Journal of Engineering and Applied Sciences 11 (3): 430-435, 2016
ISSN: 1816-949X

An Improved Method for Predicting Protein Structure Classes by Incorporating Hydropathical and Secondary Information Based on Feature Selection Technique

[1]Mohammed Hasan Aldulaimi, [2]Suhaila Zainudin and [2]Azuraliza Abu Bakar
[1]Data Mining and Optimization Research Group (DMO),
Centre for Artificial Intelligence Technology (CAIT), Al-Jazaaer High School,
General Directorate of Education (GDE), Babylon, Iraq
[2]School of Computer Science, Faculty of Information Science and Technology,
Malyasia Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia

Abstract: The prediction of the structural classes of proteins is an important classification issue in bioinformatics research. Knowledge of these classes will give a clear understanding of the protein folding process. For this reason, research into the prediction of protein classes has become a major topic of concern. This research intends to discuss new development of features based on secondary structures information of proteins and hydropathy profile that categorized proteins into all-α, all-β, α/β and α+β with each category being vital in pinpointing the proteins' structural classes. The folding patterns, functions and interactions between proteins is reliant upon the accurate prediction of its structural classes. This is especially true if one intend to synthesize new proteins possessing unique functionalities. This is however a complex undertaking, especially for structural classes of low-similarity sequences. There are a few computational methods being developed for this purpose (25-40%). The accuracy of the proposed method is on par with current methods being reported in literature.

Key wards: Protien, reason, hydropathy, profile, unique

INTRODUCTION

Levitt pioneered the classification of protein structures. Being aware of structural classes will decrease the search space for the conformations of tertiary structures (Chou and Zhang, 1995), help to analyzeproteins functions, drug designing (Zhou, 2001). Generally, globular protein domains are divided into all-α, all-β,α+β and α/β, based on the types and configuration of their respective secondary structural elements. Due to this fact, previous works proposed multiple structures of protein domain classification methods that are derived from protein sequences; however, information pertaining to this approach remains limited.

Structural Classification of Proteins (SCOP) (Murzin *et al*., 1995) is a manually noted database which is considered as the best classification approach of the structural classes of proteins. Its latest version encompasses 11 structural classes with ~90% of the domains of protein belonging to 4 main classes (all-α, all-β, α/β, α+β). With the fast increase of the proteomics and genomics, current experimental methods are regarded as being complex, time-consuming and face many limitations in determining protein structures. (X-ray crystallography, NMR "Nuclear Magnetic Resonance" and ESR "Electron Spin Resonance". It is therefore salient that anaccurate and fast computational method be developed to confirm the structural class for newly discovered proteins. These efforts are generally divided into: feature vector and classification algorithm. Current approaches have been extensively studied by (Chou, 2005; Kurgan *et al*., 2008).

In previous researh, the Amino Acid (AA) sequence always served as a platform of feature extraction (Feng *et al*., 2005; Chou and Cai, 2004; Wu *et al*., 2010; Chou, 2001). Recently, features based on predicted Secondary Structure Sequence (SSS) were proposed for the purpose of improving the prediction accuracy of low-similarity sequences (Mizianty and Kurgan, 2009; Kong *et al*., 2014) such as the length of the longest α-helices and β-strands. Feature vectors extracted from protein sequences are subsequently used as inputs into

Corresponding Author: Mohammed Hasan Aldulaimi, Data Mining and Optimization Research Group (DMO), Centre for Artificial Intelligence Technology (CAIT), Al-Jazaaer High School, General Directorate of Education, (GDE), Babylon, Iraq

multiple sets of machine learning techniques. The existing structure-driven features can be generally classified into 3 varieties content-related order and distance features. Despite the success achieved with predicted secondary structure based methods, the development of high quality prediction methods for low-similarity structures remains a challenge. This study explores current features based on the secondary structure prediction and hydropathy in the production of classifiers that will be as accurate as the new models via the utilization of lesser amounts of features.

MATERIALS AND METHODS

Data sets: In this study, three low-homology datasets ($ASTRAL_{training}$, $ASTRAL_{testing}$, 640) are used to evaluate and design the suggested method. All the three datasets have been commonlyused as standard datasets in previous studies (Kurgan *et al*., 2008; Kong *et al*., 2014; Zhang *et al*., 2011; Liu and Jia, 2010; Mizianty and Kurgan, 2009; Yang *et al*., 2010; Ding *et al*., 2012). More details of these datasets are shown in Table 1.

Generating features to represent the protein: To be used efficiently in proposed method, each amino acid in the sequence of proteinmust to be first converted into one of the three following elements ofsecondary structure: H (Helix), E (strand) and C (Coil). The sequence of elements of secondary structure is also known as protein Secondary Structure Sequence (SSS) which can be acquired from the server of Protein Structure Prediction PSIPRED (Jones, 1999). In order to reveal the general contents and spatial arrangements of the predicting the elements secondary structure of a known protein sequence particularly for α-helix and β-strand, another two simplified sequences are proposed based on SSS. One sequence is a Segment Sequence (SS) which is composed of helix segments and strand segments (Yang *et al*., 2010; Zhang *et al*., 2011; Ding *et al*., 2012; Firdaus and Harley, 2013).

First, every H, E and C segment in SSS is respectively replaced by the individual letters H, E and C. Then, all of the letters C are removed and SS is obtained. The other sequence is obtained by removing all of the letters C from SSS and the new sequence is denoted by E_H (Kong *et al*., 2014). For example, given a secondary structure sequence SSS: EECEEECCEECCCCHHHHCCHHHCCCEEEC CHHHCEE, the corresponding SS and E_H are EEEHHEHE and EEEEEEEEHHHHHHHHEEEHHHEE, respectively. Based on the above three sequences, several features are rationally constructed. The details of these features are given as follows:

Table 1: Details of the datasets

Dataset	all-α	all-β	α/β	α+β	Total
ASTRAL train	640	662	748	763	2813
ASTRAL test	640	662	747	764	2813
640	138	154	177	171	640

The contents of secondary structure elements are the most widely used structure-driven features and have been proved significantly helpful in improving prediction accuracy of protein structural class (Mizianty and Kurgan, 2009). They are formulated as:

$$p_{(i)} = \frac{n_i}{N} \quad (1)$$

Where:

n_i = Total number of occurrences of secondary structural state, i in the amino acids sequence with each $i \in \{E, H, C\}$ and

N = Sequence length of SSS

This type of features has been extended to SS (Yang *et al*., 2010). Here we further reuse them in E_H. Biosequence patterns usually reflect some important functional or structural elements in biosequences such as repeated patterns (Meysman *et al*., 2015). In SSS, the 2-symbol repeated patterns are considered here such as HH, EE, HE and EH. Since, the predicted states H and E to alternate more frequently in a protein belonging to the α/β-class than in a protein belonging to the α+β class where α-helix and β-strands are isolated (Murzin *et al*., 1995). Therefore, one of their newly developed features was the normalized alternating frequency of HE and EH in E_H. Hence, the contents of repeated patterns are proposed as follows:

$$p_{(xy)} = \frac{n_{xy}}{N} \quad (2)$$

where, n_{xy} is the number of two symbol repeated patterns, $xy \in \{EH, HE\}$. Here, we extended these features to SS and E_H.

The normalized calculations of α-helices and β-strands in SSS (12), another important structure-driven features are given as:

$$NcalcSeg_{(i)} = \frac{calcSeg_{(i)}}{N} \quad (3)$$

where, $calcSeg_{(i)}$ is the count of H or E segments, $i \in \{E, H\}$. These features have been reused in E_H (Kong *et al*., 2014). Here, we further extended to SS.

The composition moment vector, CMV, expresses both the position and the secondary structural state composition in the predicted sequence of the secondary structure. The first order composition moment vectors for the secondary structural state component α-helix (H) and β-strand (E) were calculated by:

$$CMV_i = \frac{1}{N_{(N-1)}} \Sigma_{j=1}^{ni} x_{ij} \quad (4)$$

where, i∈{E, H} and N is the number of amino acids in the sequence (length) for the protein and X_{ij} is the index of the jth position of the i_structure.

The length of α-helices or β-strands can be regarded as types of distances within similar secondary structural segments. Thus, normalized maximal, minimal and average lengths of secondary structural segments and variance of α-helices (β-strands) lengths are proposed as follows:

$$NMaxSeg_i = \frac{MaxSegi}{N} \quad (5)$$

$$NMinSeg_i = \frac{MinSeg_i}{N} \quad (6)$$

$$NAvgSeg_i = \frac{AvgSeg_i}{N} \quad (7)$$

$$NVarSeg_i = \frac{VarSeg_i}{N} \quad (8)$$

Where i∈{E, H}, $MaxSeg_i$ and $MinSeg_i$ are the lengths of the longest and shortest α-helices (β-strands) and $AvgSeg_i$ and $VarSeg_i$ denote the mean and variance of lengths of α-helices (β-strands), respectively. Below, we will further extract other type's features which are the hydopathy features.

Features set created from hydropathy and secondary structure information: The properties of physiochemical of amino acid have an important effect on the establishment of protein structures. Several physiochemical properties of amino acids like polarity, isoelectric points, hydropathy, flexibility, etc., have been used to predict structural classes (Nanni *et al*., 2014). For the proposed method, the hydropathy profile of the protein sequence was selected based on the assumption that it had a major influence on the folding of the protein. The hydropathy profile defines the hydrophilic and hydrophobic nature of the segments of a protein based on the primary structure of the protein (Liu and Wang, 2006) classified the twenty amino acids of proteins into three groups based on their hydropathy profile: External (E), Internal (I) and Ambivalent (A). The following rules by Liu and Wang (2006) were used in this study to classify amino acids according to their hydropathy profile:

$$F(S_{(i)}) = \begin{cases} lifS_{(i)} = F, I, L, M, V \\ E_iFS_{(i)} = D, E, G, K, N, Q, R \\ A_iFS_{(i)} = S, T, Y, C, W, G, P, A \end{cases} \quad (9)$$

Here, $S_{(i)}$ denotes the ith amino acid in the primary sequence of the protein and $F(S_{(i)})$ denotes its consistent replacement according to its hydropathic nature. For example, an amino acid sequence for a protein:

S = MDPFLVLLHSVSS is denoted by F(S) = IEAIIIIIEAIAAA

Using Eq. 1-8 for each protein sequence in the dataset, ith {I, E, A} were extracted hydropathic features and then combined with the (SSS) features.

Feature selection: Feature selection is defined as the approach of pinpointing and eliminating the majority of irrelevant and redundant features. This will result in increased efficiency of the model and faster computational analysis. Many feature selection methods were utilized in bioinformatics studies (Saeys *et al*., 2007) and can be divided into: filter and wrapper. Due to the combination of the feature selection method and a classifier, feature wrappers performed better than other filters. Thus, a wrapper approach based on the best first search algorithm was used to select a subset of the original features in this work. The 10 fold cross-validation on the ASTRAL training dataset with an SVM classifier was used in order to prevent over-fitting. Finally, a 23-dimensional features vector was constructed using the aforementioned features which combines information on content, position of the predicted secondary structural elements and hydropathy.

Classification algorithm: The Support Vector Machine (SVM) method is one of the most common programming techniques used in state-of-the-art systems to resolve classification problems associated with protein structures and it is used by many bioinformatics researchers as well (Kedarisetti *et al*., 2006; Dehzangi *et al*., 2013). The parameters of the regularization, C and the kernel parameter γ were adjusted based on the 10 fold cross-validation on the ASTRAL training dataset. Finally, the best values for the parameters were obtained when C = 1024 and γ = 0.25 which were selected using the grid search approach available in the LIBSVM Software.

Performance measures: In predictions using statistical methods, the 10 fold Cross Validation (10-CV) test is commonly used to ensure the statistical validity of a classifier (Liu *et al*., 2012). It was also employed to calculate the efficiency in this study. For the evaluation, the individual sensitivity (sensitivity) with The Overall prediction accuracy (OA) over the entire dataset was reported. They were defined as follows (Zhang *et al*., 2011):

$$\text{Sens}_j = \frac{TP_j}{(TP_j+FN_j)} = \frac{TP_j}{|C_j|} \quad (10)$$

$$\text{Spec}_j = \frac{TN_j}{(FP_j + TN_j)} = \frac{TN_j}{\sum_{k\neq j}|C_j|} \quad (11)$$

$$MCC_j = \frac{(TP_j*TN_j - FP_j*FN_j)}{\sqrt{(FP_j+TN_j)(TP_j+FN_j)(TN_j+FP_j)(TP_j+FN_j)}} \quad (12)$$

$$OA = \frac{(\Sigma_j TP_j)}{(\Sigma_j |C_j|)} \quad (13)$$

where, TN_j, TP_j, FN_j, FP_j and $|C_j|$ are the number of true negatives, true positives, false negatives, false positives and proteins in the structural class C_j, respectively.

RESULTS AND DISCUSSION

Structural class prediction accuracy: The 10-CV test was performed on all the benchmark datasets to evaluate and compare the proposed classification method to 8 previous methods. As mentioned earlier, the aim of the proposed method is to improve the accuracy of predictions. In order to show that experiments were performed using the 10-CV test with 23 features and only 10 reused features and the results are presented in Table 2 and 3. According to this tables it is clear that all the prediction accuracies improved after the addition of the hydropathy features (new features).

According to Table 2, the Sens, Spec and MCC values in all-α class were the best for all datasets while the values in α+β class were the lowest for example the MCC was only 90.6% in the ASTRAL training dataset. This implied that the former was the easiest to predict and the latter was the most difficult to identify.

Table 2:The prediction quality of our method on four datasets by 10-CV test

Dataset	Class	Sens (%)	Spec (%)	MCC (%)
$ASTRAL_{training}$	all-α	93.40	97.60	90.60
	all-β	80.50	96.40	79.17
	α/β	84.80	90.80	75.00
	α+β	72.30	89.00	60.90
	OA	82.40	-	-
$ASTRAL_{testing}$	all-α	94.80	98.61	93.60
	all-β	82.90	59.90	79.90
	α/β	86.80	90.40	76.12
	α+β	72.08	90.50	63.12
	OA	83.70	-	-
640	all-α	89.06	96.90	87.50
	all-β	79.00	95.11	79.30
	α/β	82.20	95.00	85.05
	α+β	87.60	93.07	70.90
	OA	84.50	-	-

Analysis and comparison with other prediction methods: As mentioned earlier, the hydrpathy features were aimed to improve the accuracies. In order to show their the contribution, the experiments were performed by the jackknife test on all mentioned datasets with 23 features and only the 10 reused features and the results were given in Table 3. According to Table 3, it was obvious that all prediction accuracies were improved after adding the novel features.

This method is compared with 4 previously published methods including the famous methods SCPRED (Kurgan *et al*., 2008) and MODAS (Mizianty and Kurgan, 2009) which are often used as a baseline for comparison. We also compared with some competing structural class prediction methods (Ding *et al*., 2012; Zhang *et al*., 2014). As shown in Table 3, the highest overall accuracies were obtained by the proposed method among all the tested methods in $ASTRAL_{training}$, $ASTRAL_{testing}$ and 640 datasets (82.4, 83.7, 84.5%) and improved by 0.6, 1.01 and 0.28% compared with previous best-performance results. As for $ASTRAL_{training}$ the α/β class accuracy was 5.25% higher than (Zhang *et al*., 2014). As for ASTRALtesting, the all-β, α/β class accuracies were 2.2 and 2.86% higher than (Zhang *et al*., 2014). As for the 640 dataset, the α+β was 13.33% higher than the

Table 3: Performance comparison of different methods on four datasets

		Sensitivity (%)				
Dataset	Reference	all-α	all-β	α/β	α+β	OA
$ASTRAL_{training}$	Experimental study	93.40	80.50	84.80	72.30	82.40
	(28)	94.06	81.72	79.55	73.79	81.80
$ASTRAL_{testing}$	Experimental study	94.80	82.90	86.80	72.08	83.70
	(13)	93.13	78.33	83.38	64.27	79.14
	(17)	94.53	77.49	87.28	71.47	82.33
	(28)	95.16	80.70	83.94	72.51	82.69
640	Experimental study	89.06	79.00	82.20	87.6	84.50
	(13)	90.60	81.80	85.90	66.70	80.80
	(11)	89.10	85.10	88.10	71.40	83.10
	(17)	94.93	76.62	89.27	74.27	83.44
	(28)	92.75	81.82	89.27	74.27	84.22

previous best-performance results (Zhang *et al*., 2014). The α/β and α+β class accuracies were satisfactory and the same as the previous best value (84.8-87.6%).

Some results were lesser to the best among the compared methods. This was partly because the features we used ignored some less common secondary structural elements such as β-turns and because the real structures of proteinsare much more complex than the theoretical model (Zhang *et al*., 2014). Although, the improvements looked small, the mean was significant to identify the protein structural class. For example, only around 38,221 PDB entries with 110,800 domains or proteins had known structural class.

Labels in SCOP (as of February, 2009) while there were >8,000,000 no redundant protein sequences in the Protein database at the National Center for Biotechnology Information (NCBI). Hence, 0.1% improvement in accuracy could help in finding the accurate structural class labels for about 8000 proteins. These prediction improvements hence clearly demonstrated that our method was very promising for recognizing protein structural class.

CONCLUSION

In this study, we proposed an accurate method that allows us to predict protein class structures. The new method relies on a 23-dimensional integrated feature vector which is the result of the combination of the information contained in both the content and position in SSS and hydropathy. The consistent results from the 10-CV tests showed that the proposed method is dependable in the case of low-similarity datasets. The proposed sequence representation is made up of 13 new features that are based upon hydropathy information which resulted in satisfactory prediction accuracy when compared to previous methods. This is attributed to the fact that hydropathy information is capable of pinpointing the link between the sequence and the protein structural class. The results reported in this research proved that the proposed method is indeed a viable tool for protein structural class prediction, especially for low similarity sequences. In future research we would like to optimize thesame features sets using one of meta-heuristic techniques such as genetic algorithm in order to get more accurate rate classification.

ACKNOWLEDGEMENT

This research was supported by the Universiti Kebangsaan Malaysia (UKM) under Grant (FRGS/1/2013/ICT07/UKM/02/3).

REFERENCES

Chou, K.C. and C.T. Zhang, 1995. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol., 30: 275-349.

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

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

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

Dehzangi, A., K. Paliwal, J. Lyons, A. Sharma and A. Sattar, 2013. Enhancing Protein Fold Prediction Accuracy Using Evolutionary and Structural Features. In: Pattern Recognition in Bioinformatics, Ngom, A., E. Formenti, J.K. Hao, X.M. Zhao and T. van Laarhoven (Eds.). Springer, Berlin, Heidelberg, ISBN: 978-3-642-39158-3, pp: 196-207.

Ding, S., S. Zhang, Y. Li and T. Wang, 2012. A novel protein structural classes prediction method based on predicted secondary structure. Biochimie, 94: 1166-1171.

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

Firdaus, S.N. and E. Harley, 2013. Protein structural class prediction using predicted secondary structure and hydropathy profile. Proceedings of the International C* Conference on Computer Science and Software Engineering, July 10-12, 2013, Porto, Portugal, pp: 49-57.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol., 292: 195-202.

Kedarisetti, K.D., L. Kurgan and S. Dick, 2006. Classifier ensembles for protein structural class prediction with varying homology. Biochem. Biophys. Res. Commun., 348: 981-988.

Kong, L., L. Zhang and J. Lv, 2014. Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. J. Theor. Biol., 344: 12-18.

Kurgan, L., K. Cios and K. Chen, 2008. SCPRED: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. BMC Bioinformatics, Vol. 9. 10.1186/1471-2105-9-226.

Kurgan, L.A., T. Zhang, H. Zhang, S. Shen and J. Ruan, 2008. Secondary structure-based assignment of the protein structural classes. Amino Acids, 35: 551-564.

Liu, N. and T. Wang, 2006. Protein-based phylogenetic analysis by using hydropathy profile of amino acids. FEBS Lett., 580: 5321-5327.

Liu, T. and C. Jia, 2010. A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. J. Theor. Biol., 267: 272-275.

Liu, T., X. Geng, X. Zheng, R. Li and J. Wang, 2012. Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. Amino Acids, 42: 2243-2249.

Meysman, P., C. Zhou, B. Cule, B. Goethals and K. Laukens, 2015. Mining the entire protein DataBank for frequent spatially cohesive amino acid patterns. BioData Min., Vol. 8. 10.1186/s13040-015-0038-4.

Mizianty, M.J. and L. Kurgan, 2009. Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. BMC Bioinformatics, Vol. 10. 10.1186/1471-2105-10-414.

Murzin, A.G., S.E. Brenner, T. Hubbard and C. Chothia, 1995. Scop A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol., 247: 536-540.

Nanni, L., S. Brahnam and A. Lumini, 2014. Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. J. Theor. Biol., 360: 109-116.

Saeys, Y., I. Inza and P. Larranaga, 2007. A review of feature selection techniques in bioinformatics. Bioinformatics, 23: 2507-2517.

Wu, J., M.L. Li, L.Z. Yu and C. Wang, 2010. An ensemble classifier of support vector machines used to predict protein structural classes by fusing auto covariance and pseudo-amino acid composition. Protein J., 29: 62-67.

Yang, J.Y., Z.L. Peng and X. Chen, 2010. Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. BMC Bioinformatics, Vol. 11. 10.1186/1471-2105-11-S1-S9

Zhang, L., X. Zhao and L. Kong, 2014b. Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou× s pseudo amino acid composition. J. Theor. Biol., 355: 105-110.

Zhang, L., X. Zhao, L. Kong and S. Liu, 2014b. A novel predictor for protein structural class based on integrated information of the secondary structure sequence. Biochimie, 103: 131-136.

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

Zhou, G.P. and N. Assa-Munt, 2001. Some insights into protein structural class prediction. Proteins Struct. Funct. Genet., 44: 57-59.