CHAPTER V

DISCUSSION

5.1 Motivation and Benefits of this study

The structures of proteins especially the three dimensional (3D) structures, of proteins determine the proteins' properties and functions. Different structures also give rise to different functions. The 3D structure of proteins can be determined by X-ray diffraction and/or NMR techniques. These techniques can be determined with very high accuracy. Accurate determination of 3D structures requires complicated and time-consuming procedures Hence, an enormous gap exists between the number of known protein sequences and the number of established 3D structures (Rost, 1995).

Information about secondary structure of protein can be helpful in determining its structural properties. In the past years, much effort has been put on the development and improvement of computer program for the prediction of protein secondary and tertiary structures from the amino acid sequence data. The computational neural networks is a one method that has been use for protein structure prediction. Numerous papers have been published in the past few years on the usability of neural networks in protein structural study because neural networks can be applied to problems even without prior knowledge of an algorithmic correlation between the input and the output data, such as the relationship between the primary sequence and the native structure or the secondary structure of proteins (Bohm, 1996). At the prediction level, neural network technology offers easier and much more rapid structural prediction than classical determination by X-ray crystallography.

Many works studies have been conducted the neural network for secondary structure prediction (Qian & Sejnowski, 1988; Holly & Karplus, 1989; McGregor et al., 1989; Kneller et al.,1990; Rost & Sander 1993; Sasagawa & Tajima, 1993; Chandonia & Karplus, 1995). Using globular proteins with already known secondary structure as models, these papers present a similar, neural networks based method for predicting the secondary structures (helix, sheet and turn). The neural networks employed in this method use a "sliding window" approach to predict the secondary structure of each residue along the polypeptide chain. The goal of the networks is to predict the correct secondary structure for the middle amino acid of the "window". The network can be considered a "window" with 7-27 sequential residues of the protein at a time (Qian & Sejnowski, 1988; Chadonia & Karplus, 1995). The range of the prediction accuracy is between 60% - 75% for a three state model *(α, β* turn or coil). Rost and Sander (1994) have described a prediction method based on neural network and multiple sequence alignments. This network system has an accuracy around 70% and this rate is also similar to Levin et al (1993).

Because secondary structure prediction methods are based on sequence-specific information of residues composition in each protein. The secondary structure is also dependent on distant interactions. Thus, the methods for secondary structure prediction cannot be expected to achieve 100% accuracy. It has been suggested that improvement for the prediction accuracy may be achieved by taking into account information of residue composition of proteins (Duchak et al., 1993; Eisenhabor et al., 1995)

For the prediction of folding classes, only amino acid sequence descriptors (amino acid composition, transition distribution and properties) were applied as inputs to neural networks. An average accuracy of 71.7% was achieve for correct positive prediction. However, the prediction accuracies reported by different authors vary from 40 to 100 percent and are difficult to compare. The prediction success depends on the size and selection of members in the learning set an of the test set of structures. It must be emphasized that the distributions of the five folding classes (all α, allβ, α+β, α/β and ζ) are not well separated in amino acid composition space, but show a considerable overlap. Although additional independent parameters other than sequence composition could increase discrimination, the folding classes are determined by complete set sequence properties and environmental conditions only for some sequences. The classification in accordance with the traditional five structural classes might not be optimal.

In this study we used coded properties of the amino acid residues as input vectors and trained the neural network for prediction of secondary structure and folding classes. The amino acid properties including hydropathy, hydrophobicity, amino acid side chain properties and helical tendencies were expected to contain information for improving the prediction of secondary structure. Information from accurate secondary structure prediction should be useful in the prediction of some folding classes.

Three layer feed forward neural networks using the computer program SNNS (Andreas et al., 1989) were utilized in this study. The predictions of the existence of helix, sheet and turn in amino acid of each protein were performed for testing the network with the prediction of easy problem. The predictions of the percent helix, sheet and turn in amino acid of each protein were performed for further folding classes prediction.

If the method in this study can be developed for secondary structure or folding classes prediction, it will provide an easy method which could predict the structure of a protein from its primary sequence of amino acid.

5.2 Evaluation of Secondary Structure Prediction Accuracy

When comparing the results of different secondary structure prediction methods, the method used for evaluation of the result is important. The most commonly used measure for secondary structure prediction accuracy is the three-state residue by residue score giving the percentage of the correct predictions, Q defined as

$$Q = \frac{(p(\alpha) + p(\beta) + p(\text{coil}))}{N}$$

where, p is number of residues predicted correctly for a given secondary structure type and N is the total number of residues. This method however was used for measurement the percent accuracy of the location of the secondary structure element. But in this study, the percent accuracy was the result from the secondary structure prediction of the whole amino acid sequence of protein. The prediction did not specify the location or position of the secondary structure in the amino acid sequence just predicted for the existence and percent of secondary structure in each protein. Thus prediction accuracy can be computed as:

$$\text{prediction accuracy} = \frac{\text{the total number of protein predicted correctly}}{\text{total number of proteins for testing}} \text{ x 100 percent}$$

5.3 Training and Testing on Neural Networks

The 98 proteins with known secondary structure in this study were collected from Protein Data Bank. These proteins have low sequence homology and the average sequence identity overall possible aligned Protein Data Bank sequence pairs is 15% (Duchack et al., 1995). Only 98 proteins were used in this study because each of them has a single amino acid sequence which allows ease of input data management easily for training and testing by neural networks.

Many works in the protein structure prediction used the “cross validation test” and “Jack-knife testing” for setting the training and testing sets. The cross validation test, the proteins are divided into classes and then subset by random permutation. One protein from the class and proteins from 1 subset of other classes were used for testing and all other proteins of the class and 9 the remaining subsets were used for training. All possible combinations of proteins from the class and 10 subsets were made for training and testing (Duchack et al., 1995). The Jack-knife testing, for example, if there are 130 proteins, 129 proteins are used for training and 1 proteins for testing. This has to be repeated 130 times until each protein has been used once for testing. The average overall 130 tests give a reasonable estimate of the prediction accuracy.

In this study the 98 proteins were set into two groups of training and testing. The training set consisted of 70 proteins and the testing set consisted of 28 proteins. This 70 number of training example was set for giving the possible highest number for training. The number of testing, 28 proteins, was basically enough to testing with the number of training, 70 examples. Because the number of training and testing were rather small, the proteins were tried to divided into these two sets by possible output. The training set should have the proteins member that have the possible outputs enough for teaching the network. The early experiment of this study, for the prediction of the prediction of helix, sheet and turn in the same network using the hydropathy (2 groups) and amino acid side chain properties, the training and testing was randomly set to have four possible groups of training and testing sets. These 4 groups were used for training and testing and gave the 4 possible percent accuracies for the testing using each properties vector. The percent accuracies of four groups using hydropathy (2 groups) as input vector, did not significantly different with 95% confidential limit. The training and testing using amino acid side chain properties also gave the same result with non significant different (data not shown). Because, the training step require time for training. Thus, most experiments in training and testing were performed only one group of training and testing for reducing time in training.

Three layer feed forward neural networks of SNNS program were used in this study. The networks consisted of input layer, output layer and hidden layer. The input layers was composed of 481 number of input units. Output layer consisted of 3 units for prediction of the existence of helix, sheet and turn structures in the same networks, 1 unit for prediction of the existence of helix or sheet or turn in separate networks, 3 units for prediction of percent helix or sheet or turn (5 groups) in separate networks and 2 units for prediction of percent helix or sheet or turn (2 groups and 3 groups). Hidden layer with various units were also tested for determining which produced the most accurate results. The hidden layer with 7, 35, 70, 10 and 120 units were used in each trial. These numbers of units expected to include the small numbers and the high numbers of hidden units for giving a good prediction.

Thus, there are 70 proteins and 28 proteins was set into training set and testing set respectively by the possible output of each experiment. There is only one set of training and testing set for training and testing in each experiment. All proteins in training and testing set had 481 hidden units and the output unit number depend on the desired output.

5.4 Prediction Accuracy

The existence of helix, sheet and turn predictions

For prediction the existence of helix, sheet and turn structure in amino acid sequence using neural networks, the prediction of these three structure was first perform in the same network. Thus, the output consisted of 3 units representing for helix, sheet and turn respectively. There were 7 possible outputs of these three structure in the same network. Hence, the probability for predicted correctly was 1/7 or approximately 14% accurate and this probability could be the baseline of percent accuracy prediction of the three structure in the same network.

All networks which were trained by the properties of amino acids and non-property coded inputs gave the predictive accuracy more than the baseline. The lowest accuracy (28%) was obtained from the networks which were trained by non-property input pattern. While, the networks which were trained by properties of input patterns gave the percent accuracy between 30-55%, which were higher than that of the non property input. These results suggest that these properties, hydropathy, amino acid side chain properties and hydrophobicity contain certain information that helps with the secondary structure prediction. The best network for the prediction is the network that give the highest accuracy. Thus, the amino acid side chain properties were the best properties which gave the highest prediction accuracy whereas hydrophobicity gave the lowest accuracy prediction. The hydrophobicity did not improve this prediction when compared with the non-properties input. The overall result of the prediction from all properties were not quite good because most results were less than 50 percent accurate. This problem might be the networks were not specific for all three structures (helix, sheet and turn). It is also possible that the number of patterns of each output may not be enough to train the network for a good prediction. Thus, a separate network for prediction of each individual structure were constructed to improve the prediction accuracy.

There are 2 possible outputs (0, 1) of the prediction of the existence of helix or sheet or turn in separate networks. Thus the probability of the correct prediction is 1/2 or 50%. This value can be used as the baseline for these predictions in separate networks. The result shows that all properties can give higher percent accuracy prediction especially for helix structure prediction.

For the existence of helix structure prediction, the baseline of the accuracy was 50%. But from the database, 91 form 98 proteins have helix structure. Thus, from this data, it can be assume that approximately 90% of proteins have helix structure and the chance for correctly predicted of the existing of helix structure in protein is also 90% if the prediction was always set to "helix". From the results after training with various properties, the percent prediction accuracy were between 80-100%. This level of accuracy was no better than the baseline. However, some networks could give 100% accuracy. From a probability consideration similar to helix case, the baselines of the sheet structure and that of the turn structure accuracy are both 50%. If the prediction was always simply set to "sheet", it would give correct results 80% of the time, according to the database. Similarity, if the prediction was always set to "turn", it would give correct result 60% of the time. Therefore, in case of sheet and turns, the accuracies of prediction obtained from random prediction and almost as high as the baseline prediction obtained from prior knowledge of the protein 3D database.

From the overall results of the existence of the secondary structure predictions, higher level of accuracy could be obtained from the network compared to random predictions. Thus, the neural networks can be use for secondary structure prediction without prior knowledge of the protein database. The prediction accuracies of the existence of helix, sheet and turn structure obtained from separate neural networks are higher than those obtained from a single multi output network. Because separating into one network can reduce the number of output units, a chance for correct prediction can increase. Thus, different outputs (helix, sheet and turn) should be trained in separate networks to give a good prediction. The number of the example for training also effect the accuracy. More number of example give good training or learning that resulting in a good prediction accuracy.

Percent helix, sheet and turn prediction

The experiment of percent of helix, sheet and turn predictions were performed for folding classes prediction because the most found of folding classes was classified by the content of secondary structure especially α-helix and β-sheet. First the output was divided into detailed six groups of 0%, 1-20%, 21-40%, 41-60%, 61-80%, and 81-100%. By chance, the baseline probability for correct prediction is 1/6 or approximately 17%.

This probability was used as the baseline for the prediction of percent 6 groups of the three secondary structures. Most properties gave the range of predicted accuracy for helix prediction higher than another structures. The range of predicted percent (6 groups) of helix is between 35-70% accurate. While, the ranges of predicted of sheet and turn were 25-50% accurate. When compare percent accuracy of all structures with baseline, these ranges of accuracy were higher than the baseline. Hence, the predictions using neural networks are better than using the random method.

When reduced the number of output from 6 groups to 3 groups, the percent of helix, sheet and turn were roughly divide into 3 groups of 0%, 1-50% and 51-100%. By chance, the baseline of these 3 groups prediction is approximately 33%. The ranges of percent accuracy for helix, sheet and turn predictions were raised to 65-85%, 60-80% and 50-65% respectively when compare with 6 groups predictions and all accuraries also higher than the baseline. This result suggest that smaller number of output give rise higher accuracy. Thus, from this reason, 2 groups of percent helix and sheet prediction were performed further for higher accuracy.

Percent of helix and sheet were divided into 2 groups by the definition of the folding classes (Levitt & Chothia, 1976). The first class is all α which has α helix more than 15% and β-sheet less than 15%. The second class is all β which has β sheet more than 15% and α helix less than 15%. The third group is $\alpha+\beta$ or α/β which the percent of helix and sheet structures are not the first or the second group. These classes were defined by mixing the criteria of Nishikawa and Ooi, Kneller et al. and Klein and Delisi (Nishikawa & Ooi, 1982; Klein & Delisi, 1986; Kneller et al., 1990) for setting the simple outputs. Thus, percent helix and sheet were divided into two groups of 1-15% and 16-100%. The hydropathy (7 groups) did not trained for both helix and sheet prediction because the result from previous prediction, it did not give the significant different accuracy result from hydropathy (2 groups). Where the helical tendencies coded amino acid sequence were used only the helix prediction because these values are specific for helix structure. The range of accuracy for the helix prediction was 65%-80% which higher than the baseline (50%). While the range of accuracy for sheet prediction was 60-75% and also higher than the baseline. The helical tendencies code amino acid sequence which were specific values of helix structure, did not gave the high percent accuracy as was expected. On the other hand these values gave the lowest accurate when compare with other properties. Because, these tendencies were divide into 5 groups by the values, thus relatively low accuracy have resulted from inappropriate grouping of these values.

The percent accuracy of all percent helix, sheet and turn predictions using neural networks were higher than the baseline. Thus, the prediction by neural method is still better than the random method. The percent of secondary structure is very helpful in the folding classes prediction. From the definition, the folding classes were classified by the content of helix and sheet structure. Thus if the prediction networks of percent helix and sheet can be improve to give a higher accuracy, it will be used for folding classes prediction by easy and the faster method.

The overall results of all predictions in this study show that, different properties coded amino acids (input), number of outputs and number of hidden unit give different prediction accuracies.

The properties coded amino acids can effect the prediction because they define the position of input in n-dimensional space where input pattern is acting as a vector in n-dimensional space. In this study, the properties were changed into symbol and then coded the amino acid. Different properties have different symbols for coding, thus, the position of the input in n-dimensional space depend on the amino acid composition and property coded. The number that coded the amino acid sequence (input) should contained the information that can define outputs. If properties specific for the desired output, all inputs which have the same output will be at the same n-dimensional space. Such a network is able to classify the group of output to give a good predictions.

Nevertheless, the number of hidden unit should be considered with the input vectors. Different hidden units resulted in different prediction accuracy. Because the classification of the input and output in the training step depends on the number of hidden units, inappropriate number of hidden unit will define inputs into the wrong group of outputs. The appropriate number of hidden units can not be predicted a priori, thus, the number of hidden units should be optimized in all training.

The number of output is another factor that should be considered. Since a small number of output units give rise to a high chance for correct prediction. Thus, the output unit should be set as small as possible. Separate networks for prediction of each individual output is one way to reduce the number of output units.

5.5 Further Study

In this study, only 98 proteins were used in this study. Seventy proteins were grouped into the training and twenty-eight were test set. The training and testing proteins were then divided into subgroups depending on desire output. Thus, there were small number of protein examples for training for prediction of desire output. This problem can lead to the low prediction accuracy because the network have not enough examples for learning. Larger number of proteins are needed for teaching the network in order to improve prediction accuracy.

Furthermore, the prediction accuracy also depends on the input vector for training. The longest sequence of protein in this study spans 481 amino acids residue. The input of the network is always a constant value, thus each amino acids sequence was set to correspond to 481 units, the value of which were initially by zeroes. The padding of zero to last amino acid residue of the sequence to make such sequence to reach 481 units long changed the position of input vector in multidimensional space and gave rise to the wrong predictions. Hence, the padding number with the smallest results in change of input position in multidimensional space is need for improving the accuracy prediction. This number may be calculated from the value of each input pattern. Moreover, the property coded amino acid as input vector are also effect the prediction accuracy. Others properties not study here also should be investigated for giving the best properties that specify the desire output. Because forming of secondary structures or folding classes of proteins are guide by many factor or amino acid attributes such as charge, side chain bulk, backbone flexibility, hydrophobic moment, amino acid neighboring and various types of descriptors. If we could train a neural network, the combination of the amino acid properties that contain the information useful in predicting the secondary structure of folding classes can be trained the neural network at that time, these properties combination training can help for improving the prediction accuracy.