Optimized Local Protein Structure with Support Vector Machine to Predict Protein Secondary Structure

Chin Yin Fai, Rohayanti Hassan, and Mohd Saberi Mohamad

Artificial Intelligence and Bioinformatics Research Group,
Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia,
81310 Skudai, Johor, Malaysia
{chinyinfai,rohayanti}@gmail.com, saberi@utm.my

Abstract. Protein includes many substances, such as enzymes, hormones and antibodies that are necessary for the organisms. Living cells are controlled by proteins and genes that interact through complex molecular pathways to achieve a specific function. These proteins have different shapes and structures which distinct them from each other. By having unique structures, only proteins able to carried out their function efficiently. Therefore, determination of protein structure is fundamental for the understanding of the cell's functions. The function of a protein is also largely determined by its structure. The importance of understanding protein structure has fueled the development of protein structure databases and prediction tools. Computational methods which were able to predict protein structure for the determination of protein function efficiently and accurately are in high demand. In this study, local protein structure with Support Vector Machine is proposed to predict protein secondary structure.

Keywords: Local Protein Structure, Support Vector Machine, Protein Secondary Structure Prediction.

1 Introduction

In recent years, human genome project has successfully generated tremendous amount of newly protein sequences in the biological database. Ironically, most of them are completely unknown in function and structure and cause complete genome sequencing gives much less understanding on the organism than initially hoped for [1]. Proteins control and mediate many of the biological activities of cells. Hence, to gain an understanding of cellular function, the structure of every protein must be understood [2]. This has shown that the study the sequence of a single protein or small complexes is no longer sufficient in helping the current genome development.

Protein structure predictions represent a key step in studying and understanding protein functions. The fact that protein function do not only depends on protein sequence but also the shape and structure induces the important goal of the proteomic studies which is identification of protein structure. Given a protein sequence, the

D. Lukose, A.R. Ahmad, and A. Suliman (Eds.): KTW 2011, CCIS 295, pp. 333–342, 2012.

secondary structure prediction problem is to predict whether each amino acid is in a helix, strand or neither. H, E and C represent helix, strand and non-routine structure, respectively [3]. The simple definition of secondary structures hides various limitations. The complexity of fundamentals for secondary structure assignments induce the creation of numerous assignment methods based on different criteria or characteristics. Due to certain limitation in secondary structure, a more precise assignment for secondary structure is presented which is local protein structure. Local protein structure is defined as the description of complete set of small prototype or protein structures. Analysis of local protein structures represents an evaluation of every parts of protein backbone. Hence, focusing on local protein structure might develop a new milestone in the future of protein secondary structure prediction.

The aim of this research is to predict protein secondary structure using machine learning algorithms based on RS126 as the dataset. RS126 is important as the core dataset to be trained and tested using machine learning algorithm because the dataset contains 126 non-redundant proteins where the number pairs of proteins in the set have more than 25% similarity over a length of 80 residues. Given the small similarity of the dataset sequences, this represents a situation that is rather close to real-world settings and it can be considered as the ideal environment for protein secondary structure prediction. The machine learning algorithm, implemented in this study is Support Vector Machine (SVM). The reason SVM is being used is because they are known to be a powerful algorithm for making binary decisions. The results will be able to show the higher accuracy of computational prediction system based on SVM for protein secondary structure prediction.

2 Materials and Methods

2.1 Materials

Materials briefly explain the dataset used and also the source of data such as the background of the dataset and how to obtain it. Details of dataset preparation and usage will be explained in the following section.

2.1.1 RS126 Dataset

The dataset used in this study is RS126. The initiation of the research is to obtain the protein sequence datasets in order to predict protein secondary structure.

RS126 is one of the oldest dataset with the longest history to evaluate for protein secondary structure prediction. The scheme is created by Rost and Sander [4]. RS126 being the most commonly used datasets to predict protein structure are applied in most of the study including this research. It contains 23,347 residues with an average protein sequence length of 185. 32% of RS126 are alpha helix, 21% as beta strand and 47% as coil.

RS126 dataset can be collected from various supplementary data files in previous research or study. Besides that, it can also be obtained from online database such as Protein Data Bank (PDB). Fig. 1 shows the list of RS126 dataset used in protein secondary structure prediction.

PDB ID	Chain	PDB ID	Chain	PDB ID	Chain	PDB ID	Chain	PDB ID	Chain
1BMV	1	2UTG	A	4SGB	I	1PYP	-	3CD4	-
4RHV	1	3GAP	A	1MCP	L	1RBP	-	3CLA	-
1BMV	2	3HMG	A	2OR1	L	1RHD	-	3CLN	-
1R09	2	3TIM	A	1GD1	O	1S01	-	3EBX	-
1LMB	3	4SDH	A	2TMV	P	1SH1	-	3ICB	-
4RHV	3	4TS1	A	2WRP	R	1UBQ	-	3PGM	-
2MEV	4	4XIA	A	5CYT	R	2AAT	-	3RNT	-
4RHV	4	5HVP	A	1ACX	-	2ALP	-	4BP2	-
1BBP	A	7CAT	A	1AZU	-	2CAB	-	4CMS	-
1CDT	A	9API	A	1BDS	-	2CYP	-	4CPV	-
1FXI	A	9WGA	A	1CBH	-	2FOX	-	4GR1	-
1GP1	A	1WSY	B	1CC5	-	2FXB	-	4PFK	-
1IL8	A	2LTN	B	1CRN	-	2GBP	-	4RXN	-
1OVO	A	2SOD	B	1ECA	-	2GCR	-	5LDH	-
1TNF	A	3HMG	B	1ETU	-	2GN5	-	5LYZ	-
1WSY	A	9API	B	1FDX	-	2I1B	-	6ACN	-
256B	A	9INS	B	1FKF	-	2LHB	-	6CPA	-
2AK3	A	1FC2	C	1FND	-	2MHU	-	6CPP	-
2CCY	A	5ER2	E	1GDJ	-	2PCY	-	6CTS	-
2GLS	A	6TMN	E	1HIP	-	2PHH	-	6DFR	-
2HMZ	A	1FDL	H	1L58	-	2SNS	-	6HIR	-
2LTN	A	1CSE	I	1LAP	-	2STV	-	7ICD	-
2PAB	A	1TGS	I	1MRT	-	3AIT	-	7RSA	-
2RSP	A	2TGP	I	1PAZ	-	3B5C	-	8ABP	-
2TSC	A	4CPA	I	1PPT	-	3BLM	-	8ADH	-
9PAP	-								

Fig. 1. List of RS126 dataset used to predict protein secondary structure

2.1.2 Dihedral Angle (DA)

Generally, dihedral angle is defined as the angle between two planes. In terms of proteomics, the backbone dihedral angles of proteins are called phi (φ), psi (ψ) and omega (ω). Every different angle has its own functions. Dihedral angle is used as feature vector in this research due to its nature form of representation, which is the numerical or integer form. Besides that, dihedral angles play a key role in defining or 'tightening' the secondary structure of protein structures during the structure refinement process. The importance of dihedral angle information tends to increase with the size of the protein being studied as the quality and quantity of other restraints.

In this study, all the dihedral angles are obtained through ramachandran function in Matlab. Ramachandran function generates the dihedral angle for the protein specified by the PDB database identifier PDBid. PDBid is a string specifying a unique identifier for a protein structure record in the PDB database. Each structure in the PDB database is represented by a four-character alphanumeric identifier. The PDBid is similar to the identifier of protein in RS126. For example, 4hhb is the identifier for hemoglobin. The results will return the dihedral angles for each protein in RS126 as 3 columns which include phi angle, psi angle and omega angle.

2.1.3 DSSP

The DSSP program was designed by Wolfgang Kabsch and Chris Sander as the standard method for assigning secondary structure to the amino acids of a protein, given the atomic-resolution coordinates of the protein. DSSP is a database of secondary structure assignments for all protein entries in the Protein Data Bank (PDB). DSSP is also the program that calculates DSSP entries from PDB entries.

DSSP has eight types of protein secondary structure, depending on the pattern of hydrogen bond. The list bellows shows the different types of protein secondary structure in DSSP:

i) H = alpha helix
ii) B = residue in isolated beta-bridge
iii) E = extended strand, participates in beta ladder
iv) G = 3-helix (3/10 helix)
v) I = 5 helix (pi helix)
vi) T = hydrogen bonded turn
vii) S = bend
viii) L = others

These eight types are usually assigned into three larger groups: helix (G, H and I), strand (E and B) and loop (all others). In this research, DSSP used as feature class are from the three classes, which is helix (H), strand (E) and coil (C). DSSP dataset can be obtained from the RS126 sequence data which contain secondary structures and will be implemented as the feature class to fit into SVM for prediction.

2.2 Methods

The study of protein secondary structure prediction will focus on its feature representation which is the local protein structure. Using the conventional methods of machine learning algorithm, which is applying only Support Vector Machine is not effective in protein structure prediction. This is due to the nature behavior where biological features are known to be dynamic rather than being taken as static data in pattern recognition problem solving. With this issue in mind, a preprocessing step is taken into consideration as an extra biological feature in order to enhance the performance of the system and accurately predict protein secondary structure from local protein structure. It is to be believed that considering biological features such as local protein structure, protein sequences information in feature selection is crucial in machine learning approaches. The reason why local protein structure is used as the additional feature in the study is because local protein structure able to analyze small sets of protein and approximate every part of protein backbone.

With DSSP and dihedral angle available in the workspace, secondary structure and DA can be segmented into different local protein structure with different segment lengths. Every local protein structure will have their own DA and DSSP after segmentation and by implementing them as feature vector and feature class, the data can now fit into SVM for classification to predict protein secondary structure.

Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. The Support Vector Machine (SVM) is a binary classification algorithm and with this attribute, it is suitable for the task of predicting protein secondary structure. SVM has shown that it is able to classify data precisely in the field of protein secondary structure prediction, functional classification of proteins, protein fold recognition, and prediction of subcellular location. SVM has previously been used in the prediction of protein secondary structure [5][6][7][8]. 10 fold cross validation is implemented in support vector machine to classify and predict protein secondary structure.

By using 10 fold cross validation, the datasets are partitioned into 10 samples. From the 10 samples, 1 of them is assigned as testing model to validate the data and

the rest are used as testing model. The process of cross validation is repeated 10 times, where each of the 10 samples is used once as the validation model. All of the results can be used to produce estimations for prediction. Kernel implemented is the RBF kernel. By using non-linear kernel, the margin hyperplanes can be optimized. The algorithm still works similarly with a linear algorithm, just that a RBF kernel is applied to every dot product.

The performance of the system is tested and output of the system will be analyzed right after it is released. The performance and accuracy of protein structure prediction is measured and evaluated by how well the system can predict protein secondary structure with higher accuracy and less false positive rate. To enhance the measurement system, widely used evaluation measurement for classification problem such as accuracy, true positive rate (sensitivity) and false positive rate will be applied.

Accuracy measures the probability of true results (true positives and true negatives) in the whole population (true positives, false positives, false negatives, true negatives). Accuracy can be calculated as follow:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

True positive rate which is also known as sensitivity or recall defines the proportion of actual positives which are correctly identified as such. It measures the probability of the true positive value among true positives and false negatives. The formula of sensitivity is shown as below:

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{2}$$

False positive rate measures the probability of the positive prediction result when the proteins are non-secondary structure. It can be calculated as follow:

$$FPR = \frac{FP}{TN + FP} = 1 - Specificity \tag{3}$$

Besides applying the evaluation method mentioned above, a statistical method, t-test is implemented for validation of the results obtained. A t-test is any statistical hypothesis test in which the test statistic follows a Student's t distribution, if the null hypothesis is supported. In the research, t-test is applied on two samples of result which represents different local protein structures

3 Results and Discussions

Initially, to understand the importance of optimizing local protein structure, the prediction is conducted using machine learning algorithm SVM without any feature representations. The native RS126 dataset is used as the dataset to fit into SVM for training and testing followed by evaluation. The native RS126 is the original sequence and protein structure obtained from the dataset without any pre-processing step being applied. The output is recorded and tabulated in Table 1.

Table 1. Evaluation results of the prediction of native RS126 dataset

	Helix	Strand	Coil	Overall
Accuracy	0.54	0.42	0.07	0.43
TPR	0.39	0.32	1	0.45
FPR	0.09	0.11	0.49	0.34

Chen proposed that by selecting numerous lengths for local protein structure, it will assist in improving the accuracy of protein secondary structure prediction [9]. The initial result shows that the accuracy of the prediction without using any feature selection or representations is very low even compare to the other existing methods. Hereby, this research proposed an optimization using local protein structure to predict protein secondary structure.

This study is carried out using 3 different segment lengths, length 13, 15 and 17. The definition of applying different segment length is taking in to account 13, 15 and 17 continuous residues or amino acids in the protein sequence. For each protein in RS126, local protein structure with 3 different segment lengths will be applied. The optimal length for local protein structure will be determined using the best overall accuracy from the results of evaluation. With t-test validation, the significance of the optimal local protein structure compare to the initial method can be observed.

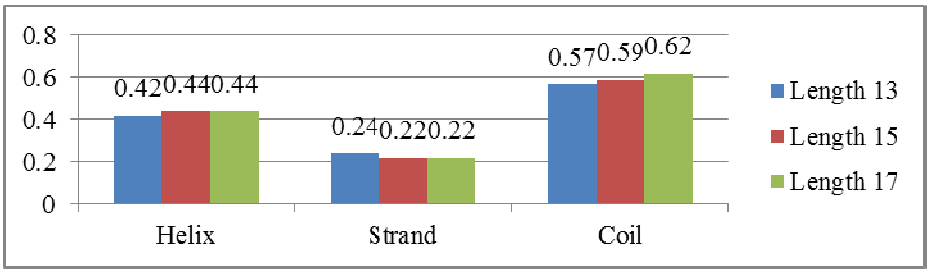

Fig. 2. Accuracy of each local protein structure based on secondary structural state

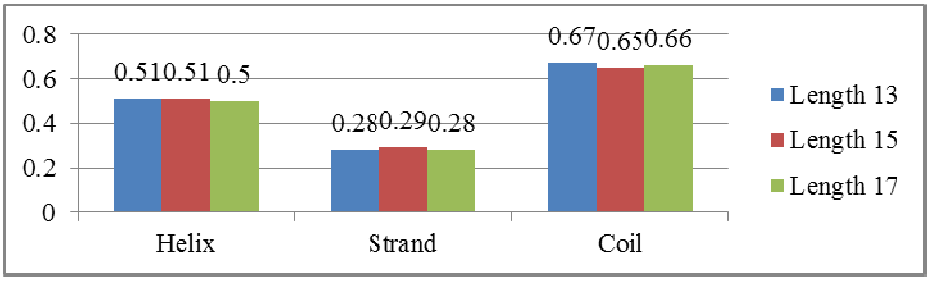

Fig. 3. True positive rate of each local protein structure based on secondary structural state

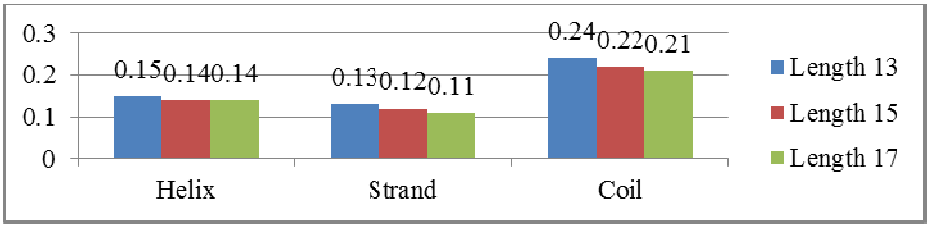

Fig. 4. False positive rate of each local protein structure based on secondary structural state

Most of the prediction results are evaluated by accuracy as depicted in Fig. 2. According to Fig. 2, for local protein structure with segment length 13, the highest accuracy is achieved by coil followed by helix and then strand. Similar results are collected from other local protein structures where coil having the highest accuracy among all secondary structural states. In terms of secondary structure, for helix, segment length 15 and 17 record the highest accuracy compare to others. Meanwhile, strand structure with length of 13 has the highest accuracy in compare to length 15 and 17. As for coil, length 17 records the highest accuracy among all.

In this research, other than accuracy, to provide a more reliable result, true positive rate and false positive rate are also used to analyze the prediction result. The results for true positive rate and false positive rate are illustrated in Fig. 3 and 4. For true positive rate, length 17 has the highest score for helix, length 15 for strand and length 13 for coil. As for false positive rate, length 17 has the lowest score for all secondary structural state.

From the tables and figures illustrated, it is obvious that generally, segment length 17 has the better accuracy compare to other local protein structures with the score of 0.44, 0.22 and 0.62. Most of the accuracies achieved is either the highest or is merely behind the highest score. Similar in true positive rate, most of the score that length 17 achieved is in the top range while in false positive rate, length 17 has the lowest rate among all local protein structures. It can be concluded that segment length 17 is the best local protein structure in this research.

A comparison of the prediction with optimal local protein structure with the prediction using native protein dataset is being conducted and analyzed. The proposed method with optimized local protein structure is expected to have better performance compare to the conventional prediction method in terms of accuracy, true positive rate and false positive rate. The comparison of the performance of both methods is illustrated in Fig. 5.

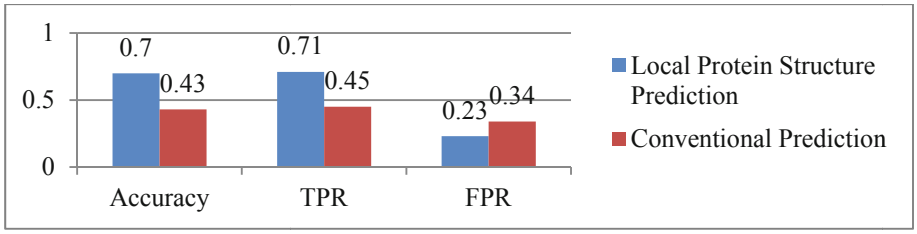

Fig. 5. Line graph of local protein structure prediction versus conventional prediction

According to Fig. 5, accuracy of local protein structure prediction is very much higher compared to conventional prediction. The score of accuracy for local protein structure prediction is 0.70 and it almost doubles the score of conventional prediction. This shows that by implementing feature selection or representation, there will be an improvement in prediction. Besides that, local protein structure prediction gives higher true positive rate and lower false positive rate. All the evaluation methods above indicate that the implementation of local protein structure achieved drastic improvement compare to the prediction method without any pre-processing or optimization.

Further validation of the results has been proposed to ensure the reliability of the prediction. A statistical validation, t-test, is conducted to test the significance of the results returned by the prediction. Table 2 shows the results of t-test for accuracy of the prediction system between optimal local protein structure and native structure. Only 11 samples are tabulated due to the large amount of protein sequence in RS126 dataset. It is noted that most of the t-test results returned h value as 1. This proves that the difference of accuracy predicted from the secondary structure prediction between optimal local protein structure and native structure is significant. The improvement of the accuracy, true positive rate and false positive rate is convincing and reliable.

Table 2. Sample of t-test results for accuracy between optimal local protein structure and native structure

	Significance	Lower Bound	Upper Bound
1azu	Yes	0.47	0.57
1bbpa	Yes	-0.39	-0.33
2aat	Yes	0.30	0.40
3ait	Yes	0.07	0.26
4bp2	Yes	-0.29	-0.17
5cytr	Yes	0.27	0.41
6acn	Yes	0.30	0.41
7cata	Yes	0.30	0.43
8abp	Yes	0.20	0.29
9apia	Yes	0.22	0.34
256ba	Yes	0.12	0.21

Finally, a comparison of accuracy between proposed method (optimal local protein structure), initial research (native structure) and other prediction methods is conducted. This is to observe the level of optimization of the proposed method compare to the conventional or other methods.

According to Table 3, it can be clearly observed that the initial research has the lowest accuracy due to lack of feature representations for the predictions. The proposed method which implement optimal local protein structure has the higher accuracy even compared to other prediction methods. This might be because by breaking down a native protein structure into small local protein structure segment, more information can be learned by the algorithm and will yield better predictions. Besides that, SVM is one of the most efficient binary classification algorithm compare to the algorithm used by other methods such as N-grams and others.

Table 3. Comparison of accuracy between different methods of protein secondary structure prediction

Methods	Reference	Accuracy
Extreme learning machine, Improved propensity score in binary scheme. Fixed window size.	Wang *et al*. [10]	69.0
Context sensitivity vocabulary, N-grams.	Yan *et al*. [11]	69.8
Initial Study: Native RS126 dataset, SVM	-	43.0
Proposed Method: Optimal Local Protein Structure, DSSP as Feature Class, DA as Feature Vector, SVM	-	70.0

4 Conclusion

Optimized local protein structure with SVM has been proposed to predict protein secondary structure. There were several interesting outcome faced during the study. The importance of protein secondary structure prediction, comparison of the study with previous work, influence of local protein structure to predict protein secondary structure, application of statistical method to enhance the reliability of evaluation methods have been conducted extensively and make great contributions to the research of protein secondary structure. Some future works are suggested to enhance the current prediction of protein secondary structure prediction such as use different datasets other than RS126, develop more feature representations and use various parameters in the classification process such as different cross validation and kernel. It is important to study more details about protein secondary structure because it will help us to understand more about their functions. With the knowledge of proteomics, contribution can be made to various fields such as development of cure in medicine sector.

Acknowledgements. We would like to thank Universiti Teknologi Malaysia for supporting this research by UTM GUP research grant (Vot number: Q.J130000.7123.00H67).

References

1. Walhout, A.J., Vidal, M.: Protein interaction maps for model organisms. Nat. Rev. Mol. Cell Biol. 2(1), 55–62 (2001)
2. Legrain, P., Wojcik, J., Gauthier, J.: Protein-protein interaction maps: a lead towards cellular functions. Trends in Genet. 17, 346–352 (2001)
3. Jing, N., Xia, B., Zhou, C.G., Wang, Y.: Protein Secondary Structure Prediction Methods based on RBF Neural Networks. Computational Methods 10, 1037–1043 (2006)
4. Rost, B., Sander, C.: Combining evolutionary information and neural networks to predict protein secondary structure. Proteins: Struct., Funct., Genet. 19, 55–72 (1994)

5. Hua, S., Sun, Z.: A novel method of protein secondary structure prediction with high segment overlap measure: Svm approach. J. Mol. Biol. 308, 397–407 (2001)
6. Guo, J., Chen, H., Sun, Z.R., Lin, Y.L.: A novel method for protein secondary structure prediction using dual-layer SVM and profiles. PROTEINS: Structure, Function, and Bioinformatics 54, 738–743 (2004)
7. Hu, H.J., Pan, Y., Harrison, R., Phang, C.T.: Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier. Nanobioscience 3, 265–271 (2004)
8. Yang, B.R., Hou, W.Z., Zhuna, Quan, H.: KAAPRO: An approach of protein secondary structure prediction based on KDD in the compound pyramid prediction model. Expert Systems with Applications 36(5), 9000–9006 (2010)
9. Chen, C.T., Lin, H.N., Sung, T.Y., Hsu, W.L.: Hyplosp: A Knowledge-Based Approach to Protein Local Structure Prediction. Journal of Bioinformatics and Computational Biology 4(6), 1287–1308 (2006)
10. Wang, G., Zhao, Y., Wang, D.: A protein secondary structure prediction framework based on the extreme learning machine. Neuorocomputing 72, 262–268 (2008)
11. Yan, L., Carbonell, J., Seetharaman, J.K., Gopalakrishnan, V.: Context sensitive vocabulary and its application in protein secondary structure prediction. ACM 1(58113), 881 (2004)