# CHAPTER II

# PREDICTION OF PROTEIN SECONDARY STRUCTURE

For four decades, the study of the secondary structure prediction has been done by using various techniques, especially machine learning theory from rule-based theory, neural network through recently used kernel methods. In this section, some detail of each technique is briefly discussed to understand the innovations of each one.

## 2.1 First Generation Methods

After Pauling and Corey had proposed the existence of $\alpha$-helices and $\beta$-sheets [1, 2], researchers aimed to study the correlation of the amino acids and protein structures [17]. The earliest one was the correlation on proline with the formation of $\alpha$-helices [16]. The main feature of these first generation methods was the use of single residue statistics for prediction [18]. There were a few numbers of known structures on database which made the data biased and the accuracy overestimated [5].

Chou and Fasman's method [19] used statistical analysis based on the known structures by assigning each amino acid the probability of forming an $\alpha$-helix or a $\beta$-sheet. To predict the secondary structure of a new sequence, contiguous regions of residues with a high probability of forming secondary structure features are considered. By example, if 4 out of 6 contiguous residues are likely to form an $\alpha$-helix or 3 out of 5 are likely to form a $\beta$-sheet, the assignment will be done to that structural element. These regions are then extended using a similar rule.

## 2.2 Second Generation

To decrease the bias, the second generation of secondary structure prediction used much more numbers of known structures and more detailed statistics than the single residue. The consideration concentrated on the consecutive stretch from $n - i$ to $n + i$ residues as contributing to the secondary structure of residue $n$, with $i$ typically in the interval {5, 10}. To make the analytical result more accurate, algorithms based on statistical information, physio-chemical properties, sequence patterns, neural networks, graph theory, multivariate statistics, expert rules and nearest neighbor algorithms [18] were applied.

Qian and Sejnowski (1988) applied the neural network for secondary structure prediction, It was a fully connected multi-layer perceptron(MLP) with one hidden layer model that Sejnowski and Rosenberg had developed for the problem of text-to-speech. [20] For the input, 13 optimal length residues were feed to the network. Before feeding, amino acids in each residue were encoded. Each amino acid was assigned a binary vector (such as (1, 0, 0 ...) or (0, 1, 0 ...)). It was unique and orthogonal for each one. Thus to encode an amino acid "alphabet" of 20 different amino acids requires that each amino acid is represented in 20 dimensional space. For the output, there were three output nodes representing the probability of the amino acid belonged to either an $\alpha$-helix or $\beta$-sheet or loop region

Firstly, the network was initialized with random weights. Secondly, it was trained by using backpropagation [20, 21] with the data set chosen from the Brookhaven Protein Data Bank (PDB). To avoid the performance oscillation because of contiguous data set [22], the input was fed randomly during the training. After the

training had been done, the system would predict the secondary structure of the amino acid in the middle position of each feeding residue. Although the prediction result of the network is better than other contemporary methods, the accuracy was around 64% which was rather low. However, using neural networks to predict secondary structure is still applicable [23, 24, 25].

## 2.3 Third Generation

Most of the previous methods faced the following problems [18]: (a) the accuracy was below 70%, (b) for $\beta$-strands structure, prediction result was only slightly better than random (c) the prediction was able to classify only the short strands and helices. Another problem was about the nature of secondary structure. When a protein folds, distal regions of the protein chain are shortened and the interactions increase. This makes the parts of the chains form the structure especially true for $\beta$-strands. With these problems, it was very difficult to obtain a high accuracy.

If a short fragment was used to train, the network was unable to capture all features in long range interaction. However, if a longer fragment was used instead, this problem still remained because the longer fragment was required more free parameters to represent its features. So there were insufficient data to train the network efficiently.

Rost and Sander [9] proposed PHD system for secondary structure prediction service via mail server. Actually, this architecture was similar to Qian and Sejnowski's work by including ensemble average and a secondary network. However, the main change on PHD was the predictions method that was based upon profiles, an

aligned sequence of homologous protein, instead of the network with single sequences.

When a mutation occurred in natural evolution, it destabilized a protein. Some mutations were successful and some were not. Normally, the successful one change only some residues, not change the main structure of protein. The evolution always preserved the structure rather than sequence.

Long range information was contained in the profile. The profile data was extracted by sequence alignments. The profile encoding method was an extension of orthogonal encoding method. Profile encoding data were an average of homologous sequence by extracting information from an alignment method. Rost and Sander method increased the accuracy up to 71 percent.

Baldi [22] used a recurrent neural network to extract feature from the sequence. Both sides of protein sequence chain were scanned by the network. Baldi method captured the long length information more effectively. It was implemented on prediction server SSPRO which is one of highly accurate prediction servers.

## 2.4 The Goals of Secondary Structure Prediction

It was impossible to achieve 100 % accuracy of the prediction. However, the actual goal of the protein prediction should be based on the three dimensional structure accuracy rather than two dimensional structure accuracy, because the three dimensional structure implies the function of protein. Normally, the three dimensional structure is rarely changed in evolution. Rost suggested the segment overlap measure (*SOV*) [28] for three dimensional structure measurement. Although

the two dimensional structure prediction was not reached 100 % percentage, the fold recognition was able to do correctly.

A comparison of the homologous protein on the same fold shows that the 12 percentage average was different on two dimensional structure. On the other hand, although the two dimensional structures of the same fold are significantly different, i.e. 12 percent, the three dimensional structures were still similar. So, it was impossible to predict the structure more accurately than 88 percent. The SOV method relied on this approach. The detail of SOV will be introduced in the next section. This method aims to the accuracy of protein segments rather than the accuracy of residues.

## 2.5 Training and testing data sets

It is not an easy task to choose a suitable dataset for the training process. Both knowledge of learning machines and specific domains are required. The idea is to choose a representative set of problems with known solutions that can be used to train the network and to test its performance. However, three problems can be encountered: (a) selected dataset that does not reflect the underlying probability distribution of real world examples is likely to occur, (b) selected data points that contain contradictory information, and (c) selected a data set that contains artificial correlations.

The first two of these will lead to problems in training a classifier. Over presenting particular patterns during training will lead to the machine being biased towards certain classifications and presenting contradictory patterns can lead to problems converging towards a classification scheme. A dataset with artificial correlations, however, will lead to over estimations of prediction ability. If a test set

contains highly correlated patterns to the training set, this will be testing the ability of classifiers to remember training patterns whereas we want to test its ability to generalize and to predict new patterns. In the context of secondary structure prediction, this means that no two proteins in the dataset can have pairwise sequence identity over 25%.

A further issue is that if prediction schemes are to be compared, then the comparison must employ the same training and test sets. Otherwise, the comparison is meaningless. The secondary structure prediction problem has, in general, standardized on several test sets. These test sets have been expertly chosen to avoid the problems mentioned above and to facilitate comparisons.

Three standard data sets were used to develop and test our prediction model. The first one is data set of 126 non-homologous protein chains proposed by Rost and Sander [9], referred to as RS126. This data set is widely used for training and testing many prediction models of protein secondary structures. The other is the larger data set contained 513 non-homologous protein chains included the first one except 11 protein chains that have an SD score of at least 5. This data set is constructed by Cuff and Barton [42], referred to as CB513. We selected these two data sets because they were classified as an appropriate representative subset of all protein in the Data Bank (PDB), and widely used for standard data sets of comparison objective.

Another larger data set used to measure the model is data set obtained from *PDB select* – the selection of a representative set of PDB chains [43]. This data set is useful for the development of prediction methods whereas it is intended to save time and effort by offering a representative selection that is currently about a factor of fifteen smaller than the entire database. The most recent data set recent released on July 2005

composes of 2810 chains with 430927 residues with percentage of homology identity less than 25%. With the larger data set, the predicting model can be trained and improved for overall accuracy.

## 2.6 Protein Secondary Structure Definition

Secondary structure elements were originally recognized by experts in the field. However, if we are to train classifiers to automatically predict secondary structure, it becomes necessary to obtain an abundance of high quality data of the known secondary structure of proteins. This is, however, a non-trivial problem.

When the structure of a protein is determined experimentally, the result is a list of co-ordinates of the positions of atoms in the protein. These co-ordinates are not perfect because of the errors occurred by the experimented process. Therefore, the positions of the molecules cannot be correctly determined by any algorithmic determination of secondary structure. Despite this, methods have been developed for secondary structure assignment.

The earliest approach to secondary structure assignment is the DSSP algorithm developed by Kabsch and Sander [5]. The program is freely available for academic use, and is currently maintained at http://swift.cmbi.ru.nl/gv/dssp/. The algorithm assigns secondary structure on the basis of the hydrogen bonding pattern between the backbone NH and carboxyl group.

In addition to DSSP, there are STRIDE [44] and DEFINE [45]. Cuff and Barton [42] undertook a comparison of the three methods. There is an agreement at only 71% of residues between all three methods. DEFINE seems to be the major cause of the

disagreement; DSSP and STRIDE agree at 95% of residues. In addition to this, there is a difference in the length distributions of the predicted structures. However, DSSP assignments will be considered for our comparisons.

The DSSP program assigns residues into eight different classes, H($\alpha$-helix), G($3_{10}$-helix), I($\pi$-helix), E($\beta$-strand), B(Isolated $\beta$-bridge), T(turn), S(bend) and – (rest or coil). To predict the secondary structure in our experiment, two prediction schemes are considered. The first scheme is "eight different classes" and the second scheme is three different classes". In case of three different classes of structure, only H(helix), E(strand) and C(coil) are considered. These three classes of structures are formed by grouping the eight basic structures as follows.

1) H, G and I to H; E to E; all other states to C.
2) H and G to H; E and B to E; all other states to C.
3) H and G to H; E to E; all other states to C.

However, the different assignment and reduction methods influence the prediction accuracy [42]. In our experiment, we used the first reduction method because it is considered to be a standard method and widely used.

**Table 2.1** Three and Eight Classes of Secondary Structure.

| SSP Class | Abbreviation | Simple Class |
|---|---|---|
| $\alpha$ -helix | H | H |
| $3_{10}$-helix | G | H |
| $\pi$ -helix | I | H |
| $\beta$ -strand | E | E |
| Isolated $\beta$ -bridge | B | E |
| turn | T | C |
| bend | S | C |
| rest or coil | - | C |

## 2.7 Performance measures

The standard performance measures for the prediction accuracy are the three-state overall per-residue accuracy measure ($Q_{index}$) and the segment overlap quantity measure ($SOV$).

**Overall Accuracy, $Q_{index}$**

This is the most obvious measure of prediction accuracy for measuring the percentage of residues predicted to be in their correct class. The index is defined as follows.

$$Q_{index} = 100 \times \frac{1}{N} \sum_{i=1}^{n} a_{ii} \tag{2.1}$$

where $a_{ii}$ is the number of elements of class $i$ predicted correctly to class $i$. This is an obvious and intuitive measure of the accuracy of a classifier. But this measure can easily be misleading when the class membership is not evenly distributed.

One of most obvious measure of $Q_{index}$ is $Q_3$ defined as the percentage of residues predicted in their correct class and shown in the following equation.

$$Q_3 = \frac{\sum_{i \in \{H,E,C\}} \text{number of residues correctly predicted in class } i}{\sum_{i \in \{H,E,C\}} \text{number of residues in class } i} \times 100 \tag{2.2}$$

Even through $Q_3$ can give us the performance of the predicting model, it does not reflect the specific goals of secondary structure prediction [28]. Therefore, the more

suitable measurement method for secondary structure, the segment overlap measure (*SOV*), had been introduced.

**Segment Overlap Measure, *SOV***

The segment overlap measure (*SOV*), firstly proposed by Rost [28] and modified by Zemla [29], is a measurement that provided for evaluation of secondary structure prediction methods by measuring the secondary structure segments rather than individual residues. The segment overlap measure (*SOV*) is defined as:

$$SOV = \frac{1}{N} \sum_{i \in \{H,E,C\}} \sum_{S(i)} \left[ \frac{minov\ (s_1, s_2) + \delta(s_1, s_2)}{maxov\ (s_1, s_2)} \times len\ (s_1) \right] \times 100 \qquad (2.3)$$

where s1 and s2 are the observed and predicted secondary structure respectively, s($i$) is the set of overlapping pairs of segments s1 and s2 in state $i$, *len*(s1) is the number of residues in segment s1, *minov*(s1,s2) is the length of two segments s1, s2 in state $i$ (the overlap segment), *maxov*(s1,s2) is the total length spanned by two segments s1 and s2, and N is the total number of residues. The $\delta$ is an integer value given by

$$\delta\ (s1,s2) = min\{(maxov(s1,s2) - minov(s1,s2)), \qquad (2.4)$$
$$(minov(s1,s2),\ int(len(s1)/2),\ int(len(s2)/2)\ \}$$

By comparing Q index and SOV score in Table 2.2, one can see how SOV does reflect the quality of the prediction.

**Table 2.2** Three different predictions that all have the same per residue accuracy.

| Experiment | Secondary Structure | $Q_H$ | *SOV* |
|---|---|---|---|
| Observed | -HHHHHHHHHH- | - | - |
| Prediction 1 | -H-H-H-H-H-- | 50.0 | 12.5 |
| Prediction 2 | -HHH---HH--- | 50.0 | 37.9 |
| Prediction 3 | ----HHHHH--- | 50.0 | 63.2 |

The predictions are qualitatively quite different. Prediction 1 indicates 5 different isolated helices, clearly a bad prediction, since helices have minimum length 3. Prediction 2 is better, with only 2 separate helices. Prediction 3 is the best since it predicts one contiguous helix. Therefore, SOV score reflects the difference in quality of the prediction.

**Matthew's Correlation Coefficient, $Co_i$**

Another measurement used to measure the correlation of results is Matthew's Correlation Coefficient [46]. This measurement is given by

$$Co_i = \frac{TP_i \times TN_i - TP_i \times FN_i}{\sqrt{(TP_i + FN_i)(TP_i + FP_i)(TN_i + FP_i)(TN_i + FN_i)}} \qquad (2.5)$$

where *TP, FP, FN, TN* are a number of true positives, false positives, false negatives, and true negatives for class $i \in \{H,E,C\}$, respectively. The result is a value between –1 and 1, such that 1 shows complete agreement, -1 shows complete disagreement, and 0 shows the prediction is uncorrelated with the results. This statistic, therefore,

allows comparison with a random baseline. A random scheme predicting according to class frequencies would have $Co_i$ close to 0. It also has the advantage that, for two class prediction problems, it encodes the quality of the prediction in a single statistic. We shall generate the correlation coefficient and use it for optimization of binary classifiers.

## 2.8 Cross Validation Method

All results reported are calculated through 7-fold cross validation. Even though a full jack-knifes test is more accurate, due to the limited computational power, the 7-fold cross validation is performed. The protein data sets are divided randomly into seven subsets with having similar size and similar content of each class of secondary structure. Then, the prediction model is trained on six subsets, and tested on the remaining subset. This process is repeated seven times, once for each subset.