

Protein Secondary Structure Prediction Accuracy versus Reduction Methods

Saad Osman Abdalla Subair
College of Computing
Al Ghurair University
Dubai, UAE

Safaai Deris
School of Graduate Studies
University of Technology Malaysia, UTM
Skudai, Johor, Malaysia

Abstract

Predicting protein secondary structure is a key step in determining the 3D structure of a protein that determines its function. The Dictionary of Secondary Structure of Proteins (DSSP) uses eight classes to describe a protein. The DSSP database is a database of secondary structure assignments for all protein entries in the Protein Data Bank (PDB) with an algorithm designed to standardize these secondary structure assignments. Five methods that reduce these eight classes into the adopted three classes: alpha helices (H) beta strands (E), and coils (C) are implemented in this research. A protein secondary structure classifier (NN-GORV-II) has been used to evaluate the five reduction methods under the same hardware, platforms, and environment to allow stringent and reliable comparison of these methods and then arrive at a clear conclusion. This paper explains and discusses the effect of these reduction methods on the prediction accuracy and quality.

Keywords: Protein Secondary Structure Prediction, Reduction Methods, Amino Acids, DSSP.

1.0 Introduction

The DSSP database is a database of secondary structure assignments for all protein entries in the Protein Data Bank (PDB). The DSSP program was designed by Kabsch and Sander to standardize these secondary structure assignments [1, 2]. Among other algorithms to conduct the same task of assigning secondary structures are STRIDE and DEFINE algorithms [2]. As described by the DSSP authors, the DSSP works by assigning potential backbone hydrogen bonds which based on the 3D coordinates of the backbone atoms and subsequently by identifying repetitive bonding patterns [1]. The DSSP algorithm classifies each residue into eight classes: H \rightarrow α alpha helix; B \rightarrow residue in isolated β bridge; E \rightarrow extended strand, participates in β ladder; G \rightarrow 3-helix; I \rightarrow 5-helix; T \rightarrow hydrogen bonded turn; S \rightarrow bend; and “.”. The majority of protein secondary structure classifiers use the three states of protein secondary structure (helices (H), strands (E), and coils(C)); this why these eight classes are collapsed or reduced into these three standard classes. The adopted reduction schemes from the mentioned eight states or classes to three classes of helices, strands, and coils are usually performed by using one of the five reduction methods or schemes as will be discussed in the methodology part. The purpose of this study is to learn the effect of the NN-GORV-II [3] algorithm on the five reduction methods and learn the effect of these reduction methods on prediction accuracy and quality. The NN-GORV-II algorithm was developed in previous experiments by combining neural networks and information theory to achieve a better performance protein secondary structure predictor [4].

2.0 Materials and Methods

Cuff and Barton's 513 non redundant proteins which contain 84,107 residues is used for these series of experiments [5]. The CB513 data sets were selected by a stringent definition of sequence similarity or non redundancy, where no two proteins in the set share more than 25% sequence identity over a length of more than 80 residues. The sequences were then filtered to permit only X-ray crystal structures with resolutions of less than or equal to 2.5 Angstroms which in turns reduced to set of 554 domain sequences [5]. A sample from the CB513 protein data set is shown in Figure 1.

```

RES:V,K,D,G,Y,I,V,D,D,V,N,C,T,Y,F,C,G,R,N,A,Y,C,N,E,E,C,T,K,L,K,G,E,S,G,Y,C,Q,W,A,S,P,Y,G,N,A,C,Y,C,Y,K,L,P,D,H,V,R,T,
K,G,P,G,R,C,H,
DSSP:_E,E,E,E,B,B,_T,T,S,_B,_,_S,_H,H,H,H,H,H,H,H,H,T,T,_S,E,E,E,E,E,E,E,T,T,E,E,E,E,E,E,E,_T,T,S,_B,_,_S,S,
_--_
DSSPACC:e,e,e,b,b,b,b,e,e,e,b,b,b,b,e,b,e,e,e,e,b,e,e,b,b,e,e,e,b,e,e,b,b,b,b,e,e,b,b,e,e,b,e,e,e,e,e,
STRIDE:C,E,E,E,E,B,B,T,T,T,T,C,B,C,B,C,C,C,H,H,H,H,H,H,H,H,H,C,C,C,E,E,E,E,E,E,E,T,T,E,E,E,E,E,E,E,T,T,T,T,C,B,
C,C,C,C,C,C,C,
RsNo:1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,4
6,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,
DEFINE:_,_E,E,E,E,E,_,_E,E,E,E,H,H,H,H,H,H,H,H,H,_E,E,E,E,_E,E,E,E,E,E,_E,E,E,E,_E,
E,E,E,E,
align1:V,K,D,G,Y,I,V,D,D,V,N,C,T,Y,F,C,G,R,N,A,Y,C,N,E,E,C,T,K,L,K,G,E,S,G,Y,C,Q,W,A,S,P,Y,G,N,A,C,Y,C,Y,K,L,P,D,H,V,R,
T,K,G,P,G,R,C,H,
align2:K,R,D,G,Y,I,V,Y,P,N,N,C,V,Y,H,C,V,P,....P,C,D,G,L,C,K,K,N,G,G,S,S,G,S,C,S,F,L,V,P,S,G,L,A,C,W,C,K,D,L,P,D,N,V,P,I,K,
D,R,K,..C,T,
.
.
align33:V,R,D,G,Y,I,A,Q,P,H,N,C,A,Y,H,C,L,K,S,S,G,C,D,T,L,C,K,E,N,G,A,T,S,G,H,C,G,H,K,S,G,H,G,S,A,C,W,C,K,D,L,P,D,K,V,G
,I,I,V,E,K,..C,H,
align34:V,R,D,G,Y,I,A,Q,P,H,N,C,V,Y,H,C,F,P,S,G,G,C,D,T,L,C,K,E,N,G,A,T,Q,G,S,S,C,F,I,L,G,R,G,T,A,C,W,C,K,D,L,P,D,R,V,G,V
,I,V,E,K,..C,H,

```

Figure 1: An example of a flat file of CB513 data base

2.1 Reduction of DSSP Secondary Structure States

The data set used in this research was downloaded from the web site <http://barton.ebi.ac.uk/> and then extracted to Linux Red Hat 9 platform where the whole experiment is conducted. The adopted reduction schemes of the mentioned eight classes to three states of helix (H), strands (E), and coil (C) is more often performed by using one of the following schemes or methods [5,6,7].

1. *Method I*: {H, G, I} → H; {E} → E; the rest → C.
2. *Method II*: {H, G} → H; {E, B} → E; the rest → C.
3. *Method III*: {H, G} → H; {E} → E; the rest → C.
4. *Method IV*: {H} → H; {E, B} → E; the rest → C.
5. *Method V*: {H} → H; {E} → E; the rest → C.

In this study, all the above mentioned schemes are attempted to learn their effects on prediction performance and quality. The 8-to 3-state reduction scheme can alter the prediction accuracy of an algorithm in a range of 1-3% [2]. PERL (Practical Extraction and Reporting Language) under Red Hat 9 Linux Environment is used to extract and parse the amino acids sequences or residues (RES) into corresponding files forming standard FASTA format. The corresponding laboratory determined DSSP predictions of the residues are extracted and parsed into other files that contain the predicted sequences from the seven algorithms. The resulting final files are flat files that contain the amino acid sequence (AA), the predicted secondary structure (PSEQ), and the observed secondary structure (OSEQ) after being reduced into three state schemes (Figure 2). PERL is used to convert these files into format that is readable by the Q₃ and SOV (Segment Overlap) program [8] to evaluate the prediction accuracy of each method. PERL is also used to convert the names of these files into format that is readable by CLUSTALW and PSIBLAST programs.

```

>OSEQ
CEEEEECCCCCECCCCCHHHHHHHHHHCCCCEEEEEEEECEEEEEEEEECCCCCECCCC
CCC
>PSEQ
CCCCEEEECCCCCEEECCCCCCCCCHHHHCCCCEEEEEEEECCCCCEEEEEEEEECCCCCECCCC
CCC
>AA
VKDGYIVDDVNCTYFCGRNAYCNEECTKLKGESGYCQWASPYGNACYCYKLPDHRVTRTKG

```

Figure 2: The FASTA format file parsed into a format readable by the Q₃ and SOV programs

2.2 Measure of Performance and Quality of Prediction

The Q_3 accuracy per residue which measures the expected accuracy of an unknown residue is computed as the number of residues correctly predicted divided by the total number of residues. The Q_H is defined as the total number of α helix correctly predicted divided by the total number of α helix. The same definitions are applied to Q_E (β strands) and Q_C (coils). The Q_3 is expressed as:

$$Q_3 = \sum_{(i=H,E,C)} \frac{\text{predicted}_i}{\text{observed}_i} \times 100 \quad (1)$$

Segment overlap (SOV) calculation [8, 9] is performed for each data set. Segment overlap values attempt to capture segment prediction. The SOV aims to assess the quality of a prediction by taking into account the type and position of secondary structure segment, the natural variation of segment boundaries among families of homologous proteins, and the deviation at the end of each segment. SOV is calculated by:

$$Sov = \frac{1}{N} \sum_s \frac{mnov(S_{obs}; S_{pred}) + \delta}{mxov(S_{obs}; S_{pred})} \times len(s_1) \quad (2)$$

Where:

N : the total number of residues,

$mnov$: the actual overlap, with $mxov$ is the extent of the segment.

$len s_1$: is the number of residues in segment s_1 .

S_{obs} : Observed Segment

S_{pred} : Predicted Segment

δ is: the accepted variation where there are only minor deviations at the ends of segments.

3.0 Results and Discussion

The mentioned reduction methods are well established for a long time. It was reported that the eight-to-three state reduction scheme can alter the prediction accuracy of an algorithm in a range of 1-3% [5] Table 1 shows the numbers of helices, strands, and coils according to each of the five reduction methods from the eight states to the three states. A PERL program under LINUX environment was developed to make these assignments and count the number of the total residues in the database and then the numbers and the ratio of each secondary structure state. It can be observed that the percentage assigned as coils increases gradually from Method I to Method V.

Table 1: Percentage of secondary structure state for the five reduction methods of the DSSP definition

<i>Reduction Method</i>	<i>Helix</i>		<i>Strands</i>		<i>Coils</i>	
	<i>Number</i>	<i>%</i>	<i>Number</i>	<i>%</i>	<i>Number</i>	<i>%</i>
<i>Method I</i>	28851	35	18951	23	35590	43
<i>Method II</i>	28881	35	17810	21	36701	44
<i>Method III</i>	28851	35	17810	21	36731	44
<i>Method IV</i>	25807	31	18951	23	38634	46
<i>Method V</i>	25807	31	17810	21	39775	48

Figure 3 shows how the 480 amino acids or proteins had been predicted and distributed through the different levels of Q_3 predictions by the five different reduction methods. Figure 3 elucidated that the performance accuracy Q_3 for Method V predicted just below 250 of the 480 proteins tested at the level of 80-90%, just above 100 proteins for the level of 70-80%, and below 100 proteins for the 90-100%. Method IV had a similar pattern of Method V, while other three reduction methods predicted just above 200 proteins at the 80-90% level.

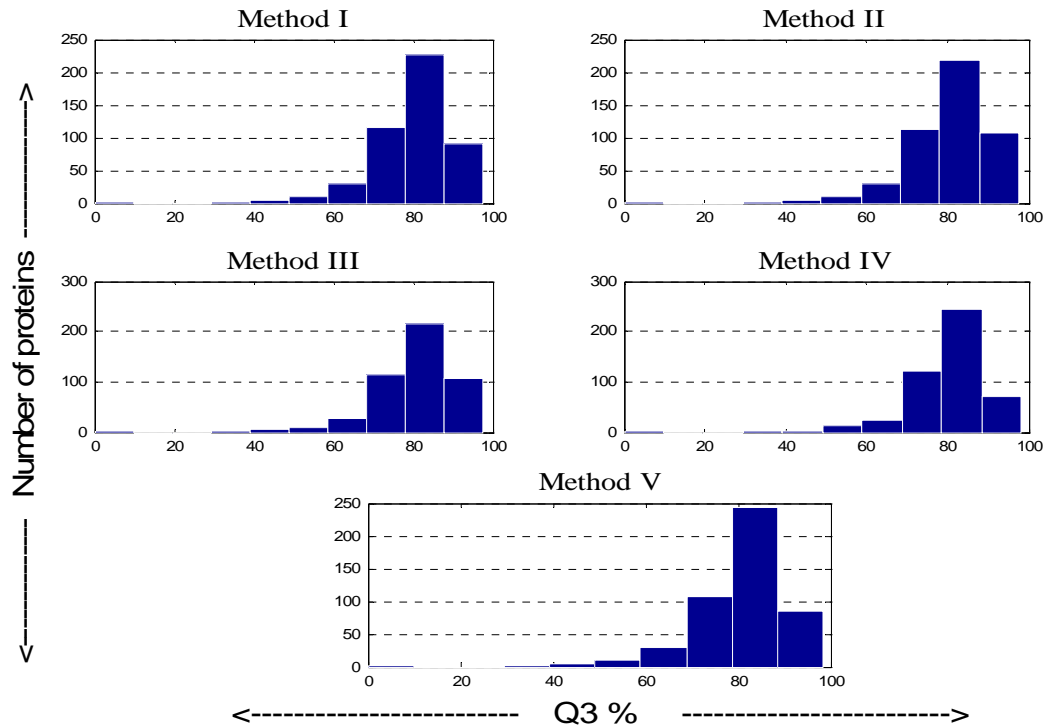


Figure 3: Five histograms showing the Q_3 distribution of the test proteins with respect to the five reduction methods

Table 2 shows the results of one way analysis of variance procedure (ANOVA) against the performance of prediction accuracy (Q_3) of the five reduction methods. The ANOVA procedure tests for the hypothesis that whether ever all means of the five methods are similar or whether there are significant differences between them. In other words, the importance of this test is to accept or reject the fact that the means of the performance of the five reduction methods differ significantly at the 0.05 or 0.01 probability level or not. The same ANOVA test has been conducted for the SOV of the five reduction methods.

Table 2: The analysis of variance procedure (ANOVA) of the Q_3 for the five reduction methods*

<i>Method</i>		<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F-test</i>	<i>Significance</i>
<i>Method II</i>	Between Groups	49578.977	252	196.742	122.356	.000
	Within Groups	365.003	227	1.608		
	Total	49943.980	479			
<i>Method III</i>	Between Groups	49633.031	252	196.956	132.267	.000
	Within Groups	338.023	227	1.489		
	Total	49971.053	479			
<i>Method IV</i>	Between Groups	44528.264	252	176.699	29.473	.000
	Within Groups	1360.915	227	5.995		
	Total	45889.180	479			
<i>Method V</i>	Between Groups	45300.225	252	179.763	24.194	.000
	Within Groups	1686.648	227	7.430		
	Total	46986.873	479			

* Method I is control

Table 2 presents the results of the five reduction methods. It shows that the means are significantly different from each others at the 0.001 probability level, as far as their performance accuracies are concerned. This probability level suggested that we are more than 99% sure that these methods differ from each others. The same conclusion applies for SOV (Table not shown), that the five reduction methods are significantly different from each other as far as their SOVs are concerned. This elucidates that the five reduction methods are significantly different in their performance, prediction quality, and usefulness.

To explore the effect of the five reduction methods on the NN-GORV-II algorithm performance, Table 3 shows the scores of the helices (Q_H), strands(Q_E), coils(Q_C), and all the states together (Q_3) with respect to each reduction method. The performances of helices (Q_H) are almost the same and about 77.4% with standard deviations of 26.53% for all the first three methods, I, II, and III. The performances of the helices (Q_H) for Method IV and Method V are 87.03 with standard deviations 20.57 for each. There is about 10% Q_H increase in predicting helices for methods IV and V compared to methods I, II, and III. This increase in Q_H accuracy is accompanied by a 6% decrease in the standard deviations for methods IV and V. This result proves that methods IV and V predicted helices more accurately and the prediction is more homogenous compared to the other three methods I, II, and III. The strands (Q_E) prediction accuracies are 77.12% with standard deviations of about 12% for methods II, III, and V while strands predictions are 69.49% with standard deviations of 27.42% for methods I and IV.

This reveals that strands predictions have higher accuracies and are more stable and homogenous for methods II, III, and V in comparison with the other two methods. It had been reported in the literature that beta strands are difficult to predict as compared to the other two states [10]

Table 3: The effect of the five reduction methods on the performance accuracy of prediction (Q_3) of the of NN-GORV-II prediction method.

<i>Reduction Method</i>	Q_3	Q_H	Q_E	Q_C
<i>Method I</i>	79.88±10.13	77.42±26.53	69.49±27.42	80.31±11.77
<i>Method II</i>	80.49±10.21	77.40±26.53	77.12±24.19	79.99±11.75
<i>Method III</i>	80.48±10.21	77.42±26.53	77.12±24.19	79.96±11.77
<i>Method IV</i>	80.38±9.79	87.03±20.57	69.49±27.42	78.34±11.78
<i>Method V</i>	80.98±9.90	87.03±20.57	77.12±24.19	78.07±11.76

As for the coils states prediction accuracy (Q_C), Table 3 shows that methods I, II and III scored about 80% prediction accuracies with standard deviations of 11% each while the prediction for the coil states scored about 78% with standard deviations of about 11% for methods IV, V each. This result proves that methods IV and V predicted the coil states with less accuracy but with the same stabilities and homogeneities compared to the other three methods.

Considering the overall prediction accuracies (Q_3) for the five reduction methods, Table 3 shows that Method I recorded the least accuracy of 79.88% while Method V recorded the highest accuracy which is 80.98%. The other three methods recorded accuracies of 80.49%, 80.48%, and 80.38% for methods II, III, and IV, respectively. The standard deviations for all the five methods are almost the same and are around 10% which showed comparatively small standard deviations that reflected homogenous and stable predictions for all the five reduction methods. This observation is confirmed in Figure 4 which shows the trend of predicting the 480 proteins using the different five reduction methods. The graph portrays that the five reduction methods performed in more or less similar trend and the margin differences between the five methods are very small.

By further elaboration to Table 3, it is clear that Method I records the most rigorous and least accurate performance in assessing the NN-GORV-II algorithm. In contrast, Method V shows the highest accuracy demonstrating that it is the most optimistic method of assessing prediction algorithms. The difference in accuracy prediction (Q_3) between Method I and V is 1.1% which is a considerable difference. It reflects a true difference in evaluating prediction algorithms since this difference resulted from experiments conducted in exactly the same environments. This result is consistent with [5, 11] in leading the conclusion that different reduction methods can affect the prediction accuracy of an algorithm. Method II had a medium score between methods I and V, while having similar pattern score to methods III and IV.

Similar to Table 3, the overall segment overlap (SOV_3) measures for the five reduction methods can be summarized as follows: Method II and III achieve overall SOV_3 of 76.3% with standard deviations of 17.5 each. Method I and IV score SOV_3 of 75.8% with standard deviations of about 16% each while Method V achieves an overall SOV_3 of 74.93% with standard deviations of 18.78% (Table not shown).

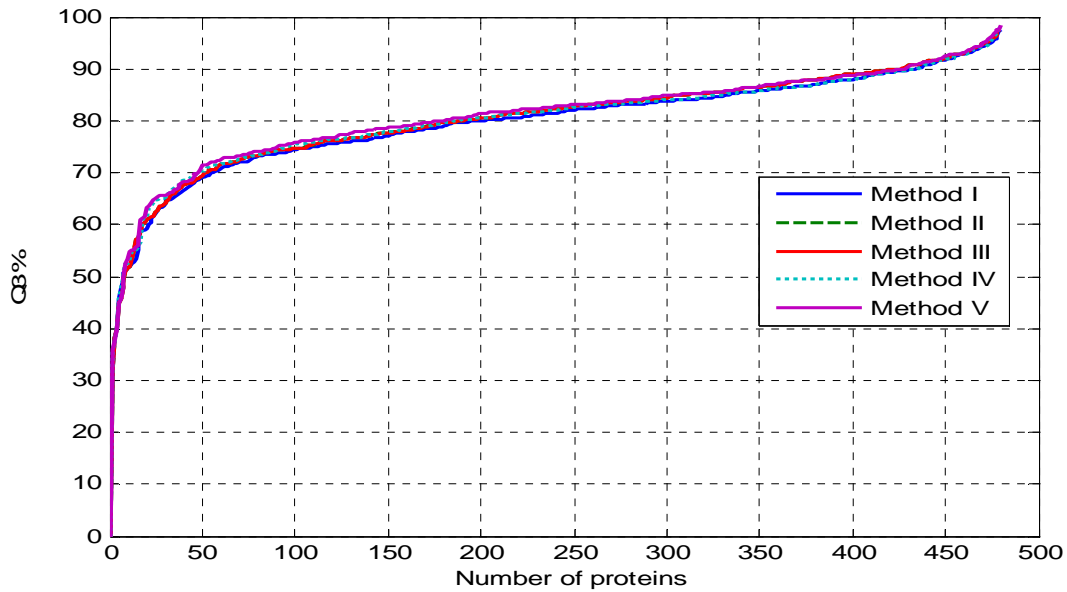


Figure 4: The performance accuracy (Q_3) of the five reduction methods on the test proteins

The SOV table concluded similar results that indicate methods IV and V predictions for the helices states are of higher quality and stability compared to the other three methods. The results are further elucidated in Figure 5. The graph of Figure 5 portrays that the SOV of the five reduction methods performed in a similar trend to the Q_3 performance.

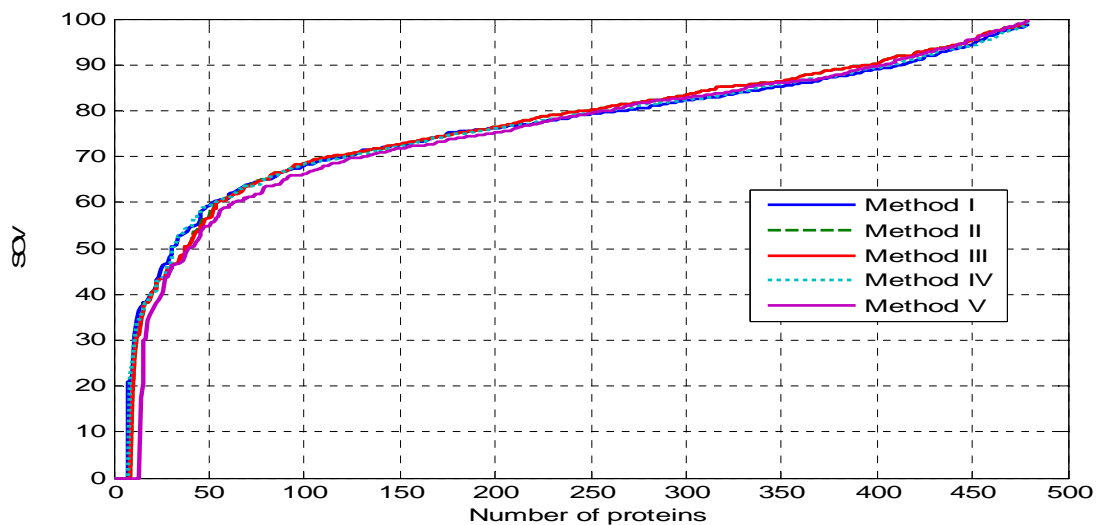


Figure 5: The SOV measure of the five reduction methods on the test proteins

Although these two graphs (Figure 4 and Figure 5) show similar results for Q_3 and SOV_3 , the ANOVA analysis proved that they are significantly different. These results reveal that methods II and III predict the secondary structures of proteins with high quality and more usefulness while methods I and IV predict the proteins with comparatively less quality and usefulness. However, Method V had achieved the highest apparent performance (Q_3) in prediction accuracy (Table 3). Method V as well had achieved the least SOV_3 and hence the least quality of prediction compared to the other five reduction methods. The above results also conclude that Method I and Method II showed higher quality and more usefulness.

4.0 Conclusion

Five reductions methods that assign the DSSP eight protein secondary structural classes into the commonly used three structural classes or states are attempted in this study. The number of helices, strands, and coil states are affected by different reduction methods. The one way analysis of variance (ANOVA) procedure showed that the five reduction methods varied significantly in their performance (Q_3) and quality (SOV_3) of predicting protein secondary structures. Method I is the most pessimistic in its performance result while Method V is the most optimistic. Using method I will make a reliable and practical evaluation of a novel protein secondary structure classifier rather than using Method V. Method II is in middle performance between method I and V.

5.0 Acknowledgments

The idea of this research was originated at the University of Technology Malaysia (UTM). The authors would like to thank Al-Ghurair University in Dubai for sponsoring this research work. The authors would like to thank Dr Mohamed Elfatih Hamad, Public Health Department, Dubai Municipality for reviewing the manuscript

6.0 References

- [1] Kabsch, W. and C. Sander, "On The Use of Sequence Homologies to Predict Protein Structure: Identical Pentapeptides Can Have Completely Different Conformations", *Proceedings of the National Academy of Science, USA*, 81: 1075-1078. 1984.
- [2] Cuff, J.A. and G. J. Barton, "Application of Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction", *Proteins: Structure, Function and Genetics*, 40: 502-511. 2000.
- [3] Subair S.O. and S. Deris, "Combining Artificial Neural Networks and GOR-V Information Theory to Predict Protein Secondary Structure from Amino Acid Sequences", *International Journal of Intelligent Information Technologies*, Idea Group Publishing, USA, 53-72. 2005.
- [4] Abdalla (Subair), S. O. and Deris, S. "An Improved Method for Protein Secondary Structure Prediction by Combining Neural Networks and GOR V Theory". *Second Middle East Conference on Healthcare Informatics (MECHCI 2005)*. Dubai Knowledge Village, Dubai, UAE. 9-10 April 2005.
- [5] Cuff, J. A. and Barton, G. J. "Evaluation and Improvement Of Multiple Sequence Methods For Protein Secondary Structure Prediction". *Proteins: Structure, Function and Genetics*. 34: 508-519.1999.
- [6] Pollastri, G., Przybylski, D., Rost, B., Baldi, P. "Improving The Prediction Of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles". *Proteins: Structure, Function, and Genetics, Supplement*. 47: 228-235. 2002.
- [7] Frishman, D. and Argos, P. "Knowledge-Based Protein Secondary Structure Assignment". *Proteins: Structure, Function, and Genetics, Supplement*. 23:566-579. 1995.
- [8] Zemla, A., C. Venclovas, K. Fidelis, and B. Rost, "A Modified Definition of SOV: A Segment Based Measure for Protein Secondary Structure Prediction Assessment", *Proteins: Structure, Function, and Genetics, Supplement*, 34: 220-223. 1999.
- [9] Rost, B. R., Sander, C. and Schneider, R. "Redefining the Goals of Protein Secondary Structure Prediction". *Journal of Molecular Biology*. 235: 13-26. 1994.
- [10] Ouali, M. and King, R. D "Cascaded Multiple Classifiers For Secondary Structure Prediction". *Protein Science*, 9: 1162-1176. 2000.
- [11] Rost, B. "Protein Secondary Structure Prediction Continues to Rise". *J. Struct. Biol*, 134: 204-21. 2001.