or 310 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). Typically they are reduced to three groups: helix (H that includes H and G), strand (E that includes E and B), and coil (C that includes remaining types) (Moult et al., 1997). Recent years show increasing interest in the computational prediction of the secondary structure content. This task is concerned with prediction of percentage amount of the α-helix and β-strand content in a given protein sequence. It concerns three granularities of the secondary structure: 3 states (structure types), 8 states, and 10 states in which the extended strand is divided into subtypes including strand, parallel strand and anti-parallel strand. The existing prediction methods first convert the primary protein sequence into feature space representation and use it to predict the content values. Most methods use composition vector-based representations and apply regression to perform prediction (Chou, 1999; Zhang et al., 1998; Liu and Chou, 1999; Lin and Pan, 2001; Pilizota et al., 2004; Lee et al., 2006; Homaeian et al., 2007). Table 1 shows chronological comparison of recent content prediction methods, which includes information about the prediction algorithms, number of predicted structure types, and sequence representations.

Another popular sequence representation is based on hydrophobicity. The usefulness of this representation is supported by its extensive prior use and desirable properties. In order to incorporate the sequence-order information into a non-sequential discrete model for statistical/machine learning prediction, instead of the classical composition vector Chou proposed the pseudo AA composition to represent a protein sequence (Chou, 2001). The concept of the Chou's pseudo AA composition has stimulated a series of follow-up studies in which different types of pseudo AA composition have been developed to improve the prediction quality of protein attributes, such as protein structural class (Shen and Chou, 2005a; Chen et al., 2006a; Chen et al., 2006b; Xiao et al., 2006a; Lin and Li, 2007a), membrane protein type (Guo, 2002; Wang et al., 2004; Liu et al., 2005a; Liu et al., 2005b; Chou and Cai, 2005; Wang et al., 2005; Shen and Chou, 2005a; Shen et al., 2006; Shen and Chou, 2006b; Wang et al., 2006), enzyme family class (Chou, 2005b; Cai et al.,

Table 1
Comparison of recent secondary structure content prediction methods; MLR stands for multiple linear regression and NN stands for neural network

Reference	Prediction algorithm	♯ predicted types	Sequence representation
(Eisenhaber et al., 1996)	vector decomposition	3	Composition vector
(Zhang et al., 1998)	MLR	3	Composition vector, autocorrelation based on hydrophobicity (Fauchere–Pliska index)
(Chou, 1999)	MLR	8	Pair coupled composition vector
(Liu and Chou, 1999)	MLR	10	Pair coupled composition vector
(Zhang et al., 2001)	MLR	3	Composition vector, autocorrelation based on hydrophobicity (Fauchere–Pliska index)
(Lin and Pan, 2001)	MLR	3	Composition vector, autocorrelation based on hydrophobicity (Fauchere–Pliska index), side chain mass interaction functions
(Cai et al., 2002a)	NN	10	Pair coupled composition vector
(Cai et al., 2002b)	NN	8	Pair coupled composition vector
(Pilizota et al., 2004)	MLR	3	Composition vector
(Ruan et al., 2005)	NN	3	Composition moment vector
(Lee et al., 2006)	MLR, NN, SVR	8	PSI-BLAST-based composition vector
(Homaeian et al., 2007)	MLR	3	Composition and composition moment vectors, autocorrelation based on hydrophobicity (Fauchere–Pliska's and Eisenberg's indices), property groups

2005; Cai and Chou, 2005), GPCR type (Chou and Elrod, 2002; Chou, 2005c; Guo et al., 2006; Wen et al., 2006), protein quaternary structure (Chou and Cai, 2003a; Guo et al., 2006; Wen et al., 2006; Zhang et al., 2006), protein subcellular localization (Pan et al., 2003; Chou and Cai, 2003b, 2004b; Xiao et al., 2005a, b; Shen and Chou, 2005b; Gao et al., 2005; Xiao et al., 2006b; Shen et al., 2007; Shi et al., 2007; Chou and Shen, 2006a–d, 2007; Shen and Chou, 2007a; Shen and Chou, 2007b, c), as well as other protein attributes (Chou and Cai, 2006; Du and Li, 2006; Mondal et al., 2006; Zhou and Cai, 2006; Lin and Li, 2007b). In most of these approaches, the hydrophobicity or its combination with other AA properties have been used to formulate different types of pseudo AA composition, fully indicating the importance of the hydrophobicity scale to the protein science.

Hydrophobicity is not only one of the major structural forces, but is also able to show periodicity of the secondary structure (Cornette et al., 1987). In an aqueous environment, hydrophobic molecules, including the hydrophobic AA side chains, are forced together to minimize the disruptive effect on the hydrogen-bonded water molecules network. Thus, distribution of hydrophilic and hydrophobic AA side chains has a significant impact on the protein structure. The hydrophobic side chains cluster in the protein interior, while hydrophilic side chains arrange themselves near the protein outside where they can form hydrogen bonds with water and with other polar molecules. The hydrophobic AAs are usually hydrogen-bonded to other hydrophobic AAs or to the polypeptide protein backbone.

Zhang and colleagues studied 14 hydrophobicity indices with respect to the content prediction and concluded that the best results are obtained by using Fauchere–Pliska's index (Zhang et al., 2001). The 14 indices were proposed by Fauchere and Pliska (1983), Wold et al. (1987), Wertz and Scheraga (1978), Sweet and Eisenberg (1983), Ponnuswamy et al. (1980), Parker et al. (1986), Nishikawa and Ooi (1980), Miyazawa and Jernigan (1985), Cid et al. (1992), Bull and Breese (1974), and Biou et al. (1988), respectively. Majority of existing hydrophobicity indices are stored in the AA index database (Kawashima et al., 1999). The Fauchere–Pliska's index was used in several content prediction methods (see Table 1) and in structural class prediction (Kurgan et al., 2006). A substantial drawback of the Zhang's contribution is that the indices were tested only with respect to prediction using a multiple linear regression method, and not with respect of their overall correlation with the content. A few years earlier, in 1998, Juretic and Lucin showed that among 87 different AA indices that they examined, the Eisenberg's hydrophobicity index is the most suitable to identify periodicity of the secondary structure (Juretic and Lucin, 1998). Although their results have shown the benefits associated with this index, it received little attention among the researchers in the protein structure prediction community, i.e., it was used to predict structural classes (Kedarisetti et al., 2006; Kurgan and Homaeian, 2006) and in one content prediction method (Homaeian et al., 2007). Finally, a novel hydrophobicity index, so-called Heuristic Molecular Lipophilicity Potential, was recently introduced (Du et al., 2006). This index, which is based on quantum mechanical and statistical mechanical principles introduced in Du et al. (2005), quantifies lipophilic and hydrophilic properties of AA side chains.

2.2. *Proposed hydrophobicity scales*

The prediction methods described in Section 2.1 use only raw hydrophobic index values, i.e. hydrophobicity values of corresponding AAs. To this end, we propose two novel hydrophobic scales that transform the raw values into a

Table 2
The Eisenberg's, Fauchere–Pliska's, and Cid's hydrophobicity indices

Index AA	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Eisenberg	0.62	0.29	−0.90	−0.74	1.19	0.48	−0.40	1.38	−1.50	1.06	0.64	−0.78	0.12	−0.85	−2.53	−0.18	−0.05	1.08	0.81	0.26
Fauchere–Pliska	0.42	1.34	−1.05	−0.87	2.44	0.00	0.18	2.46	−1.35	2.32	1.68	−0.82	0.98	−0.30	−1.37	−0.05	0.35	1.66	3.07	1.31
Cid	0.17	1.24	−1.07	−1.19	1.29	−0.57	−0.25	2.06	−0.62	0.96	0.60	−0.90	−0.21	−1.20	−0.70	−0.83	−0.62	1.21	1.51	0.66

new representation that provides better correlation with the secondary structure content. Similarly to the raw values, the new scales produce a single value for each residue along the sequence. At the same time, raw values are identical for the same AAs, while the scale values differ for each AA along the protein sequence. We investigate the added-value of the new scales by applying them with the most commonly used Fauchere–Pliska's index and the Eisenberg's index. We also use the Cid's index, which is the most recent index studied by Zhang and colleagues (Zhang et al., 2001), in order to contrast and compare the other two indices. Table 2 shows the three selected indices.

The proposed scales are based on aggregation of raw index values along the sequence, i.e. for the tth residue they sum the hydrophobicity index values for the first t residues. The accumulation of the first t index values to compute the value for the tth residue can be seen as injection of aggregated (global) information stored in the first t residues. This allows for representing long-range trends in hydrophobicity, which may help to represent the existing long-range interactions in the secondary structures. For instance, β-sheets are often formed by segments that are far apart in the sequence. The proposed aggregation procedure is motivated by the translation process that occurs in the ribosome. During that process, the peptide synthesis occurs from N-terminus (beginning of the sequence) to C-terminus (end of the sequence), i.e., incoming AAs are added to the growing C-terminus, which motivates summation starting from the N-terminus. At the same time, we note that the residues behind the residue at position t also have impact on the folding.

The first proposed scale is based on the above described summation, while the second additionally incorporates information about local interactions, i.e., it computes 3-point moving average of the summed values. In contrast, the raw hydrophobicity index values used in the recent contributions does not consider long-range interactions. The index values and the two proposed scales are defined as

- Raw hydrophobicity index (H_x) $= \{X_t\}$
- Cumulative hydrophobicity scale (CH_x) $= \left\{\sum_{i=1}^{t} X_i\right\}$
- 3-points moving average of cumulative hydrophobicity scale ($3\mathrm{CH}_x$) $= \left\{\sum_{i=t-2}^{t}\sum_{j=1}^{i} X_i\right\}$

where $t = 1, 2, \ldots, N$, N is the protein sequence length; X_t is the corresponding hydrophobicity index value for tth AA in the sequence; subscript x denotes the corresponding index which will be used with the scale, i.e., E for the Eisenberg's index, F for the Fauchere–Pliska's index, and C for the Cid's index.

Additionally, for each primary protein sequence the content values have been computed based on known secondary sequence:

- *%helix* (%h)—$|H|_t/t$, where $|H|_t$ is the number of residues in helical conformation in the substring of protein sequence from the position 1 to t;
- *%strand* (%s)—$|E|_t/t$, where $|E|_t$ is the number of residues in strand conformation in the substring of protein sequence from the position 1 to t.

The two proposed scales, the raw index values, and the above two secondary protein content variables constitute a real-valued function with respect to the primary protein sequences.

2.3. Fuzzy cognitive maps

The analysis of the proposed hydrophobicity scales and the raw index values is performed with the use of fuzzy cognitive maps (FCMs). FCMs, which were introduced by Kosko (1986) as an extension to cognitive maps (Axelrod, 1976), are convenient tool for modeling and simulation of dynamic systems. They describe a given system as a collection of concepts that are connected by cause-effect relations, which is depicted with a graph. The graph's nodes represent concepts and the causal relations between them are depicted by directed edges. Each edge is associated with a weight value that reflects strength of corresponding relation, i.e., it determines the degree of considered causal relation between the two concepts. This value is usually normalized to the interval [−1,1]. Positive values reflect promoting effect, whereas negative describe inhibitory effect. The value of −1 represents full negative, +1 full positive, and 0 denotes no causal effect. Other values correspond to the intermediate levels of causal effect. The graph can be equivalently expressed by a square matrix, called *connection matrix*, which stores all weight values for edges between corresponding concepts represented by corresponding rows and columns. The FCM model is simulated by calculating its state during a number of consecutive iterations. This state is represented by a *state vector,* which determines values of each node. The FCM iteratively updates the state of the system, i.e., value of a node is calculated in each iteration from values in the preceding iteration of nodes that exert influence on the given node (nodes that are connected to the given node)

$$C_j(t+1) = f\left(\sum_{i=1}^{N} e_{ij} C_i(t)\right),$$

where $C_i(t)$ is the value of ith node at the tth iteration, e_{ij} is the edge weight (relationship strength) from the concept C_i to the concept C_j, t is the iteration number (time point), N is the number of concepts, and f is the transformation (transfer) function.

The transformation function is used to reduce unbounded weighted sum to a certain range, which is usually set to [0,1]. This allows for comparisons between nodes, which can be defined as active (value of 1), inactive (value of 0), or active to a certain degree (value between 0 and 1). Three most commonly used transformation functions are bivalent, trivalent, and logistic. There are several advantages of modeling dynamic systems using FCM models. FCMs are very simple and intuitive to understand. They are also flexible in terms of system design and applications since they have comprehensible structure and operation, are adaptable to a given domain, and capable of abstract representation (Koulouriotis et al., 2003).

Development of FCM models almost always relies on human knowledge (Aguilar, 2005). As a result, the developed models strongly depend on experts beliefs, which imply subjectivity of the model and problems with unbiased assessment of its accuracy. The main difficulty is to accurately establish weights (strength) of the defined relations. A novel method for learning the weights based on real-coded genetic algorithm (RCGA) (Herrera et al., 1998) was recently proposed (Stach et al., 2004, 2005). The method allows for automated generation of FCM models from data based on a genetic algorithm based optimization. Considering the fact that FCM model can be fully described by its connection matrix, the learning goal is to find N^*N parameters. The RCGA algorithm exploits input data to find the parameters. Input data are a sequence of states described by state vectors at a particular time (iteration). In this paper, the input data are the values of the two considered hydrophobicity scales, the raw index values, and the protein content along the primary protein sequence (each residue corresponds to a time-point). The learning objective is to generate the same state vector sequence for the same initial state vector defined by the input data. At the same time, the learned matrix generalizes the inter-relations between concept nodes, which are inferred from the input data. The FCM model is suitable to perform simulation for different initial state vectors, and quantify the degree and type of cause–effect relations between the concepts. The learned parameters quantify the strength of the relation between the scales, raw index and content, i.e., higher absolute values correspond to stronger relations while values close to zero indicate weak relations. Although relations between all pairs of concepts (scales, raw index, content) are computed (see Appendix A), the experimental section concentrates on analysis of relations between the proposed scales/raw index values and the secondary structure content.

2.4. Goals

Our goals are four-fold:

GOAL 1: To quantify and compare strength of relation between the hydrophobicity scales and the raw index values, i.e., Hx, CHx, and 3CHx, and the secondary structure content. FCMs are used to quantify the degree of relation between the scales that use each of the three selected hydrophobicity indices and the content values. This goal allows concluding which of the considered scales is the most useful with respect to capturing the secondary structure when considering different indices.

GOAL 2: To evaluate two types of directed relations between hydrophobicity-based scales/raw index values and the secondary structure content, i.e. impact of the content on the hydrophobicity and impact of the hydrophobicity on the content. This analysis is possible due to inherent capabilities of FCMs, which in contrast to commonly used correlation based analysis reflect both strength and direction of a given relation.

GOAL 3: To quantify and compare the three hydrophobicity indices, i.e., Eisenberg's, Fauchere–Pliska's, and Cid's, with respect to their ability to correlate with the secondary structure content. Each of the indices is compared for each of the considered scales and when using the raw index values, and on average between all scales.

GOAL 4: To evaluate effectiveness of the FCM based analysis. The results are compared with results when using correlation analysis.

3. Experiments and results

The FCMs were used to evaluate the quality of the two hydrophobicity scales and the raw index values (for the three hydrobhobicity indices) with respect to their relation to the secondary structure content. To do that, a protein set has been carefully selected and used to generate input data consisting of the scale values and the corresponding content values along the sequences.

3.1. Experimental setup

The input data has been generated based on a set of 210 low homology proteins, which were used in two recent studies concerning secondary structure content prediction (Zhang et al., 2001; Kurgan and Homaeian, 2005). However, similarly as in Kurgan and Homaeian (2005), 11 of them (1MBA_, 1MDC_, 1OPAA, 4SBVA, 1FBAA, 1ETU_, 1GP1A, 3ADK_, 1CSEI, 1ONC_, and 1FUS_) have been excluded from experiments. These proteins include at least one unknown amino acid type in their sequence in the newest release of the PDB. As a result, experiments have been performed with 199 proteins. The

Table 3
Summary of input data for FCM modeling

H	CH	3CH	3CCH
H_x %helix %strand	CH_x %helix %strand	$3CH_x$ %helix %strand	CH_x $3CH_x$ %helix %strand

Table 4
Comparison of the secondary structure content between the original data set and data set used in the experimentation

Structural class	Average content values for the original data set			Average content values for the first 41 AAs		
	Helix	Strand	Coil	Helix	Strand	Coil
α	39.8	59.9	0.3	41.0	58.5	0.4
β	2.7	70.4	26.9	2.0	73.0	25.0
αβ	22.2	60.9	16.9	19.2	61.6	19.3

proteins are divided into three structural classes: α-class that contains 55 proteins, β-class with 72 proteins, and αβ-class with 72 proteins. The α-class contains proteins that have majority of helix structures, i.e. >15% helices, and<10% strands, β -class contains proteins that have majority of strand structures, i.e. >15% strands, and<10% helices, while the αβ-class contains remaining proteins (Zhang et al., 2001). To improve presentation of the results, the analysis of the relations has been performed separately for each of the structural classes. Table 3 describes the input data that was computed for each protein using the formulas defined in Section 2.2. H series exploit relation between hydrophobicity scale and the protein content, whereas series CH, 3CH, and 3CCH investigate the two new cumulative scales, both separately and together. The input data has been generated for each examined index, i.e. Eisenberg's, Fauchere–Pliska, and Cid.

Total of $3 \times 4 \times 199 = 2388$ series have been computed and used to learn FCM models. Considering the restrictions imposed by FCMs, the input data has been linearly normalized to the interval [0,1]. The average length of sequences belonging to the different structural classes equals 131.8, 150.2, and 164.8 for the α, β, and αβ classes, respectively. Differences in sequence length result in learning bias, i.e., structural classes, which express different secondary structure characteristics, have different average sequence lengths and input data length is shown to have impact on the quality of FCM learning (Stach et al., 2004). Therefore, to remove the bias the initial 41 points (corresponding to the first 41 AAs) have been used for each series, which corresponds to the length of the shortest protein in the data set (1LTSC). We note that these 41 points are representative of the entire data set, i.e., the secondary structure content of the original data set for each structural class is shown to be virtually identical to the content of the first 41 residues, see Table 4. In other words, the selected subset of points can be used to express the relation between the secondary structure and the hydrophobicity scales/raw values.

Example input data for *1YSAC* protein, which belong to α-class, is shown using black lines in Fig. 1. The 2388 connection matrices (FCM models) have been computed and average values for different series and structural classes are reported.

3.2. Experimental results

A sample result for 1YSAC protein using Cid hydrophobic index and for the four considered data series is shown in Fig. 1.

The original hydrophobicity index values produce sharply changing curves (see Fig. 1a), which are hardly correlated with the smooth curves of the content values. On the other hand, both proposed hydrophobicity scales are much smoother. The generated FCM models were able to reproduce the smooth signals, whereas in case of the original scale, their smoothing tendency is observed. The trends generated by FCMs for the proposed hydrophobicity scales exhibit correlation with the smooth content curves, which in turn shows that there is a strong relation between protein structural content and the hydrophobicity.

Detailed experimental results are included in Appendix A. The strength of the relations of interest with respect to the defined goals is summarized in Fig. 2 using bar-plots. The higher is the bar, the stronger is the corresponding relation. The figure allows performing visual analysis of relations from three different perspectives: scale/raw values-wide, index-wide, and direction-wide. The bars shown in Figs. 2a and c, and e depict the strength of relation that the protein content exerts on the corresponding scales/raw index values, whereas remaining subfigures show the strength of relation that the scales/raw index values exerts on the protein content. Rows in the figure show the strength of the relation for different hydrophobicity indices, while columns correspond to different directions of the relation. Finally, each figure compares the results for the different hydrophobicity scales/raw values, i.e., the non-cumulative, H, raw index values are shown in white, and the proposed cumulative scales are shown using other colors, and is divided into the three structural classes. The CH1 and 3CH1 bars correspond to CH and 3CH (cumulative) scales from the 3CCH series.

3.3. Discussion of results for goal 1

This section concentrates on comparison of relation strengths between different hydrophobicity scales/raw index values and the protein content. Based on the results shown in bold and underline in Table 5, the average values are 0.35 for the new cumulative (CH and 3CH) scales and

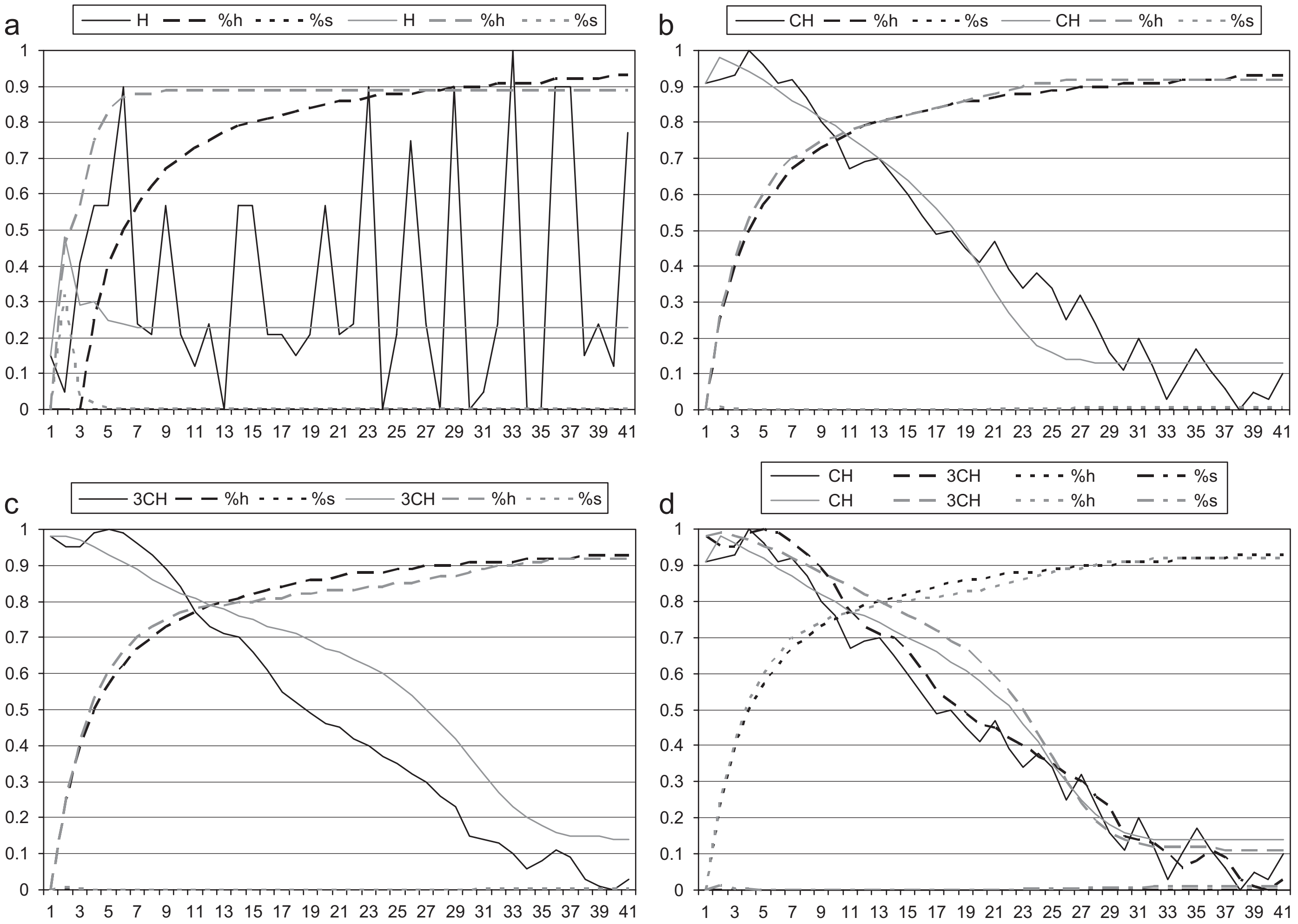

Fig. 1. Experimental results for the 1YSAC protein: (a) H series, (b) CH series, (c) 3CH series, and (d) 3CCH series; black lines correspond to the data used for learning, gray lines were obtained from simulation of the best-found FCM model.

0.143 for the currently used raw index values (H) in case of relation that the content exerts on the scale. Therefore, the proposed cumulative scales are characterized by about 2.5 times stronger relation when compared with the current non-cumulative raw values. On the other hand, for the opposite direction of the relation, relatively similar 0.482 and 0.592 values were recorded for cumulative scales and non-cumulative raw values, correspondingly. The strongest average strength of relation (over different hydrophobicity indices and directions) equals $(0.511+0.392+0.507)/3=0.469$ and it was observed for the 3CH scale. The average strength for the CH scale is comparable and equals 0.443, while the strength of the relation for the H raw values is the lowest and equal to 0.367. The cumulative, i.e. CH and 3CH, scales exhibit stronger relation irrespective of the applied hydrobhobicity indices when compared with the raw index values, except for the Cid's index when the relation from the scale/raw values to the content is considered. This shows that the cumulative scales, which include information about long-range interactions, are significantly better with respect to their relation with the secondary structure content.

3.4. Discussion of results for goal 2

We focus on evaluation of two types of directed relations between hydrophobicity based scales/raw index values and the protein content, i.e. relation that the content exerts on the scales/raw values and the scales/raw values exert on the content. In general, the strength of the relation varies for the two directions:

- In the case of relation that the scales/raw values exert on the protein content on average both the cumulative scales and the non-cumulative index values are characterized by similar relation strength. The two rightmost bars in Figs. 2b, d, and f show that series 3CCH, which considers both con-cumulative scales together, results in weaker relation. This may be caused by strong correlation between the two scales, which diminishes strengths of relation between them and the content, and therefore only one of them in separation should be used.
- In the case of relation that the content exerts on the scales/raw values, the cumulative scales exhibit virtually always significantly stronger relation when compared to

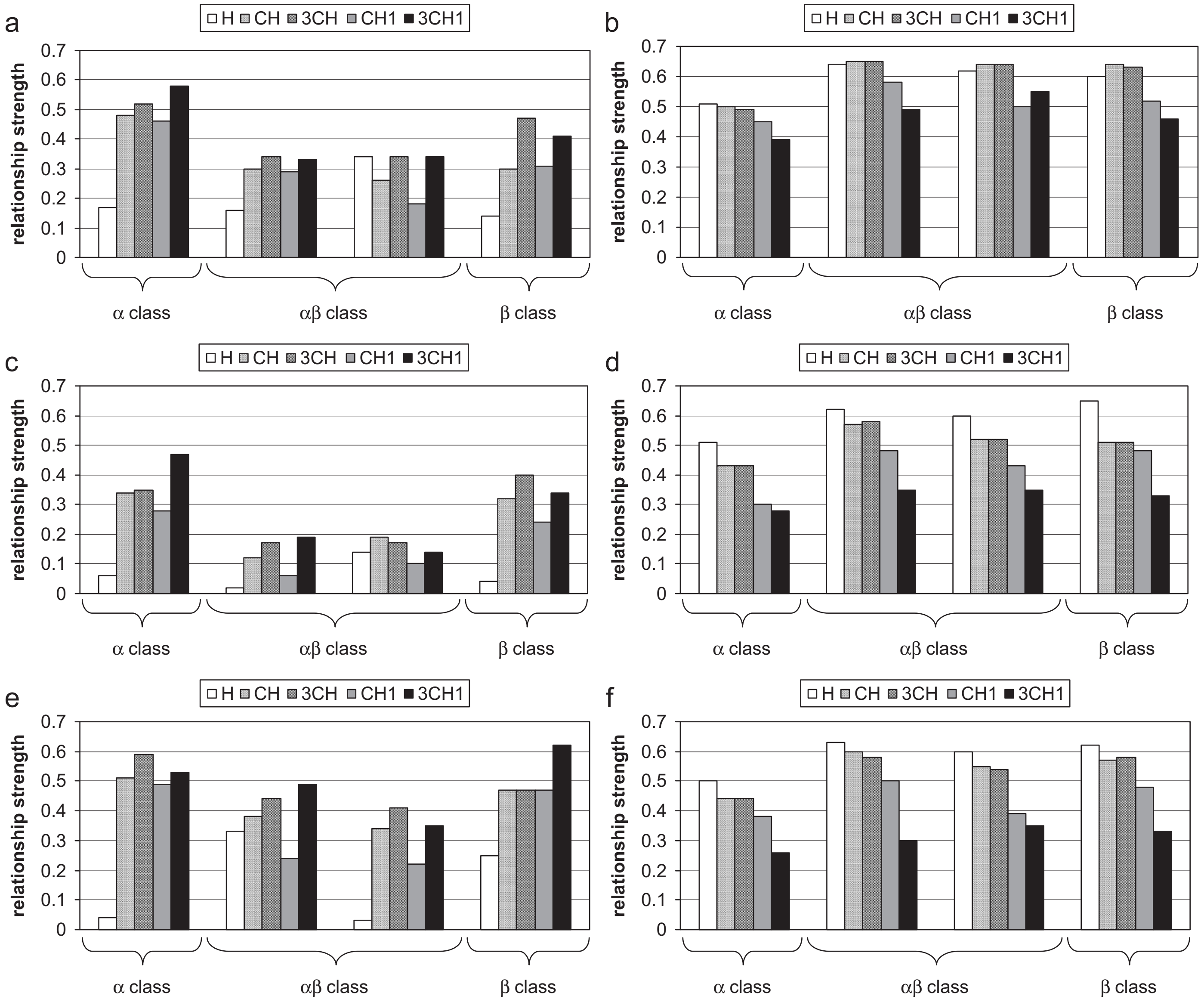

Fig. 2. Strength of relations that the content exerts on the hydrophobicity scales/raw index values (a) for the Eisenberg's index, (c) for the Fauchere–Pliska index, (e) for the Cid index; strength of relations that the hydrophobicity scales/raw index values exert on the content (b) for the Eisenberg's index, (d) for the Fauchere–Pliska index, and (f) for the Cid index.

Table 5
Summary of experiments with 199 proteins

Scales	Relations that the content exerts on the scale				Relations that the scale exerts on the content				Avg both directions		
	E index	F index	C index	Avg	E index	F index	C index	Avg	E index	F index	C index
CH	0.335	0.243	0.425	0.334	0.608	0.508	0.540	0.552	0.472	0.376	**0.483**
3CH	0.418	0.273	0.478	0.389	0.603	0.510	0.535	0.549	**0.511**	0.392	0.507
CH1	0.310	0.170	0.355	0.278	0.513	0.423	0.438	0.458	**0.412**	0.297	0.397
3CH1	0.415	0.285	0.498	0.399	0.473	0.328	0.310	0.370	**0.444**	0.307	0.404
Avg cumulat. scales	0.369	0.243	0.439	**0.350**	0.549	0.442	0.456	**0.482**	**0.459**	0.343	0.448
H	0.203	0.065	0.163	**0.143**	0.593	**0.595**	0.588	**0.592**	**0.398**	0.330	0.376

the non-cumulative raw values, which is shown using white bars in Figs. 2a, c, and e. The results for $\alpha\beta$-class show weaker strength for this direction of relation when compared with α- and β-classes, while in case of the other direction there is no significant difference between different structural classes.

Table 6
Correlation coefficients

		Eisenberg's index			Fauchere−Pliska's index			Cid's index		
		H scale	CH scale	3CH scale	H scale	CH scale	3CH scale	H scale	CH scale	3CH scale
α	%h	0.096	0.431	0.449	0.097	0.500	0.506	0.129	0.426	0.446
$\alpha\beta$	%h	0.118	0.487	0.508	0.097	0.583	0.595	0.085	0.465	0.484
	%s	0.126	0.403	0.419	0.128	0.603	0.604	0.130	0.380	0.382
β	%s	0.105	0.471	0.487	0.093	0.639	0.640	0.104	0.419	0.445
Avg		0.111	0.448	0.466	0.104	0.581	0.586	0.112	0.422	0.439

- Each of the two cumulative scales exhibit twice stronger relation with protein content for the relation that the protein content exerts on the hydrophobicity scales, while comparable strength of relation (about 0.55 when CH and 3CH scales are used separately vs. 0.59 for H raw values) is reported for the opposite direction.
- Finally, the average strength of relation that the content exerts on the scales/raw values is about twice weaker than the strength of the relation that the scale exerts on the content.

3.5. *Discussion of results for goal 3*

In this section, we compare the strength of relation between the hydrophobicity and the protein content with respect to the three hydrophobicity indices, i.e. Eisenberg's, Fauchere–Pliska's, and Cid's. The comparison among different hydrophobicity indices is performed based on average, over different scales, strengths of relations. To facilitate the comparison, average values of the cumulative scales are reported in Table 5, i.e., *avg cumulat. scales* row gives average for all cumulative scales and *avg both directions* column gives average over both directions. The bolded values (without underline) indicate the best performing indices for both the cumulative scales and the non-cumulative raw index values.

The table reveals that the Eisenberg's index provides the best results, i.e., the average strength of relation equals 0.511, 0.459 and 0.398 for the 3CH, over all cumulative scales, and for the H raw values, respectively. This confirms the conclusions from Juretic and Lucin (1998) and application of this index for a cumulative scale performed in Homaeian et al. (2007).

In case of the non-cumulative, H, raw values and the relation that the scales exert on the content all three indices are characterized by comparable strength. At the same time, the strongest relation (0.595) corresponds to the Fauchere–Pliska index. This is consistent with the reported results, i.e., the raw values with this index were used in Lin and Pan (2001), Zhang et al. (1998, 2001). The strength of the relation that the content exerts on the H raw values is relatively low, i.e., below 0.2, for all three indices.

In short, the results show that, on average, the Eisenberg's index with the cumulative scales is better than the currently used Fauchere–Pliska index and non-cumulative raw index values. As expected, the Cid index remains in the shadow of the other two indices.

3.6. *Discussion of results for goal 4*

The analysis concentrates on evaluation of effectiveness of the FCM-based results by comparing them with results using commonly applied correlation-based analysis. To perform comparison, Pearson correlation coefficients between corresponding concepts (hydrophobicity scales/raw values, and helix and strand content) for each data series have been computed. Absolute values of the correlation coefficients were used to compute averages, see Table 6.

The results show that the average, over the three indices, correlations for the cumulative scales equal $(0.466+0.586+0.439)/3=0.50$, $(0.448+0.581+0.422)/3=0.48$ and for the 3CH and CH scales, while for the H raw values the correlation equals $(0.111+0.104+0.112)=0.11$. Very similar values are observed when each index is analyzed separately. This confirms the results achieved with FCMs, although a much bigger difference is observed. The coefficient values around zero are usually associated with weak or no correlation, which indicates that the non-cumulative index values are weakly correlated with the secondary structure content. On the other hand, correlation of about 0.5 and higher are associated with strong correlation, which gives further confirmation of the quality of the two proposed scales.

The average, over the two scales and the raw index values, correlation for each index equals $(0.111+0.448+0.466)/3=0.34$ for Eisenberg's index, $(0.104+0.581+0.586)/3=0.42$ for the Fauchere–Pliska's index, and $(0.112+0.422+0.439)/3=0.32$ for the Cid's index. This shows that Fauchere–Pliska's index is better correlated with protein content when compared with the Eisenberg's index. Also, the average correlation for non-cumulative, H, raw values and Fauchere–Pliska's index is lower than for both Cid's and Eisenberg's indices, which is in disagreements with results published in Zhang et al. (2001). These results also do not agree with our conclusions that were drawn using FCMs. We argue that the correlation results are deceptive since they consider correlations between the hydrophobicity and only one of the content (helix or strand) values at the time. It is clear that since a protein contains both helices and strand and they are inherently related, only methods that can consider relation between all three concepts can give reliable results. Also, the correlation based analysis does not allow studying the strength of relations between concepts with respect to the direction of relation, which is one of the inherent features of the FCMs.

4. Conclusions

This paper proposes and performs analysis of two novel hydrophobicity scales. In contrast to the currently used raw index values, the new scales incorporate long-range interactions along the protein sequence. The two new scales and the raw hydrophobicity index values were compared using three hydrophobicity indices, i.e. Eisenberg's, Fauchere–Pliska's, and Cid's. The degree of the relation between the scales and the secondary structure content was quantified using fuzzy cognitive maps (FCMs) and a comprehensive set of 200 low homology proteins.

The results show that the new cumulative hydrophobicity scales are characterized by much stronger relation with the secondary protein content when compared to the currently used non-cumulative raw values. This conclusion holds true irrespective of the applied hydrobhobicity index. The strength of relation indicates that Eisenberg's index with the proposed scales is better than the currently used Fauchere–Pliska index and the raw values.

Acknowledgments

The authors would like to thank anonymous reviewers for their constructive and helpful criticism. This work was supported by the Alberta Ingenuity Fund and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Appendix A

The below table includes nine connection matrices (for the three corresponding hydrophobicity indices, i.e., Eisenberg's, Fauchere–Pliska's and Cid's, and three protein structural classes, i.e., α, β, and $\alpha\beta$) for each series. The matrices correspond to average, over the corresponding protein sets, FCM models. Each cell contains an average strength of relation across the proteins in the same structural class and for a given series computed using the same hydrophobicity index. The values reflect the strength of directed relation between the two corresponding concepts, which is normalized to the [−1,1] interval. For instance, value of −0.51 in the first row and the second column in table for H series shows that the strength of the cause–effect relation from concept H to concept %h for the H series. This value shows that the strength of the relation that the raw index values computed using Eisenberg's index exhibits on helix content is relatively high. In contrast, value of 0.17 in the second row and the first column shows that strength of the relation that the helix content exhibits on the raw hydrophobicity values based on Eisenberg's index is relatively low. The values shown in bold describe relations of our interest, which are summarized in Fig. 2 in the paper.

H series		Eisenberg's index			Fauchere-Pliska's index			Cid's index		
		H	%h	%s	H	%h	%s	H	%h	%s
α	H	0.07	**−0.51**	−1.00	−0.08	**−0.51**	−1.00	−0.25	**−0.50**	−0.90
	%h	**0.17**	0.11	−0.76	**−0.06**	0.03	−0.81	**−0.04**	0.06	−0.80
	%s	0.03	0.15	0.19	0.22	−0.05	0.05	0.06	0.16	0.09
$\alpha\beta$	H	0.04	**−0.64**	**−0.62**	−0.13	**−0.62**	**−0.60**	−0.21	**−0.63**	**−0.60**
	%h	**0.16**	0.27	−0.57	**−0.02**	0.15	−0.52	**−0.33**	0.00	−0.60
	%s	**0.34**	−0.60	0.24	**0.14**	−0.58	0.11	**0.03**	−0.57	0.05
β	H	0.07	−0.95	**−0.60**	−0.09	−0.94	**−0.65**	−0.22	−0.94	**−0.62**
	%h	0.09	0.14	0.11	−0.01	0.06	−0.01	0.15	0.10	−0.03
	%s	**0.14**	−0.77	0.00	**−0.04**	−0.65	−0.04	**−0.25**	−0.72	0.00

CH series		Eisenberg's index			Fauchere-Pliska's index			Cid's index		
		CH	%h	%s	CH	%h	%s	CH	%h	%s
α	CH	0.44	**−0.50**	−0.99	0.56	**−0.43**	−0.92	0.50	**−0.44**	−0.88
	%h	**−0.48**	0.03	−0.78	**−0.34**	0.22	−0.82	**−0.51**	0.03	−0.77
	%s	0.00	0.04	0.18	0.09	0.16	0.12	0.03	0.01	0.21
$\alpha\beta$	CH	0.37	**−0.65**	**−0.64**	0.47	**−0.57**	**−0.52**	0.41	**−0.60**	**−0.55**
	%h	**−0.30**	0.31	−0.44	**−0.12**	0.39	−0.26	**−0.38**	0.23	−0.42

	%s	**−0.26**	−0.50	0.24	**−0.19**	−0.40	0.24	**−0.34**	−0.56	0.25
β	CH	0.32	−0.95	**−0.64**	0.45	−0.92	**−0.51**	0.32	−0.92	**−0.57**
	%h	−0.01	0.12	0.02	0.24	0.28	0.14	−0.08	0.12	0.02
	%s	**−0.30**	−0.74	0.02	**−0.32**	−0.63	0.12	**−0.47**	−0.71	0.05

3CH series		Eisenberg's index			Fauchere-Pliska's index			Cid's index		
		3CH	%h	%s	3CH	%h	%s	3CH	%h	%s
α	3CH	0.53	**−0.49**	−0.99	0.60	**−0.43**	−0.90	0.56	**−0.44**	−0.93
	%h	**−0.52**	0.07	−0.89	**−0.35**	0.24	−0.82	**−0.59**	0.04	−0.72
	%s	−0.05	0.16	0.16	0.10	0.07	0.27	−0.02	0.08	0.07
αβ	3CH	0.47	**−0.65**	**−0.64**	0.52	**−0.58**	**−0.52**	0.53	**−0.58**	**−0.54**
	%h	**−0.34**	0.28	−0.45	**−0.17**	0.37	−0.41	**−0.44**	0.17	−0.52
	%s	**−0.34**	−0.55	0.24	**−0.17**	−0.37	0.39	**−0.41**	−0.60	0.22
β	3CH	0.39	−0.95	**−0.63**	0.49	−0.92	**−0.51**	0.41	−0.91	**−0.58**
	%h	0.03	0.13	0.07	0.19	0.04	0.15	−0.05	0.25	−0.03
	%s	**−0.47**	−0.78	−0.00	**−0.40**	−0.62	0.15	**−0.47**	−0.70	0.12

3CCH series		Eisenberg's index				Fauchere-Pliska's index				Cid's index			
		CH	3CH	%h	%s	CH	3CH	%h	%s	CH	3CH	%h	%s
α	CH	0.33	0.13	**−0.45**	−0.98	0.42	0.20	**−0.30**	−0.72	0.43	0.24	**−0.38**	−0.64
	3CH	0.12	0.44	**−0.39**	−0.99	0.16	0.45	**−0.28**	−0.68	0.06	0.38	**−0.26**	−0.68
	%h	**−0.46**	**−0.58**	0.02	−0.75	**−0.28**	**−0.47**	0.21	−0.57	**−0.49**	**−0.53**	0.19	−0.69
	%s	−0.04	0.11	0.11	0.14	0.17	−0.03	0.06	0.19	0.17	−0.14	0.07	0.22
αβ	CH	0.35	0.16	**−0.58**	**−0.50**	0.34	0.11	**−0.48**	**−0.43**	0.47	0.24	**−0.50**	**−0.39**
	3CH	0.03	0.34	**−0.49**	**−0.55**	0.12	0.43	**−0.35**	**−0.35**	−0.07	0.32	**−0.30**	**−0.35**
	%h	**−0.29**	**−0.33**	0.34	−0.30	**−0.06**	**−0.19**	0.49	−0.29	**−0.24**	**−0.49**	0.27	−0.51
	%s	**−0.18**	**−0.34**	−0.44	0.30	**−0.10**	**−0.14**	−0.34	0.35	**−0.22**	**−0.35**	−0.48	0.24
β	CH	0.36	0.10	−0.94	**−0.52**	0.29	0.03	−0.81	**−0.48**	0.37	0.08	−0.78	**−0.48**
	3CH	−0.04	0.31	−0.93	**−0.46**	0.13	0.44	−0.71	**−0.33**	−0.06	0.36	−0.68	**−0.33**
	%h	0.16	−0.02	0.10	0.09	0.05	0.11	0.27	0.18	0.02	−0.04	0.12	0.03
	%s	**−0.31**	**−0.41**	−0.70	0.08	**−0.24**	**−0.34**	−0.47	0.18	**−0.47**	**−0.62**	−0.56	0.18

References

Aguilar, J., 2005. A survey about Fuzzy cognitive maps papers. Int. J. Comput. Cogn. 3 (2), 27–33.

Axelrod, R., 1976. Structure of Decision: the Cognitive Maps of Political Elites. Princeton University Press.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Biou, V., Gibrat, J.F., Levin, J.M., Robson, B., Garnier, J., 1988. Secondary structure prediction: combination of three different methods. Protein Eng. 2, 185–191.

Birzele, F., Kramer, S., 2006. A new representation for protein secondary structure prediction based on frequent patterns. Bioinformatics 22 (21), 2628–2634.

Bull, B.M., Breese, K., 1974. Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. Arch. Biochem. Biophys. 161, 665–670.

Cai, Y.D., Chou, K.C., 2005. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J. Proteome Res. 4, 967–971.

Cai, Y., Liu, X-J., Xu, X., Chou, K-C., 2002a. Artificial neural network method for predicting protein secondary structure content. Comput. Chem. 26, 347–350.

Cai, Y., Liu, X-J., Chou, K-C., 2002b. Prediction of protein secondary structure content by artificial Neural network. J. Comput. Chem. 24 (6), 727–731.

Cai, Y.D., Zhou, G.P., Chou, K.C., 2005. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J. Theor. Biol. 234, 145–149.

Chen, C., Zhou, X., Tian, Y., Zou, X., Cai, P., 2006a. Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357, 116–121.

Chen, C., Tian, Y.X., Zou, X.Y., Cai, P.X., Mo, J.Y., 2006b. Using pseudo-amino acid composition and support vector machine to predict protein structural class. J. Theor. Biol. 243, 444–448.

Chou, K.C., 1999. Using pair-coupled amino acid composition to predict protein secondary structure content. J. Protein Chem. 18 (4), 473–480.

Chou, K.C., 2000a. Review: prediction of tight turns and their types in proteins. Anal. Biochem. 286, 1–16.

Chou, K.C., 2000b. Review: prediction of protein structural classes and subcellular locations. Curr. Protein Peptide Sci. 1, 171–208.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43 (3), 246–255 [Erratum: Proteins, 2001, 44, 60].

Chou, K.C., 2005a. Review: progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr. Protein Peptide Sci. 6, 423–436.

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

Chou, K.C., 2005c. Prediction of G-protein-coupled receptor classes. J. Proteome Res. 4, 1413–1418.

Chou, K.C., Elrod, D.W., 2002. Bioinformatical analysis of G-protein-coupled receptors. J. Proteome Res. 1, 429–433.

Chou, K.C., Cai, Y.D., 2003a. Predicting protein quaternary structure by pseudo amino acid composition. Proteins 53, 282–289.

Chou, K.C., Cai, Y.D., 2003b. Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. J. Cellular Biochem. 90, 1250–1260 [*Addendum*, *Journal of Cellular Biochemistry*, 2004, 91, 1085].

Chou, K.C., Cai, Y.D., 2004a. Predicting protein structural class by functional domain composition. Biochem. Biophys. Res. Commun. 321, 1007–1009 [*Corrigendum*: *Biochem. Biophys. Res. Commun.*, 2005, 329, 1362].

Chou, K.C., Cai, Y.D., 2004b. Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. J. Cellular Biochem. 91, 1197–1203.

Chou, K.C., Cai, Y.D., 2005. Prediction of membrane protein types by incorporating amphipathic effects. J. Chem. Inform. Model. 45, 407–413.

Chou, K.C., Cai, Y.D., 2006. Prediction of protease types in a hybridization space. Biochem. Biophys. Res. Comm. 339, 1015–1020.

Chou, K.C., Shen, H.B., 2006a. Predicting protein subcellular location by fusing multiple classifiers. J. Cellular Biochem. 99, 517–527.

Chou, K.C., Shen, H.B., 2006b. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. 347, 150–157.

Chou, K.C., Shen, H.B., 2006c. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J. Proteome Res. 5, 1888–1897.

Chou, K.C., Shen, H.B., 2006d. Large-scale predictions of Gram-negative bacterial protein subcellular locations. J. Proteome Res. 5, 3420–3428.

Chou, K.C., Shen, H.B., 2007. Large-scale plant protein subcellular location prediction. J. Cellular Biochem. 100, 665–678.

Cid, H., Bunster, M., Canales, M., Gazitua, F., 1992. Hydrophobicity and structural classes in proteins. J. Protein Eng. 5, 373–375.

Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A., DeLisi, C., 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in protein. J. Molec. Biol. 195, 659–685.

Ding, C.H., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358.

Du, P., Li, Y., 2006. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physico-chemical features of segmented sequence. BMC Bioinform. 7, 518.

Du, Q.-S., Liu, P.-J., Mezey, P.G., 2005. Theoretical derivation of heuristic molecular lipophilicity potential: a quantum chemical description for molecular solvation. J. Chem. Inform. Model 45 (2), 347–353.

Du, Q.-S., Wang, S.-Q., Chou, K.C., 2006. Heuristic molecular lipophilicity potential (HMLP): lipophilicity and hydrphilcity of amino acid side chains. J. Comput. Chem. 27, 685–692.

Eisenhaber, F., Imperiale, F., Argos, P., Frommel, C., 1996. Prediction of secondary structural content of proteins from their amino acid composition alone. I. New analytic vector decomposition methods. Proteins 25 (2), 157–168.

Fauchere, J.L., Pliska, V., 1983. Hydrophobic parameters-Pi of amino-acid side-chains from the partitioning of *N*-acetyl-amino-acid amides. Eur. J. Med. Chem. 18, 369–375.

Feng, K.Y., Cai, Y.D., Chou, K.C., 2005. Boosting classifier for predicting protein domain structural class. Biochem. Biophys. Res. Commun. 334, 213–217.

Gao, Y., Shao, S.H., Xiao, X., Ding, Y.S., Huang, Y.S., Huang, Z.D., Chou, K.C., 2005. Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28, 373–376.

Guo, Z.M., 2002. Prediction of Membrane protein types by using pattern recognition method based on pseudo amino acid composition. Master Thesis, Bio-X Life Science Research Center, Shanghai Jiaotong University.

Guo, Y.Z., Li, M., Lu, M., Wen, Z., Wang, K., Li, G., Wu, J., 2006. Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30, 397–402.

Herrera, F., Lozano, M., Verdegay, J.L., 1998. Tackling real-coded genetic algorithms: operators and tools for behavioural analysis. Artificial Intel. Rev. 12 (4), 265–319.

Homaeian, L., Kurgan, L.A., Cios, K.J., Ruan, J., Chen, K., 2007. Prediction of protein secondary structure content for the Twilight zone sequences. Proteins, in print.

Jones, D., 2000. Protein structure prediction in the postgenomic era. Curr. Opin. Struct. Biol. 10 (3), 371–379.

Juretic, D., Lucin, A., 1998. The preference functions method for predicting protein helical turns with membrane propensity. J. Chem. Inform. Comput. Sci. 38 (4), 575–585.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22 (12), 2577–2637.

Kawashima, S., Ogata, H., Kanehisa, M., 1999. AA index: amino acid index database. Nucleic Acids Res 27, 368–369.

Kedarisetti, K., Kurgan, L.A., Dick, S., 2006. Classifier ensembles for protein structural class prediction with varying homology. Biochem. Biophys. Res. Commun. 348 (3), 981–988.

Kosko, B., 1986. Fuzzy cognitive maps. Int. J. Man-Machine Stud. 24, 65–75.

Koulouriotis, D.E., Diakoulakis, I.E., Emiris, D.M., Antonidakis, E.N., Kaliakatsos, I.A., 2003. Efficiently modeling and controlling complex dynamic systems using evolutionary fuzzy cognitive maps. Int. J. Comput. Cogn. 1 (2), 41–65.

Kurgan, L.A., Homaeian, L., 2005. Prediction of Secondary Protein Structure Content from Primary Sequence Alone—a Feature Selection Based Approach. International Conference on Machine Learning and Data Mining, Leipzig, Germany, LNAI 4587, pp. 334–345.

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 Recognition 39 (12), 2323–2343.

Kurgan, L.A., Rahbari, M., Homaeian, L., 2006. Impact of the Predicted Protein Structural Content on Prediction of Structural Classes for the Twilight Zone Proteins. In: Fifth International Conference on Machine Learning and Applications, Orlando, FL, pp. 180–186.

Lee, S., Lee, B.C., Kim, D., 2006. Prediction of protein secondary structure content using amino acid composition and evolutionary information. Proteins 62 (4), 1107–1114.

Liu, W., Chou, K.C., 1999. Protein secondary structural content prediction. Protein Eng. 12, 1041–1050.

Lin, Z., Pan, X., 2001. Accurate prediction of protein secondary structural content. J. Protein Chem. 20 (3), 217–220.

Lin, H., Li, Q.Z., 2007a. Using Pseudo Amino Acid Composition to Predict Protein Structural Class: Approached by Incorporating 400 Dipeptide Components. J. Comput. Chem. doi:10.1002/jcc.20554.

Lin, H., Li, Q.Z., 2007b. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem. Biophys. Res. Commun. 354, 548–551.

Lin, K., Simossis, V.A., Taylor, W.R., Heringa, J., 2005. A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 21 (2), 152–159.

Liu, H., Yang, J., Wang, M., Xue, L., Chou, K.C., 2005a. Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. Protein J. 24, 385–389.

Liu, H., Wang, M., Chou, K.C., 2005b. Low-frequency Fourier spectrum for predicting membrane protein types. Biochem. Biophys. Res. Commun. 336, 737–739.

Luo, R.Y., Feng, Z.P., Liu, J.K., 2002. Prediction of protein strctural class by amino acid and polypeptide composition. Eur. J. Biochem. 269, 4219–4225.

Miyazawa, S., Jernigan, R.L., 1985. Estimation of elective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18, 534–552.

Mondal, S., Bhavna, R., Mohan Babu, R., Ramakumar, S., 2006. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J. Theor. Biol. 243, 252–260.

Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K., Pedersen, J.T., 1997. Critical assessment of methods of protein structure prediction (CASP): round II. Proteins 29, 2–6.

Niu, B., Cai, Y.D., Lu, W.C., Zheng, G.Y., Chou, K.C., 2006. Predicting protein structural class with AdaBoost learner. Protein Peptide Lett. 13, 489–492.

Nishikawa, K., Ooi, T., 1980. Prediction of the surfaceinterior diagram of globular proteins by an empirical method. Int. J. Peptide Protein Res. 16, 19–32.

Pan, Y.X., Zhang, Z.Z., Guo, Z.M., Feng, G.Y., Huang, Z.D., He, L., 2003. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. J. Protein Chem. 22, 395–402.

Parker, J.M.R., Guo, D., Hodges, R.S., 1986. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigencity and X-ray-derived accessible sites. Biochemistry 27, 5425–5432.

Pilizota, T., Lucic, B., Trinajstic, N., 2004. Use of variable selection in modeling the secondary structural content of proteins from their composition of amino acid residues. J. Chem. Inf. Comput. Sci. 44 (1), 113–121.

Ponnuswamy, P.K., Prabhakaran, M., Manavalan, P., 1980. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. Biochim. Biophys. Acta 623, 301–316.

Rost, B., 2001. Review: protein secondary structure prediction continues to rise. J. Struct. Biol. 134 (2–3), 204–218.

Ruan, J., Wang, K., Yang, J., Kurgan, L., Cios, K., 2005. Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. Artif. Intell. Med. 35 (1–2), 19–35.

Shen, H.B., Chou, K.C., 2005a. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem. Biophys. Res. Commun. 334, 288–292.

Shen, H.B., Chou, K.C., 2005b. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem. Biophys. Res. Comm. 337, 752–756.

Shen, H.B., Chou, K.C., 2006a. Ensemble classifier for protein fold pattern recognition. Bioinformatics 22, 1717–1722.

Shen, H.B., Chou, K.C., 2006b. Using ensemble classifier to identify membrane protein types. Amino Acids, doi:10.1007/s00726-00006-00439-00722.

Shen, H.B., Chou, K.C., 2007a. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng. Design Select. 20, 39–46.

Shen, H.B., Chou, K.C., 2007b. Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85, 233–240.

Shen, H.B., Chou, K.C., 2007c. Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem. Biophys. Res. Commun. 355, 1006–1011.

Shen, H.B., Yang, J., Liu, X.J., Chou, K.C., 2005. Using supervised fuzzy clustering to predict protein structural classes. Biochem. Biophys. Res. Commun. 334, 577–581.

Shen, H.B., Yang, J., Chou, K.C., 2006. Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J. Theor. Biol. 240, 9–13.

Shen, H.B., Yang, J., Chou, K.C., 2007. Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction, Amino Acids, doi:10.1007/s00726-00006-00478-00728.

Shi, J.Y., Zhang, S.W., Pan, Q., Cheng, Y.-M., Xie, J., 2007. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids, doi:10.1007/s00726-00006-00475-y.

Stach, W., Kurgan, L., Pedrycz, W., Reformat, M., 2004. Learning fuzzy cognitive maps with required Precision using genetic algorithm approach. Electron. Lett. 40 (24), 1519–1520.

Stach, W., Kurgan, L., Pedrycz, W., Reformat, M., 2005. Genetic learning of fuzzy cognitive maps. Fuzzy Sets and Systems 153 (3), 371–401.

Sweet, R.M., Eisenberg, D., 1983. Correlation of sequence hydrophobicities measures similarity in three dimensional protein structure. J. Mol. Biol. 171, 479–488.

Wang, M., Yang, J., Liu, G.P., Xu, Z.J., Chou, K.C., 2004. Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng. Design Select. 17, 509–516.

Wang, M., Yang, J., Xu, Z.J., Chou, K.C., 2005. SLLE for predicting membrane protein types. J. Theor. Biol. 232, 7–15.

Wang, S.Q., Yang, J., Chou, K.C., 2006. Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. Journal of Theoretical Biology 242, 941–946.

Wen, Z., Li, M., Li, Y., Guo, Y., Wang, K., 2006. Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32, 277–283.

Wertz, D.H., Scheraga, H., 1978. Infuence of water on protein structure. An analysis of the preferences of amino acids residues for the inside or outside and for specific conformations in a protein molecule. Macromolecules 11, 9–15.

Wold, S., Eriksson, L., Hellberg, S., Jonsson, J., Sjöström, M., Skagerberg, B., Wikström, C., 1987. Six non-natural amino acids and their application to a structure activity relationship for oxytocin peptide analogues. Can. J. Chem. 65, 1814–1820.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Huang, Y., Chou, K.C., 2005a. Using complexity measure factor to predict protein subcellular location. Amino Acids 28, 57–61.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., Chou, K.C., 2005b. Using cellular automata to generate Image representation for biological sequences. Amino Acids 28, 29–35.

Xiao, X., Shao, S.H., Huang, Z.D., Chou, K.C., 2006a. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J of Comput. Chem 27, 478–482.

Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D., Chou, K.C., 2006b. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30, 49–54.

Zhang, C., Lin, Z-S., Zhang, Z., Yan, M., 1998. Prediction of Helix/Strand content of globular proteins based on their primary sequences. Protein Eng. 11 (11), 971–979.

Zhang, Z., Sun, Z., Zhang, C., 2001. A New approach to predict the Helix/Strand content of globular proteins. J. Theor. Biol. 208, 65–78.

Zhang, S.W., Pan, Q., Zhang, H.C., Shao, Z.C., Shi, J.Y., 2006. Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30, 461–468.

Zhou, G.P., Cai, Y.D., 2006. Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. Proteins 63, 681–684.