VOL4 NO. 1 March 2015 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: www.scjournal.ius.edu.ba

IUS Soft Computing Research Group

Prediction of Protein Structural Classes for Low-Similarity Sequences Based On Predicted Secondary Structure

Betul Akcesme
International University of Sarajevo,
Faculty of Engineering and Natural Sciences,
HrasnickaCesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina
bcicek@iuso.edu.ba

Article Info

Article history:
Article received on march 2015
Received in revised form April 2015

Keywords:
Protein structural class prediction;
Secondary structure;
Alternating frequency;
Parallel and anti-parallel β-sheets;
Support vector machine

Abstract

Knowledge about structural classes of proteins plays an important role in inferring tertiary structure and function of a protein. One of the major problems with the existing algorithm for the prediction of protein structural classes is low accuracies for proteins from α+β and α/β classes. To improve accuracies, one needs to extract features with high representation power. Several authors proposed enormous number of features. Some of them redundant, most of them overlapping. In this paper, most prominent features proposed in the literature are reviewed. Features extracted from Position Specific Scoring Matrices (PSSM) are excluded and left as the subject matter of another paper. Also some combinations of these features are used to classify a low-homology dataset, 25PDB, and 30FB, with sequence similarity lower than 25% and 30%, respectively. Comparison of our results with others shows that to find the best combination is very important and may provide a cost-effective alternative to predict protein structural class in particular for low-similarity datasets.

1. INTRODUCTION

In many fields of bioinformatics the knowledge about the structural classes of proteins is important in (Chou,2004, 2005; Kurgan and Homaeian, 2006; Costantini andFacchiano, 2009). As of Februrary 2015, in SCOPe 2.05, SCOP experts manually classified 71,000 of the 110,000 total PDB entries, and about 90% of them belong to the four major classes;all-α, all-β, α+β and α/β classes (Andreeva et al., 2004; Murzinet al., 1995). The classification of protein structures in SCOP isdone manually based on proteins with known tertiary structures.

Since the rapid development of the genomics and proteomics,there has been enormous accumulation of data on the amino acidsequences of proteins. Therefore, the manual method apparentlycannot cope with the demand for rapid classification. Hence overthe past two decades, researchers have made unremitting effortsin computational prediction of structural classes on the basis ofthe amino acid sequences of the proteins. There are generally twoaspects in the computational prediction: feature vector andclassification algorithm.

A Large number of sequence features have been applied to representprotein sequences, including amino acid composition(Nakashima, et. al., 1986; Zhou, 1998; Chou, 1999),pseudo amino acid composition(Chou, 2001; Lin, and Li, 2007; Xiao et. al., 2006; Zhang, and Ding, 2007;

Zhang et. al., 2008), polypeptidecomposition (Luo, et. al., 2002; Sun, and Huang, 2006), functional domain composition (Chou, andCai, 2004), PSIBLASTprofile (Chen, et. al., 2008;Liu, et. al., 2010), predicted secondary structure information(Kurgan, et. al., 2008a, 2008b; Mizianty, and Kurgan, 2009) and amino acid sequence reverse encoding (Deschavanne, andTuffery, 2008; Yang, et. al, 2009).

Meanwhile, many machine learning algorithms have been alreadyused to implement the protein structural class predictions, such asneural network (Cai, G.P. Zhou, 2000), support vector machine (Anand, et. al., 2008; Cai, et. al, 2001, 2002; Chen, et. al., 2006; Qiu, et. al., 2009),fuzzy clustering (Shen, et. al., 2005), Bayesian classification (Wang, and Yuan, 2000), rough sets (Cao, et. al., 2006),information discrepancy (Jin, et. al., 2003; Kedarisetti, et. al., 2006; Zhang, et. al., 2009) and classifier fusion technique(Feng, et. al., 2005; Kedarisetti, et. al., 2006; Cai, et. al., 2006). The existing sequence representation methods and classificationalgorithms have been extensively reviewed (Chou, 2005; Kurgan and Homaeian, 2006).

Prediction accuracies of various different structural classprediction methods are given in Table 1.

Table 1
Comparison of prediction accuracies among various different structural classprediction methods.

Data	Method	Accuracy (%)				
		α	β	α/β	α+β	O.all
25PDB	SCPRED	92.6	80.1	74.0	71.0	79.7
	MODAS	92.3	83.7	81.2	68.3	81.4
	LIU-JIA	92.6	81.3	81.5	76.0	82.9
	ZHANG	95.0	85.6	73,2	81.5	83.9
	KONG	94.1	87.1	74.4	81.1	85.0
D675	SCPRED	89.1	81.8	90.4	58.2	79.5
	MODAS	89.9	81.8	84.2	65.9	80.0
	LIU-JIA	90.8	81.4	84.7	68.6	82.0
	ZHANG	–	–	–	–	–
	KONG	–	–	–	–	–

In this paper,we propose a newcombination of feature set. When testingwas made on the 25PDB dataset including 1670 (three of the 1673 proteins in 25PDB dataset are removed because of their deficiencies) proteins withtwilight-zone similarity, an overall average prediction accuracy of80.6% was obtained with whichis 17.9% higher than the most competing methods using onlyinformation extracted from amino acid sequences (Kurgan andChen, 2007; Kurgan et al., 2008b; Mizianty and Kurgan, 2009). On the 30FB dataset including 10289 proteins with less than 30% similarity, an overall average prediction accuracy of 80.6% was obtained, whichis also 17.9% higher than methods using onlyinformation extracted from amino acid sequences

Though the overall prediction accuracy has been improved, theprediction accuracies for α+β and α/βclasses are still unsatisfactory(73.6% on average) (Table 1). In this study, we tried to furtherimprove the prediction accuracy in a different way. The 11features were selected by knowledge-based rational design ratherthan random screening. Three of the features were speciallydesigned to improve the prediction accuracies for proteins fromα+βand α/β classes. The prediction performed with an optimizedsupport vector machine revealed an improvement in predictionaccuracy, especially for proteins from α+βand α/β classes.

2.FORMULATION OF THE PROBLEM

(a) Database

Three widely used datasets with low sequence identity wereused in this study to compare the accuracy of our prediction withthose of existing prediction methods. The 25PDB dataset is thetouchstone for the prediction of protein secondary structuralclasses due to a comparatively larger number of proteins, 1670 proteins anddomains, and low identity among samples, average identity of25% (Kurgan and Homaeian, 2006). The 30FB dataset, a huge number of 10289 proteinsand domains with average identity of 30%,which is created from SCOPE database by the authors and their research group at International University of Sarajevo, Faculty of Engineering and Natural Sciences.

(b) Features

SCOP scientists manually classify proteins into different structural classesaccording to their 3-D structures; hence the featuresderived from these structures might directly be applied to theprediction of protein structural classes. In this paper, most prominent features proposed in the literature are reviewed. Features extracted from Position Specific Scoring Matrices are excluded and left as the subject matter of another paper.In this article three of these feature sets will be presented: Liu, and Jia (Liu, and Jia,2010), Zhang (Zhang, et. al., 2011), and Kong, and Zhang (Kong, and Zhang,2014).

Liu, and Jia (2010) Features

From the beginning, a standard for protein structure classificationis the content of the secondary structural elements (Chou,2005), ConH and ConEthat reflect the contents of Hand E residues, respectively (Kurgan et al., 2008a, 2008b).The sequence length was denoted by N.

(1) P1 and P2 represent the content of residues H (ConH) and E (ConE), respectively, in the secondary structural sequence.
(2) P3 and P4 represent normalized length of the longest a-helix (MaxSegH/N) and b-strand (MaxSegE/N), respectively.The sequence length was denoted by N.

(3) P5 and P6 represent the normalized average length of a-helices (AvgSegH/N) and b-strands (AvgSegE/N), respectively.

(4) P7(CMVH) and P8(CMVE) represent the composition moment vectors H and E, respectively, which are formulated as

$$P_7 = \frac{\sum_{j=1}^{nH} nHj}{N(N-1)}, \quad P_8 = \frac{\sum_{j=1}^{nE} nEj}{N(N-1)} \qquad (1)$$

wherenH and nE are the total number of H and E residues in thesequence of the secondary structure, respectively; nHj and nEj are the jth position (in the secondary structure sequence) of H and Eresidues, respectively.

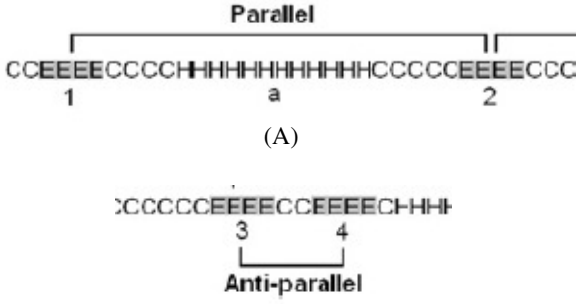
(5) P9 represents the normalized alternating frequency of a-helices and β-strands (Altn/N).

(6) P10 and P11 represent the proportion of parallel β-sheets and anti-parallel β-sheets, respectively, which can be calculated as follows:

$$P_{10} = \frac{PnE}{PnE+APnE}, \quad P_{11} = \frac{APnE}{PnE+APnE}, \qquad (2)$$

In (5) and (6) three novel features of the secondary structure were proposedon the basis of the structural characteristics of proteins from α/βand α+β classes.(Liu, and Jia, 2010)

α-helices and β-strands are usually separated in α/β proteins,but are usually interspersed in α+β proteins. In α/β proteins,α-helices and β-strands alternate more frequently than in α+βproteins. Therefore, the first feature was chosen as thealternating frequency of α-helices and β-strands (Altn). As an example in Fig. 1A, α-helices andβ-strands alternate two times (Altn=2).



(A)

(B)

Fig. 1. (A) and (B), amethod for the determination of β-strands composingparallel β-sheets or anti-parallel β-sheets.

Consider that the β-strands in α/βproteins are usually composedof parallel b-sheets, while the β-strands in α+βproteins are usuallycomposed of anti-parallel β-sheets, the second and the third featuresare based on the number of β-strands that form parallel β-sheets(PnE) and the number of β-strands that form anti-parallel (APnE)β-sheets, respectively (Fig. 1A, B). Liu, and Jia proposed that if two β-strands(segments of E) are separated by α-helix (segments of H), these twoβ-strands would form parallel β-sheets. Otherwise, they would formanti-parallel β-sheets.

3. CLASSIFICATION ALGORITHM CONSTRUCTION

Support vectormachine (SVM) has been successfully used in theprediction of protein secondary structural class because of its highaccuracy (Kurgan et al., 2008a, 2008b). The SVM classifier mapsfeature vectors into multi-dimensional space by using kernelfunction K(x), as a x-insensitive loss function and regulatory parameterC. x-insensitive loss function and regulatory parameterC. Here, Guassian kernel function

$$K(x_i, x_j) = exp(-\gamma \|x_i, x_j\|^2) \qquad (3)$$

ischosen for its superiority for solving nonlinear problems compared with other kernel functions (Yuan et al., 2005). The parameterization of SVM was performed through a grid search over γ and C values based on cross-validation on datasets. The final classifier uses C = 362 andγ= 0.7.

RESULTS

The prediction method was examined with 25PDB, and 30FB datasets.Our results on the two datasets with these features are as in Table 2

Table 2
Comparison of prediction accuracies among various different structural class prediction methods.

Data	Method	Accuracy (%)				
		α	β	α/β	α+β	O.all
25PDB	SVM	85.3	83.2	69.7	58.3	74.1
	ANN	90.2	84.1	74.0	64.4	78.2
	EUCLID	85.3	80.0	44.9	71.7	70.5
	MAHAL	81.8	72.7	45.5	81.8	70.5
	CAMBE	88.0	75.3	70,5	81.9	78.9
	MANH	89.4	80.3	72.3	54.2	74.1
30FB	SVM	72.5	69.0	60.2	46.1	62.0

Zhang,Ding, and Wang (2011) Features

Zhang,Ding, and Wang (Zhang, et. al., 2011) proposed eleven features where eight of them are the same as in Liu, and Jia (Liu, and Jia, 2010), and three newly-designed features are rationally utilized to reflect thegeneral contents and spatial arrangements of the secondarystructural elements of a given protein sequence.

The three novel features of *Zhang,Ding, and Wang* are derived from the secondarystructure sequences of proteins to characterize the distributions ofα helices and β strands, and hopefully that they could be used todistinguishα+β, and α/βclasses.

As the first step, they reduced a secondary structuresequence into a segment sequence, which is composed of helixsegments and strand segments denoted by αand β, respectively.

Here, α-helixsegment refers to a continuous segment of allHsymbols, and β-strandrefers to a continuous segment of all E symbols in the secondary structure sequence. In orderto focus on the arrangement of α-helix and β-strand segments, thecoil segments are ignored in the reduced segment sequence.

For example, given a secondary structure sequence

CCEEECCCHHHEEEHHHHCCCCCHHHCCEEEEEEC

its reduced segmentsequence is βαβααβ, in which the α-helices and β-strands arelargely interspersed, suggesting that the corresponding proteinmore likely belongs to the α/β class rather than α+β class.

The transition probability matrix (TPM) of the reduced segmentsequence can be defined as follows:

$$TPM = \begin{bmatrix} P_{\alpha\alpha} & P_{\alpha\beta} \\ P_{\beta\alpha} & P_{\beta\beta} \end{bmatrix} \quad (4)$$

They are computed by the following formula:

$$P_{ij} = \begin{cases} \frac{N_{ij}}{N_{i1}+N_{i2}}, & if N_{i1} + N_{i2} \neq 0 \\ 0, & if N_{i1} + N_{i2} = 0 \end{cases} \quad (5)$$

where$\boldsymbol{N_{ij}}$ represents the number of transitions from the ith element, to the jth element of state space {α,β}.
In order to measure the degree of segment aggregation, wechoose $\boldsymbol{P_{\alpha\beta}}and\boldsymbol{P_{\beta\alpha}}$, the two of the above four transition probabilitiesto be included into our feature set. Note that

$$P_{\alpha\alpha} + P_{\alpha\beta} = 1, \ and P_{\beta\alpha} + P_{\beta\beta} = 1. \quad (6)$$

The third feature to be extracted isthe probability of helixor strand segments occurring in a segmentsequence, denoted by $\boldsymbol{p_\alpha}$or $\boldsymbol{p_\beta}$. Only $\boldsymbol{p_\alpha}$ is used in this workdue to $\boldsymbol{p_\alpha}$+$\boldsymbol{p_\beta}$= 1.

RESULTS

The prediction method was examined with 25PDB, and 30FB datasets.Our results on the two datasets with these features are as in Table 3.

Table 3. Comparison of prediction accuracies among various different structural class prediction methods.

Data	Method	Accuracy (%)				
		α	β	α/β	α+β	O.all
25PDB	SVM	85.3	83.2	69.7	58.3	74.1
	ANN	90.2	84.1	74.0	64.4	78.2
	EUCLID	85.3	80.0	44.9	71.7	70.5
	MAHAL	81.8	72.7	45.5	81.8	70.5
	CAMBE	88.0	75.3	70,5	81.9	78.9
	MANH	89.4	80.3	72.3	54.2	74.1
30FB	SVM	72.5	69.0	60.2	46.1	62.0

Kong, and Zhang (2014) Features

The secondary structure sequences (SSS)which can be obtained fromproteinstructure prediction server PSIPRED (Jones, 1999) consist of three secondary structural elements: H(helix), E(strand) and C(coil).

In *Kong, and Zhang (2014)* for the predicted secondarystructural elements of a given protein sequence, another two simplified sequences are proposed. One sequence is a segment sequence (SS), which is composedof helix segments and strand segments (Yang, et. al., 2010; Zhang, et. al., 2011; Zhang, et. al., 2013). First, every H, E and C segment in SSS is respectively replaced by the individual lettersH, E and C. Then, all of the letters C are removed and SS is obtained. Theother sequence is obtained by removing all of the letters C fromSSS, andthe new sequence is denoted by E-H (Ding, et. al., 2012). For example, given a secondarystructure sequence

SSS:
EECEEECCEECCCCHHHHCCHHHCCCEEECCHHHC

the corresponding SS and E-H are

EEEHHEH

and

EEEEEEEHHHHHHHEEEHHHE,

respectively. Based on the above three sequences,several structure-driven features are rationally constructed.

The details of these features are given as follows:

Content-Related Structure-Driven Features

1. The contents of secondary structure elements are the most widelyusedstructure-driven features, and have been proved significantlyhelpful in improving prediction accuracy of protein structural class (Kurgan, et. al., 2008a, b). They are formulated as:

$$p(x) = \frac{N(x)}{N_1}, \ x \in \{H, E, C\} \quad (7)$$

where N(x) is the number of secondary structural element H, E or C inSSS; N_1 denotes the sequence length of SSS. This type of features hasbeen extended to SS (Mizantry, and Kurgan, 2009). Here we further reuse them in E-H.This feature is extended to SS (Yang 2010) and to E-H (Kong, and Zhang 2014).

2. Biosequence patterns usually reflect some important functional orstructural elements in biosequences such as repeated patterns (Chen, and Liu, 2013).

In SSS, the 2-symbol repeated patterns are considered here, such asHH, EE, HE and EH. Hence, the contents of repeated patterns are proposedas follows:

$$p(xy) = \frac{N(xy)}{N_1},\ xy \in \{HH, EE, HE, EH\} \quad (8)$$

where N(xy) is the number of 2-symbol repeated patterns*HH, EE, HE*or *EH*. These features are extended to E-H and SS sequence. (Kong, and Zhang 2014).

3. The normalized counts of α-helices and β-strands in SSS (Ding, et. al., 2012), another

important structure-driven features, are given below:

$$NCountSeg(x) = CountSeg/N_1,\ x \in \{H, E\} \quad (9)$$

where*CountSeg(x)* is the number of H or E segments. These featureshave been reused in E-H (Ding, et. al., 2012). Here we further extended to SS.

The 25 features shown above characterize the contents of the predictedsecondary structure from different aspects. They can be categorizedinto content-related structure-driven features. Below, wewill further extract other types of structure-driven features such asorder-related and distance-related features.

Order-related Structure-driven Features

4. Second order composition moment of H, E and C are specially proposedto reflect the spatial arrangement of secondary structural elementsin SSS (Liu, and Jia, 2010), which are formulated as:

$$CMV(x) = \frac{\sum_{j=1}^{N(x)} n_{xj}}{N_1(N_1-1)},\ x \in \{H, E, C\} \quad (10)$$

wheren_{xj} is the jth order (or position) of the corresponding secondarystructural element in SSS. As these features reflect the orderrelatedcharacteristic of secondary structure, they can be categorizedinto order-related structure-driven features. This feature is reused in E-H (Ding, et. al., 2012)and in SS sequences(Kong, and Zhang 2014)

5. Classification of protein structures is based on the contents and spatialarrangements of secondary structural elements especially for theα/β and α+β classes. While proteins in the α/β and α+β classescontain both α-helices and β-strands, they are usually separated inthe α/β class but are usually interspersed in the α+β class. Thedistribution information of secondary structure segments will behelpful to inferring spatial arrangement of secondary structural elements.

As distance information of secondary structural elementscan reflect the distributions of α-helices and β-strands, we proposeseveral distance-related structure-driven features. The length ofα-helices or β-strands can be considered as a type of distance inthe same secondary structural segment. Thus normalized maximal,minimal and average lengths of secondary structural segments andvariance of α-helices (β-strands) lengths are proposed as follows:

$$NMaxSeg(x) = \frac{MaxSeg(x)}{N_1}, \quad (11)$$

$$NMinSeg(x) = \frac{MinSeg(x)}{N_1}, \quad (12)$$

$$NAvgSeg(x) = \frac{AvgSeg(x)}{N_1}, \quad (13)$$

$$NVarSeg(x) = \frac{VarSeg(x)}{N_1}, \quad (14)$$

where x ∈ {H, E}, MaxSeg(x) and MinSeg(x) are the lengths of thelongest and shortest α-helices (β-strands) and AvgSeg(x) andVarSeg(x) denote the mean and variance of lengths of α-helices(β-strands), respectively. Similarly, we consider the distance betweenthe same secondary structural segment, and the features aredefined as:

$$NMaxD(x) = \frac{MaxD(x)}{N_1}, \quad (15)$$

$$NMinD(x) = \frac{MinxD(x)}{N_1}, \quad (16)$$

$$NAvgD(x) = \frac{AvgD(x)}{N_1}, \quad (17)$$

$$NVarD(x) = \frac{VarD(x)}{N_1}, \quad (18)$$

where x ∈ {H, E}, MaxD(x) and MinD(x) are the maximal and minimaldistances between adjacent α-helices (β-strands) and AvgD(x)and VarD(x) denote the mean and variance of distances between adjacentα-helices (β-strands), respectively. In addition, the distancebetween different secondary structural segments is further considered.

The normalized maximal, minimal and average, variance ofthe distances between adjacent α-helices and β-strands are computedby the following formulas:

$$NMaxD(xx) = \frac{MaxD(xx)}{N_1}, \quad (15)$$

$$NMinD(xx) = \frac{MinxD(xx)}{N_1}, \quad (16)$$

$$AvgD(xx) = \frac{AvgD(xx)}{N_1}, \quad (17)$$

$$NVarD(xx) = \frac{VarD(xx)}{N_1}, \quad (18)$$

where xx ∈ {HE, EH}; HE denotes segment from α-helices to the adjacentβ-strands, and EH denotes segment from β-strands to theadjacent α-helices. As there are only letters H and E in SS and E-H,the similar features of NMaxD(xx), NMinD(xx), NAvgD(xx) andNVarD(xx) in SS and E-H are always 0. Hence, we only extend another8 distance-related structure-driven features (Eqs. (5)–(12)) to SSand E-H.

A total of 88 structure-driven features are proposed here. Amongthese features, 25 of them belong to content-related

features, 7 ofthem belong to order-related features, and 56 of them are distancerelated.

In addition, 56 out of 88 features are first proposed in Kong, and Zhang (Kong, and Zhang, 2014).

RESULTS

The prediction method by Kong, and Zhang (Kong, and Zhang,2014) is examined by them with four datasets in Table 4

Table 4. Comparison of the accuracies between twofeature sets by Kong, and Zhang (2014) that include 27 features and only 12reused features.

Data	Features	Accuracy (%)				
		α	β	α/β	α+β	O.all
25PDB	All	94.1	87.1	84.1	74.4	85.0
	Reused	91.4	82.8	82.7	72.8	82.4
1189	All	93.7	86.1	86.8	73.9	85.2
	Reused	80.2	84.0	81.1	62.7	79.5
640	All	91.3	77.3	91.5	76.0	83.9
	Reused	87.7	76.0	88.7	73.7	81.4
PC899	All	96.9	94.8	97.1	78.0	94.5
	Reused	92.3	93.3	96.8	69.5	92.4

The prediction method was examined with 25PDB, and 30FB datasets by authors of this review.Results on the two datasets with these features are as in Table 5.

Table 5
Accuracies in 25PDB. Only transition matrix is used by authors.

Data	Features	Accuracy (%)				
		α	β	α/β	α+β	O.all
25PDB	Transition	86	83	48	61	68.3

5. DISCUSSION

Kong, and Zhang (2014) introduced a novel computationalmethod forpredicting protein structural class solely using the predicted secondarystructure information. The 27 structure-driven featureswhich are rationallydivided into three groups (CR, OR and DR) are extracted to reflectgeneral contents and spatial arrangements of the predicted secondarystructural elements of a given protein sequence. Based on a comprehensivecomparison with other existingmethods on four widely-used lowhomologybenchmark datasets, the proposedmethod is shown to be aneffective computational tool for protein structural class prediction. Asfor the intrinsically disordered proteins which contain regions with nostable structure and may have specific sequence characteristics, itwould be more difficult to predict their structural class. Therefore,investigations about how the proposed method performs on the lowsimilarityaswell as disordered protein datasetswill constitute an interestingsubject for future work.

6. CONCLUSION

In this review article, three novel methods for protein structural classprediction arestudied. In Kong, and Zhang (Kong, and Zhang, 2014), not only the overall prediction accuracybut also the accuracies for proteins from α+βand α/β classes arehigher than the other two methods. Furthermore, rationaldesign on the basis of protein spatial structural information isproved to be a successful approach to obtain new features and,consequently, to improve the prediction accuracy.

REFERENCES

Anand A., Pugalenthi G, Suganthan PN, (2008) Predicting protein structural class by SVM with class-wise optimized features and decision probabilities, J. Theor. Biol. 253,pp. 375-380.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., Murzin, A.G.,(2004) SCOP database in 2004: refinements integrate structure and sequencefamily data. Nucleic Acids Res. 32, pp. 226–229.

Cai, D., and Zhou, G.P. (2000) Prediction of protein structural classes by neural network, Biochimie 82, pp. 783-785.

Cai, D., Liu, X. J., Xu, X., and Zhou, G.P. (2001) Support vector machines for predicting protein structural class, BMC Bioinform. 2,3.

Cai, D., Liu, X. J., Xu, X. B., and Chou, K. C. (2002) Prediction of protein structural classes by support vector machines, Comput. Chem. 26,pp. 293-296.

Y.D. Cai,Y.D. , Feng, K. Y. Lu, W.C., and Chou, K. C. (2006) Using LogitBoost classifier to predictprotein structural classes, J. Theor. Biol. 238 172-176.

Cao,Y.F.,Liu, S., Zhang, L.D., Qin, J., Wang, J., and Tang, K. X. (2006) Prediction of protein structural class with rough sets, BMC Bioinform. 7, 20.

Chen, L., Liu, W.(2013) Frequent patterns mining inmultiple biological sequences, Comput.Biol. Med. 43, 1444–1452.

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

Chen,K.,Kurgan,L.A., andRuan, J.S. (2008)Prediction of protein structural class using novel evolutionary collocation-based sequence representation, J. Comput. Chem. 29,pp. 1596-1604.

K.C. Chou, A key driving force in determination of protein structural classes, Biochem. Biophys. Res. Commun. 264 (1999) 216e224.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins 43,pp. 246-255.

Chou, K.C., 2004. Structural bioinformatics and its impact to biomedical science.Curr. Med.Chem. 11, 2105–2134.

Chou, K.C., (2005) Progress in protein structural class prediction and its impact tobioinformatics and proteomics. Curr. Protein Pept. Sci. 6, pp. 423–436.

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

Costantini, S., Facchiano, A. M.(2009) Prediction of the protein structural class byspecific peptide frequencies. Biochimie 91, pp. 226–229.

Deschavanne, P.,Tuffery, P. (2008) Exploring an alignm-ent free approach for protein classification and structural class prediction, Biochimie 90,pp. 615-625.

Ding, S., Zhang, S.,Li,Y., andWang, T. (2012) A novel protein structural classes prediction method based on predicted secondary structure, Biochimie 94,pp. 1166–1171.

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

Jin,L.X., Fang, W.W., and Tang, H. W. (2003) Predict-ion of protein structural classes by a new measure of information discrepancy, Comput. Biol. Chem. 27,pp. 373-380.

Jones, D.T.(1999) Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 292(2),pp. 195-202.

Kedarisetti, K.D., Kurgan, L. A., and Dick, S. (2006) A comment on e Prediction of protein structural classes by a new measure of information discrepancy, Comput. Biol.Chem. 30,pp. 393-394.

Kong, L., and Zhang, L.(2014)Novel structure-driven features for accurate prediction of protein structural class, Genomics 103,pp. 292–297.

Kurgan, L. A., Homaeian, L., (2006) Prediction of structural classes for proteinsequences and domains—impact of prediction algorithms, sequence representationand homology, and test procedures on accuracy. Pattern Recognit.39, pp. 2323–2343.

Kurgan, L.A., Chen, K., 2007. Prediction of protein structural class for the twilightzone sequences. Biochem. Biophys. Res. Commun. 357, pp. 453–460.

Kurgan, L.A., Zhang, T., Zhang, H., Shen, S., Ruan, J. (2008a) Secondary structurebasedassignment of the protein structural classes. Amino Acids 35, pp. 551–564.

Kurgan, L.A., Cios, K., Chen, K., (2008b) SCPRED: accurate prediction of proteinstructural class for sequences of twilight-zone similarity with predictingsequences. BMC Bioinformat. 9, 226.

Lin,H.,Li,Q.Z. (2007) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components, J. Comput. Chem. 28,pp. 1463-1466.

Liu T., and Jia C., A (2010) high-accuracy protein structural class prediction algorithm using predicted secondary structural information, Journal of Theoretical Biology 267, 272–275.

Liu,T.,Zheng,X., and Wang, J. (2010) Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile, Biochimie 92,pp. 1330-1334.

Luo, R.Y., Feng,Z.P., and Liu,J.K. (2002) Prediction of protein structural class by amino acid and polypeptide composition, Eur. J. Biochem. 269, pp4219-4225.

Mizianty, M.J., Kurgan, L.A. (2009) Modular prediction of protein structural classesfrom sequences of twilight-zone identity with predicting sequences. BMCBioinformat. 10, 414.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., (1995) SCOP: a structuralclassification of proteins database for the investigation of sequences andstructures. J. Mol. Biol. 247, pp. 536–540.

Nakashima, H., Nishikawa, K.,Ooi, T. (1986) The folding type of a protein is relevant to the amino acid composition, J. Biochem. 99,pp. 153-162.

Qiu,J.D., Luo, S.H.,Huang, J.H.,and Liang,R.P.(2009) Using support vector machines forprediction of protein structural classes based on discrete wavelet transform,J. Comput. Chem. 30 pp. 1344-1350.

Shen, H.B.,Yang, J., Liu,X.B., and Chou,K.C., (2005) Using supervised fuzzy clustering topredict protein structural classes, Biochem. Biophys. Res. Commun. 334, pp. 577-581.

X.D. Sun, R.B. Huang, Prediction of protein structural classes using support vector machines, Amino Acids 30 (2006) 469e475.

Wang, Z. X.,and Yuan, X.(2000) How good is prediction of protein structural class by thecomponent-coupled method? Proteins 38,pp. 165-175.

Xiao,X., Shao, S.H.,Huang, Z.D., andChou,K.C.(2006) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor, J. Comput. Chem. 27,pp. 478-482.

Yang, J.Y.,Peng,Z.L., Yu,Z.G., Zhang,R.J., Anh, V.,and Wang, D. S. (2009) Prediction of protein structural classes by recurrence quantification analysis based on chaos gamerepresentation, J. Theor. Biol. 257,pp. 618-626.

Yang,J., Peng, Z., and Chen, X. (2010) Prediction of protein structural classes for low-homology sequences based on predicted secondary structure, BMC Bioinforma. 11,pp. 59.

Yuan, Z., Bailey, T.L., Teasdale, R.D., (2005) Prediction of protein B-factor profiles.Proteins 58, pp. 905–912.

Zhang,T. L., Ding, Y. S., and Chou, K. C. (2008) Prediction protein structural classes with pseudo-amino

acid composition: approximate entropy and hydrophobicitypattern, J. Theor. Biol. 250, pp.186-193.

Zhang, S., Ding, S., and K.C. Wang, T.(2011) High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure, Biochimie 93,pp. 710-714

Zhang,T.L.,and Ding,Y.S. (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes, Amino Acids 33 pp. 623-629.

Zhang, S. Yang, L, and Wang, T. (2009)Use of information discrepancy measure to compareprotein secondary structures, J. Mol. Struct. Theochem. 909,pp. 102-106.

Zhang, L., Zhao, X., and Kong L., (2013) A protein structural class prediction method based on novel features, Biochimie 95,pp. 1741–1744.

Zhou, P.(1998) An intriguing controversy over protein structural class prediction, J. Protein Chem. 17,pp. 729-738.

.