CHAPTER IV

RESULTS AND DISCUSSION

4.1 Experimental Results

The experiment has been performed on the 1st 2nd 3rd and 4th orders Markov model to examine the effect of higher orders to the performance of the classifier. In Table 4.1, the results clearly show that the performance of binary classifiers can be improved by increasing the order of Markov model. However, increasing the order of the model has some limitations. Higher order means longer sequence patterns have been used to create the Markov Transition Matrix. The multi-dimensional array of those matrices requires multiple numbers of free parameters to be estimated. Therefore, the problem of not having enough data to estimate the model is inevitable.

From the experiment, the 3rd order Markov model seems to be the optimum model for the limited estimated data available. The zero elements in those matrices are considered. We found that in 4th order transition matrix, the numbers of zero elements are more than 50% of the number of all elements in the matrix whereas the 3rd order transition matrix has numbers of zero elements less than 10%. Nevertheless, the information in the 4th order transition matrix is still useful because those zero elements can be removed by interpolation technique.

Although constructing of high-order Markov transition matrix is not a complicated task, the matrix consumes huge resources of memory. Due to its multi-dimensional array representation, the more increasing orders of the model, the more number of array dimensions will be extended. With the limited physical memory space in the system hardware, we performed the experiments only up to 4th order Markov model. The results from Table 4.3 shows that the performance of classifiers significantly increases from 1st to 3rd order, but in the 4th order model, it becomes unreliable. These unreliable results may be affected by the limited estimated data.

Table 4.1 Performance of binary classifiers (%) respected to step and order of Markov model.

Binary Classifier	step	Order of Markov model			
		1st order	2nd order	3rd order	4th order
H/~H	single	79.37	84.85	88.61	88.67
E/~E		77.63	82.68	87.90	88.15*
C/~C		74.24	77.62	82.95*	82.83
H/~H	double	77.31	84.02	85.64	85.30
E/~E		76.64	80.99	86.25	86.94
C/~C		74.15	75.82	80.88	79.93
H/~H	triple	79.49	85.50	88.72	88.75*
E/~E		75.46	81.81	84.14	85.95
C/~C		72.22	75.42	79.97	79.89

*Results on RS126 set, 1st-layer network with window size = 9, * indicates the optimum model*

For our experiment on multi-step Markov model, the result shows that the Extend (E) and Coil (C) structures have the best performance on single-step Markov model; whereas, the Helix (H) structure obtains the best performance on the triple-step Markov model. This finding suggests that some informative data can be obtained from the jumping sequence of amino acid. If we consider the triple-step model, the jumping step jumps from the position *i* to *i*+4 associated with the bonding position of H-bond between amino acid residues on the helix structure. In addition, the helix structure (H) and extend structure (E) have the best performance on 4th order whereas the coil structure (C) has the best result on 3rd order Markov Model.

With those results from Table 4.3, the optimal Markov model's structure for the binary classifiers can be obtained (* indicates the optimal values of those models). Based on the optimal structure, the optimal step and optimal order of Markov model, the further experiment regarding the window size of input sequence must be investigated.

Table 4.2 Performance of binary classifiers (%) on 1st-layer with respect to window size.

Binary classifier	Window size					
	5	7	9	11	13	15
H/~H	81.18	81.67	81.75*	81.37	80.93	80.05
E/~E	65.69	67.32	68.15	68.54	69.10*	68.61
C/~C	79.89	80.05*	79.95	78.97	78.45	76.36

Results on RS126 set, multi-step, multi-order Markov model, using optimum step and order
** indicates the optimum window size.*

Table 4.3 Performance of binary classifiers (%) on 2nd-layer with respect to window size.

Binary classifier	Window size				
	9	11	13	15	17
H/~H	80.19	80.21	81.97	82.05*	80.67
E/~E	68.64	68.73	69.09	70.12*	69.31
C/~C	79.71	80.90*	80.02	79.36	72.10

Results on RS126 set, multi-step, multi-order Markov model, using optimum step, order, and window size on 1st-layer, (*) Indicates the optimum window size.

Tables 4.2 and 4.3 show the accuracies of the binary classifiers from 1st layer and 2nd layer, respectively. The increasing of accuracy in 2nd layer confirms the improving efficiency of double layer network over the single one. By comparing the results for each classifier respect to the window size, it is possible to find the optimal configuration. The optimal size of window for each binary class classifiers is not the same. The optimal size of input window for Helix, Extend and Coil structural classes for 1st layer are 9, 13 and 7, respectively, and for 2nd layer are 15, 15 and 11, respectively (the optimal value indicate by *). We use these optimal sizes of input windows to prepare the specific input patterns for each classifier to train and test the learning model.

Table 4.4 Ratio of number of SVs to all training samples.

Network's Layer	Binary Classifier	Ratio of SVs : all samples (%)	
		SVM(freq)	SVMmer
1st-layer	H/~H	50.46	37.75
	E/~E	43.92	36.06
	C/~C	59.02	46.67
2nd-layer	H/~H	-	10.65
	E/~E	-	10.07
	C/~C	-	17.34

Results on RS126 set, 3rd-order Markov model, using optimum window size
The results of SVM(freq) obtained from (Hua and Sun, 2001);
The results of SVMmer obtained from the method proposed in this research.

The number of support vectors used by a classifier is a good indication of the difficulty of a classification problem. If a large number of support vectors are needed then the problem is more difficult to be classified, If a low number of support vectors are needed, it is simpler. The ratio of SVs to all training samples is shown in Table 4.6. We found that the ratio for each classifier in 1st layer is in rank of 30-50%. It means only 30-50% of the training samples could represent the information of all samples. Comparing to the ratio of 40-60 % from the orthogonal encoding scheme with SVM approach reported by Hua and Sun [34], the result clearly indicates that the Markov transition matrix encoding scheme makes the classification easier than the conventional one. In 2nd layer, the ratio for each classifier also reduces to rank of 10-20%. This evidence confirms for the simplification of classifying in this layer.

Ternary Classification Results

The result shown in Table 4.7 indicates that there is a significant improvement of the reclassifying technique in the double layer network model. In this model, the intermediate result obtained from the 1st layer classifiers is used as input patterns to the 2nd layer's classifiers. Then, the 2nd layer's classifiers reclassify and produce the final result. With this process, very short fragments of structure can be eliminated. Therefore, the secondary structure output will be smoother and likely to be more accurate.

Table 4.5 Results from ternary classification, comparing of single and double layer network model.

Data Set	Network Model	Q_3 (%)	SOV (%)	Q_H (%)	Q_E (%)	Q_C (%)	Co_H	Co_E	Co_C
RS126	Single	77.81	72.50	81.75	69.10	80.05	0.60	0.54	0.56
	Double	78.42	73.10	82.05	70.12	80.90	0.61	0.55	0.57
CB513	Single	79.88	73.23	82.12	69.95	83.53	0.68	0.55	0.58
	Double	80.52	74.26	83.64	71.25	85.34	0.63	0.56	0.60
PDB- Select	Single	80.18	75.78	84.20	71.34	82.63	0.61	0.54	0.56
	Double	81.92	78.21	85.22	71.73	84.13	0.61	0.55	0.59

Results on multi-step, multi-order Markov model, using optimal step, order and window size

The combining step of Markov model to make the combining input vector of single, double and triple-step Markov model can assist to add an extra informative feature form the amino acid sequence pattern. This assumption can be confirmed with the result shown in Table 4.8. The overall accuracy of the predicting model (Q_3 and SOV) can be increased for all data set when the combining input of multi-step Markov model have been presented.

Table 4.6 Results from ternary classification, comparing of optimum and combine step of Markov Model.

Data Set	Step of Model	Q_3 (%)	SOV (%)	Q_H (%)	Q_E (%)	Q_C (%)	Co_H	Co_E	Co_C
RS126	Optimum	78.42	73.10	82.05	70.12	80.90	0.61	0.55	0.57
	Combine	79.04	73.13	82.75	71.02	81.94	0.62	0.55	0.58
CB513	Optimum	80.52	74.26	83.64	71.25	85.34	0.63	0.56	0.60
	Combine	81.10	75.71	84.20	71.98	85.16	0.64	0.57	0.60
PDB- Select	Optimum	81.92	78.21	85.22	71.73	84.13	0.61	0.55	0.59
	Combine	82.18	78.52	86.18	72.01	85.41	0.64	0.56	0.60

Results on multi-step, multi-order Markov model, using optimal order and window size

Table 4.7 Results from ternary classification, comparing of filtered and non-filtered result.

Data Set	Filter	Q_3 (%)	SOV (%)	Q_H (%)	Q_E (%)	Q_C (%)	Co_H	Co_E	Co_C
RS126	non-filter	79.04	73.13	82.75	71.02	81.94	0.62	0.55	0.58
	filter	80.15	75.22	83.21	72.12	82.63	0.63	0.57	0.58
CB513	non-filter	81.10	75.71	84.20	71.98	85.16	0.64	0.57	0.60
	filter	82.49	77.18	85.81	72.05	85.49	0.65	0.59	0.61
PDB- Select	non-filter	82.18	78.52	86.18	72.01	85.41	0.64	0.56	0.60
	filter	83.10	80.02	86.54	73.10	88.35	0.64	0.58	0.62

Results on multi-step, multi-order Markov model, using combining step, optimal order and window size

The filter process based on the value of reliability index (Ri) is the final process to make the improvement of the predicting accuracy. Table 4.9 shows the significantly improving of the accuracy (both Q_3 and SOV). The results from the filtered predicting model are higher about 2% on SOV and 1% on Q_3. Obviously, the performance of our model can be improved with the error correction of the filter algorithm based on information from the reliability index.

Table 4.8 Comparison with the results of other approach for three class classification.

Method	Data Set	Q_3 (%)	SOV (%)	Q_H (%)	Q_E (%)	Q_C (%)
PHD[1]	RS126	70.8	73.5	72.0	66.0	72.0
SVM(freq)		71.2	74.6	73.0	58.0	73.0
SVM(psi)		76.1	79.6	77.2	63.9	81.5
SVMmer		80.2	76.2	85.2	70.1	82.6
SVM(freq)	CB513	73.5	76.2	75.0	60.0	79.0
SVM(psi)		75.2	80.0	80.4	71.5	72.8
SVMmer		82.5	77.2	85.8	72.1	85.5
PSIPRED	Private data	76.5	73.5	-	-	-
BRNN		75.1	-	-	-	-
SVMmer	PDB-select	83.1	80.0	86.5	73.1	88.4

The results of PHD obtained from (Rost and Sander, 1993) and (Rost et al., 1994);
The results of SVM(freq) obtained from (Hua and Sun, 2001);
The results of SVM(psi) obtained from (Kim and Park, 2003);
The results of PSIPRED, and BRNN obtained from (Jones, 1999) and (Baldi et all., 1999) respectively, reported on different data set;
The results of SVMmer obtained from new SVM approach with multi-order Markov encoding scheme proposed in this research. Combined results of 7-fold cross validation are shown.

Table 4.10, comparing the results of several predicting models, show that our present method (SVMmer) can obtain a remarkable result that out-performs other advanced methods at present. SVMmer achieves Q3 of 80.2%, SOV of 76.2% on RS126 and achieves Q3 of 82.5%, SOV of 77.2% on CB513. Our method produces a very good accuracy on Q3 but fair accuracy on SOV. The lower accuracy of SOV may result from the nature of the input features, yielding fragmentation of predicting structures. Because only local information of connected residues is used, the unreliable classification for some regions of amino acid pattern may occur. This may be the cause of shot fragments on prediction results.

For the accuracy of binary classification, the helix structure (H) achieves the highest result whereas the strand structure (E) gets the lowest result. This evident shows the difficulty for the predicting of each structure. In fact, the strand structure (E) is more affected from distant amino acid residue than any others structure.

Therefore, only the input patterns of connected residues may not provide enough information to increase the accuracy of the strand structure. Even the result of E class is lowest when compared to other classes, the result is satisfactory when compare to other predicting methods in all data set. The very high accuracy of H class may obtain from the local information extracted form the jumping step of Markov model relating to the relationship of H-bond between amino acid residues.

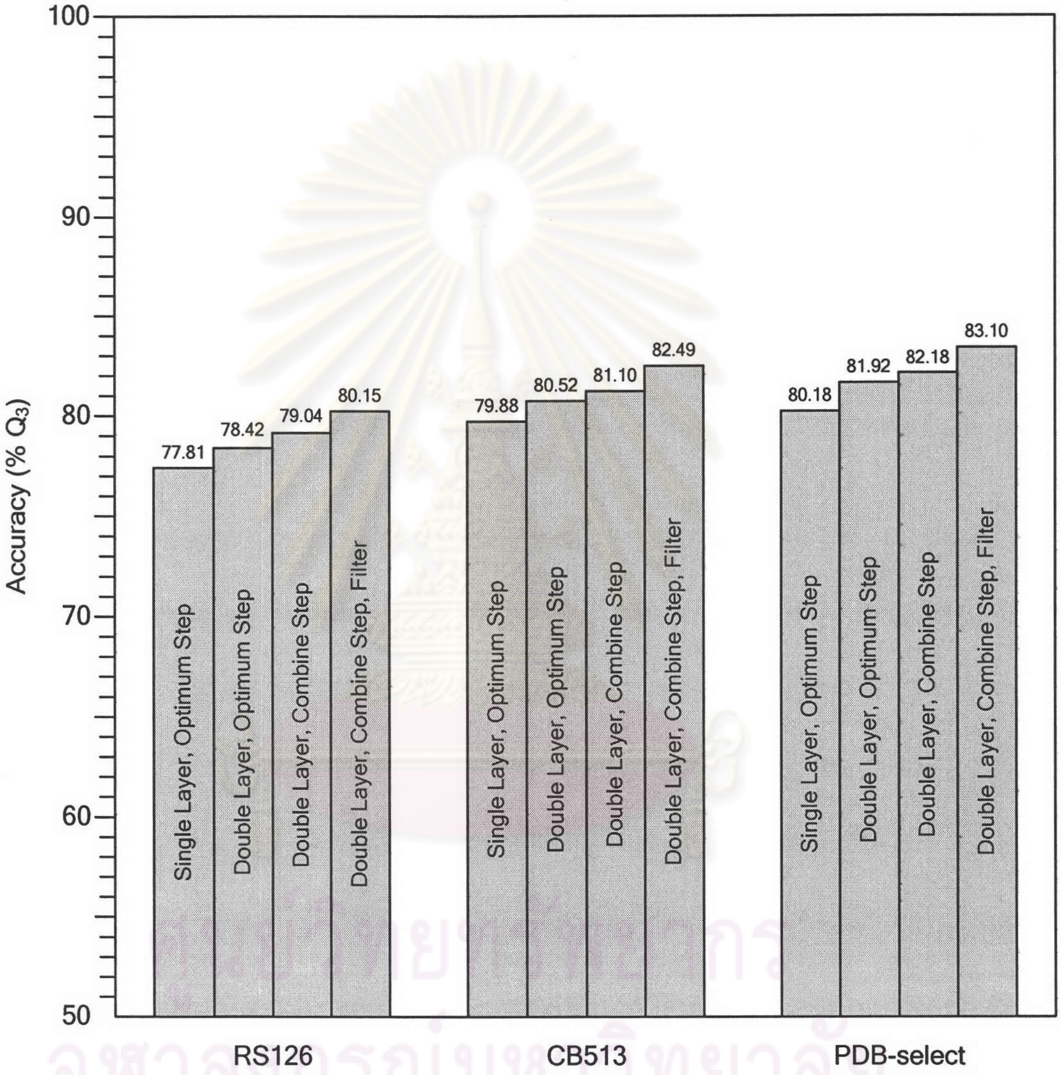

Figure 4.1 Comparison of three-state overall per-residue accuracy (Q_3).

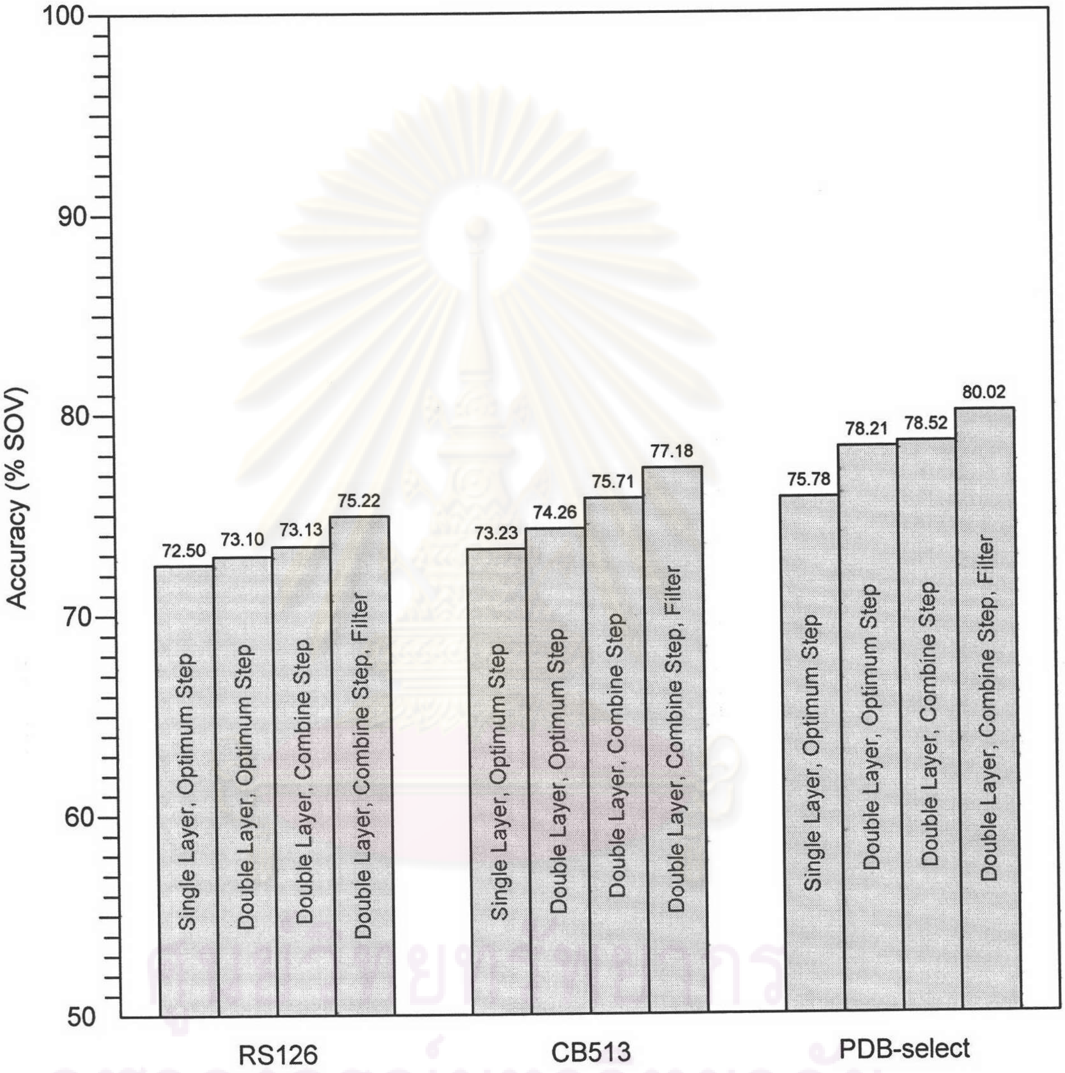

Figure 4.2 Comparison of segment overlap accuracy (*SOV*).

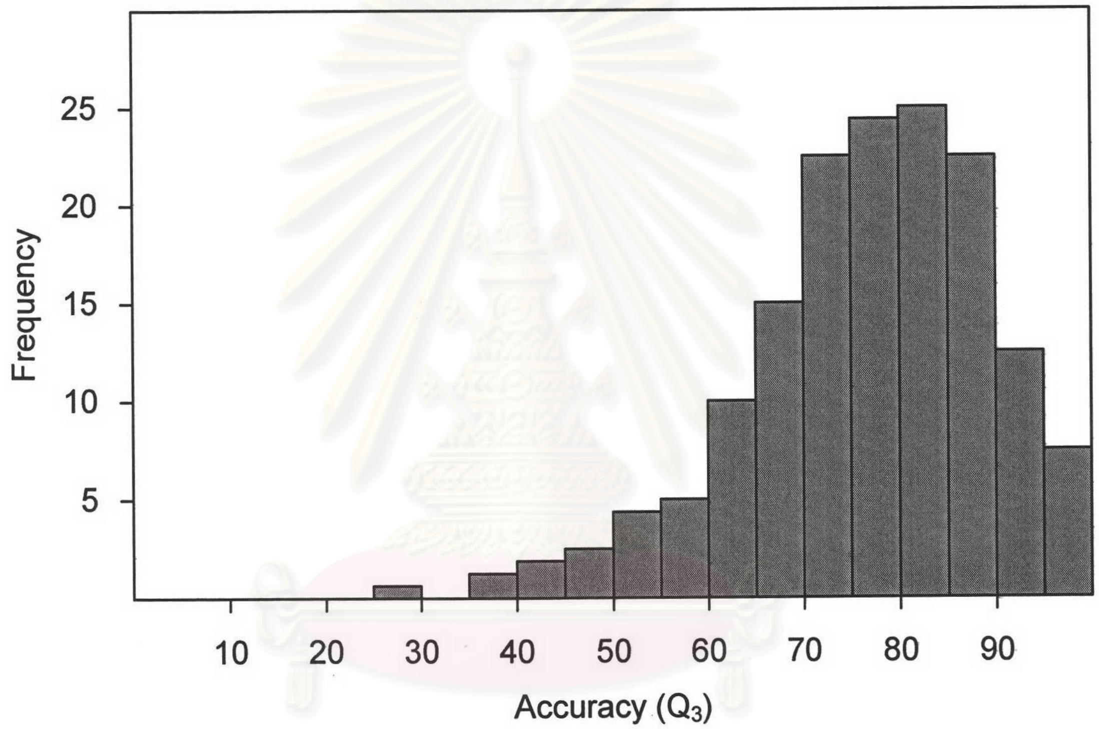

Figure 4.3 Distribution of proteins predicted with different accuracies.

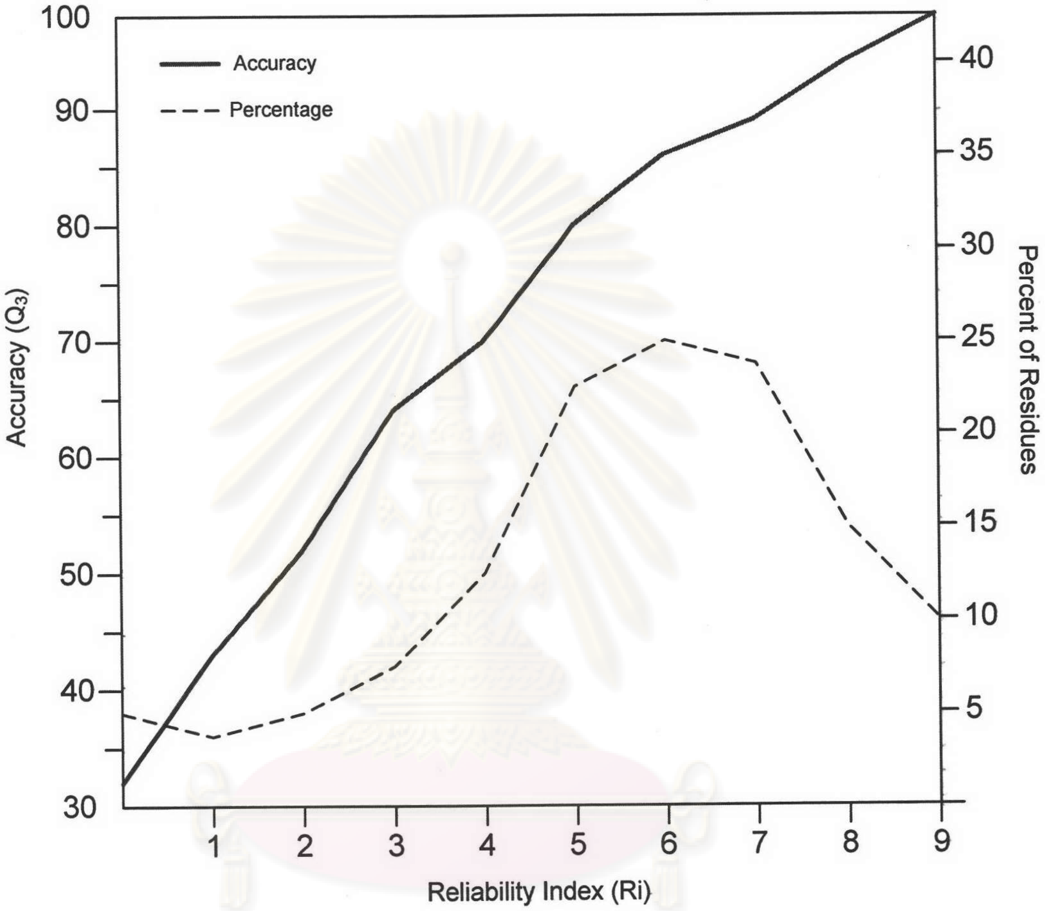

Figure 4.4 Distribution of accuracy (Q3) and number of residues with different reliability index (Ri).

4.2 Conclusion

A novel approach to the prediction of protein secondary structure from sequence data is proposed. The approach employs a new encoding method based on the Markov transition matrix to produce the input vectors that are efficiently used to train a SVM-based neural network. Designed network structure as well as optimized parameters of the prediction model is presented. Our approach achieves the highest accuracy on the widely used benchmark data sets. The SVM together with Markov transition matrix encoding scheme produces the performance of three-state overall per-residue accuracy measure Q_3 = 82.49% and segment overlap accuracy measure SOV = 77.18% through a seven-folded cross validation on the data set of 513 non-homologous protein chains (CB513). From the larger data set, 2810 non-homologous protein chains from *PDB select*, we obtain the Q_3 accuracy up to 83.10% and the SOV is 80.02%. These results significantly confirm the improved accuracy of the new encoding method over the conventional ones.

A system to predict protein secondary structure that utilizes this approach has been implemented. This system has been utilized in the Shrimp Genome Project (http://pmonodon.biotec.or.th) to predict the structure of protein sequence translated form unknown gene. The web-service developed from the SVMmer method is available online at http://pmonodon.biotec.or.th/svmmer/.

4.3 Discussion

The major improvement of our approach obtains from the efficiency of our new encoding scheme based on Markov transition matrix. The transition matrix built specifically for each classifier on the secondary structure class plays the significant role to capture the essential local information for mapping patterns of amino acid sequences to protein structures. On another hand, the probability of sequence pattern in *n*-mers that appears in the structural classes of protein is an essential feature used for training the classifiers. This very simple but powerful method uses statistical data for the preprocessing of input vectors. Actually, the extremely attractive feature of the Markov transition matrix is the utilization of structural data from all currently available protein structure via the process of maximum likelihood estimation. Clearly, the continuous growth of such databases should extend further advantage to this approach compared to others and possibly enhance the accuracy.

4.4 Future Work

The further improvement of classifying performance could be increased by two main factors. Firstly, the concept of ensemble network of classifiers as well as the advanced method of multi-class classification could be effectively applied to improve the performance of the predicting model. Secondly, the natural probability of each entry in Markov transition matrix must be achieved by increasing the data samples of amino acid sequence. If the value of each entry in the matrix is close to its ideal value then, the prediction accuracy can obviously be realized. In addition, the optimal length of the amino sequence pattern in the transition matrix is still unknown and further investigation is required.