CASP Targets Are Reliable Test

For a Protein Secondary Structure Classifier

Saad Subair

College of Computer and Information Sciences,
Princess Nourah bint Abdulrahman
University, Riyadh KSA

ABSTRACT

A classifier for predicting protein secondary structure from amino acid sequences has been proposed and implemented in a previous experiment. NN-GORV-II classifier utilizes the power of Artificial Neural Network and GOR method of protein secondary structure prediction. The Critical Assessment of techniques for Structure Prediction of proteins (CASP) experiments aim at establishing the current state of the art in protein structure prediction. The NN-GORV-II classifier is tested using CASP targets proteins. This test is based on testing a new protein classifier with proteins targets (amino acids) that were never used by the classifier at any prior training or testing stages, hence it's known as blind test. This type of prediction was described as true prediction. The performance of the NN-GORV-II method on the CASP targets: (Q_3) is 76.9% with 7.5% standard deviation while the quality of the prediction (SOV_3) of the method reached 75.4% with 9.8% standard deviation. The Correlation Coefficients are 0.68, 0.63, and 0.62 for helices, strands, and coils, respectively, indicating strong relationship between predicted and observed secondary structures states.

Keywords: Bioinformatics, Protein Secondary Structure Prediction, Blind Test, Independent Test CASP.

1.0 Introduction

Proteins are series of amino acids known as polymers linked together into contiguous chains [1]. Protein has three main structures: *primary structure* which is essentially the linear amino acid sequence. *Secondary structure* which are α helices, β sheets, and coils that are formed when the sequences of primary structures tend to arrange themselves into regular conformations*Tertiary or 3D structure*: where secondary structure elements are packed against each other in a stable configuration [2,3]

Advances in molecular biology in the last few decades lead to the rapid sequencing of considerable genomes of several species. The need for computational methods rather than laboratory techniques alone to predict protein structure becomes inevitable. GOR method was first proposed by Garnier et al. [4]and named after its authors Garnier, Osguthorpe, and Robson. The GOR method is based on the information theory and naive statistics and it has been a standard method for many years [5, 6]

Artificial Neural networks have been used successfully in the prediction of proteins secondary structures. Since the neural network can be trained to map specific input signals or patterns to a desired output, information from the central amino acid of each input value can be modified by a weighting factor, then grouped together and sent to a second level (hidden layer) where the signal is clustered into an appropriate class [7,8,9,10]. Several secondary structure classifiers or predictor use neural network alone or neural network combined with other method or methods [11, 12].

The NN-GORV-II prediction method developed in this work depends on the statistical assumption that combining relevant information in different prediction or classification methods will possibly increase the prediction accuracy of the combined method [13]. The NN-GORV-II method is a protein secondary structure classifier developed by combing the GOR method and neural networks using a filtering mechanism [14].

As described by their founder, the Critical Assessment of techniques for protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused. The goal of CASP experiments is to obtain an in depth and objective assessment of the current abilities and inabilities in the area of protein structure prediction. In the competition, participants will predict as much as possible about a set of soon to be known structures. This type of prediction was described by CASP initiators as true prediction than prediction made on already known proteins[15,16,17]

2.0 Materials and Methods

The development of NN-GORV-II algorithm was a long process described in details in [14]. The NN-GORV-II developed as an improved version of NN-GORV-I by using a filtering mechanism to the data base. The performance of the NN-GORV-II classifier was found of superior performance and quality when compared to several methods and classifiers studied[14]. .Since it is based on data or targets that have never been seen by the NN-GORV-II classifier, the test on these targets is essential to assess the reliability and partiality of this classifier.

CASP3 targets are used in this independent or blind test which represents protein sequences that have never been used in training or testing the NN-GORV-II classifier. The importance of these CASP3 proteins is that they are classified by the CASP organizers as proteins with no homologous sequences of known structure.

In this experiment, 42 CASP3 target proteins are extracted with their secondary structure predicted using the PHD [18] program. It is not possible for this experiment to find predicted or observed CASP4 or CASP5 targets which are more recent and hence CASP3 was used to give an idea about the independent test set performance.

Several assessment measures and methods are used in this work to estimate the prediction accuracy of the NN-GORV-II algorithm developed and studied in this work.. The methods implemented to assess the accuracy of performance and quality of the prediction using CASP targets.

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 ratio 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 factor is expressed as:

$$Q_3 = \sum_{(i=H,E,C)} \frac{predicted_i}{observed_i} x100 \quad (1)$$

Segment overlap measure (SOV) calculation [19, 20] is calculated for the CASP targets. Segment overlap values attempt to capture segment prediction, and vary from an ignorance level of 37% (random protein pairs) to an average of 90% level for homologous protein pairs. The SOV aims to assess the quality of a prediction rather than performance. Segment overlap 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* is the total number of residues,mnov is the actual overlap, and mxov is the extent of the segment.

*len*s1 is the number of residues in segment s1. δ is the accepted variation in segments.

The Q_3 and SOV measures are calculated using the SOV program downloaded from the web site: http://PredictionCenter.llnl.gov[21].

Matthews' correlation coefficient (MCC) is performed for each of the three states. Calculating the four numbers (p_i, r_i, u_i, and o_i). The formula of Matthews's correlation [22] can be rewritten as:

$$C_i = \frac{p_i r_i - u_i o_i}{\sqrt{(p_i + u_i)(p_i + o_i)(r_i + u_i)(r_i + o_i)}} \quad (3)$$

Where:

p_i number of correctly predicted residues in conformation.,

r_i number of those correctly rejected.

u_i number of the incorrectly rejected (false negatives).,

o_i number incorrectly predicted to be in the class (false positive),

i = is one of the confirmation states H, E, or C.

3.0 Results and Discussion

The results of the blind test on the 42 CASP targets are discussed in details in this section. Figure 1 and Table 1 show the distribution of the 42 CASP proteins predicted using the NN-GORV-II algorithm for all the threesecondary structure states. For the helices states, the histogram of Figure 1 shows that about 18 proteins (targets) are predicated at Q_H of above 95% and more than 5 proteins predicted at 85%, 75%, and 65% each.

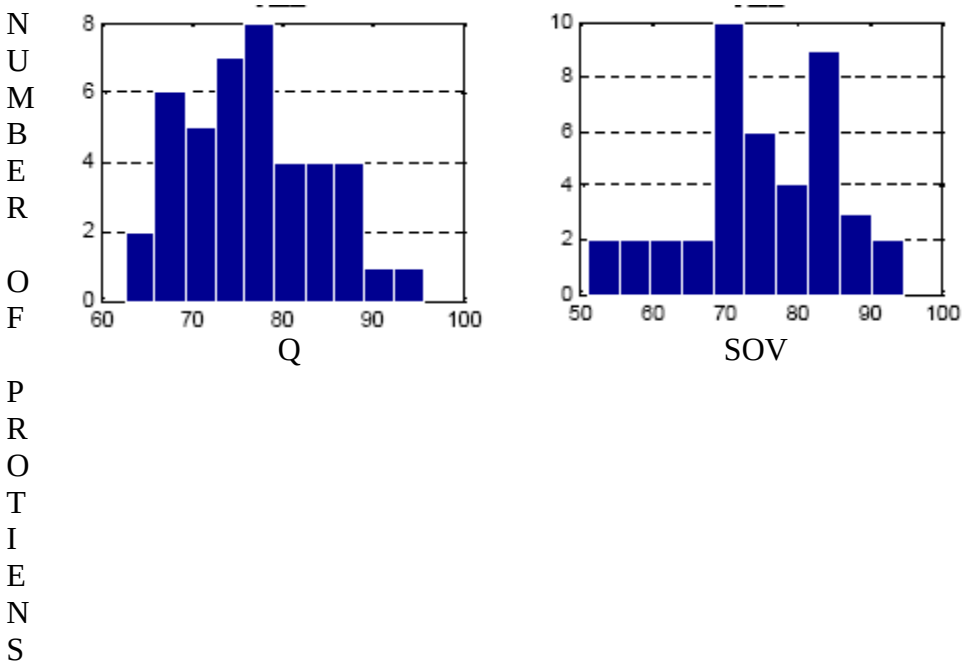

Figure 1: The distribution of prediction actuariesQ_3 and SOV of the 42 CASP targets

Less than three proteins are predicted at 55% and about two proteins predicted at 45%, 35%, and 5%. The strands prediction accuracies (Q_E)are 8 proteins predicted at 95%, 6 proteins predicted at 85% and 5% each, and 7 proteins are predicted at 75%, and 65%. The rest of the proteins are predicted at 55% Q_E level and below. As for coils, Figure 1 shows that the about 15 proteins are predicted at level of 70-80% Q_C, about 13 proteins at level of 60-70%, and about 10 proteins at level of 80-90%. The rest three proteins are predicted at level 90-100%. The SOV distributions show similar results to what is seen in Figure 1.

Figure 2 shows the results of the blind test in a line graph. The figure elucidates that the helices (Q_H) and strands (Q_E) lines travelled from the zero prediction while coils (Q_C) and the overall performance (Q_3) travelled from below 60% and above 60%, respectively.The histogram of Figure 1 and the line graph of Figure 2 show that the strands states are predicted by the NN-GORV-II in a more scattered distribution followed by the helices states while the overall prediction (ALL) was more homogenous and continuous followed by the coils states prediction. The results elucidated that the majority of protein are predicted at Q_3 accuracies between 70-80%.

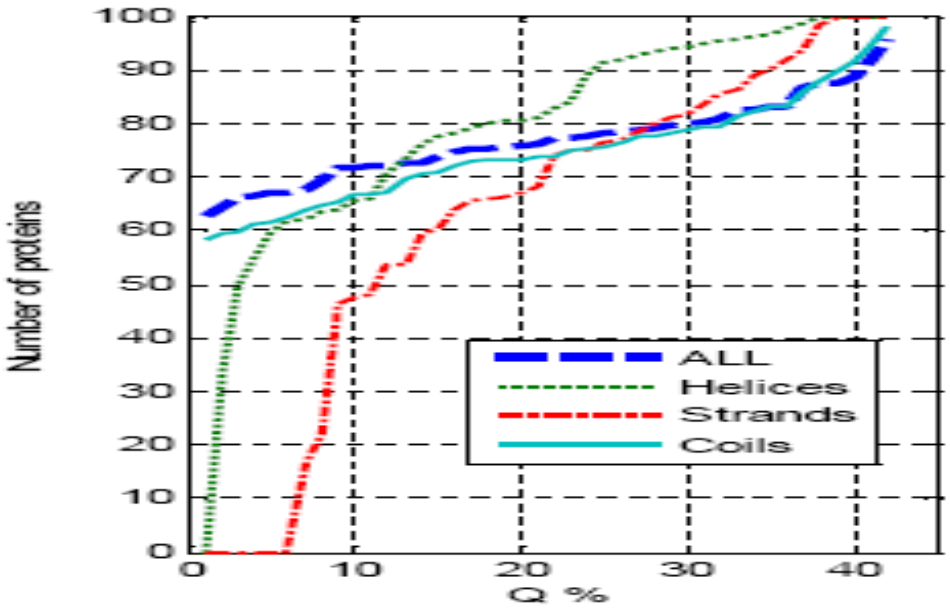

Figure 2: The performance of the 42 CASP targets with respect to Q_3 prediction measures

Table 1 shows all the values of Qand SOV. The overall prediction accuracies (Q_3 or ALL) for the 42 CASP targets are shown in Table 1. About 8 proteins are predicted at Q_3 accuracy between 60% and below 70%, about 20 proteins predicted at accuracy of 70-80%, about 12 proteins are predicted at Q_3 of 80-90%, and about two proteins predicted at accuracies above 90% and below 100%. It is clear that there is no protein predicted at accuracy below 60% of Q_3. These results are supported by the line graph of Figure 2 where each line indicates a secondary structure state travelling towards the 100% accuracy through the 42 CASP targets.

Table 1: Percentages of prediction accuracies and SOV measures for the 42 CASP3 proteins targets

ID	Protein Name	Q_3	Q_H	Q_E	Q_C	SOV_3	SOV_H	SOV_E	SOV_C
T0085	Cytochrome C554, Nitrosomonaseuropaea	72	65.8	22.2	83.3	72.9	82.7	27.8	73.5
T0084	RLZ, artificial construct	91.9	100	100	62.5	69.5	68	100	75
T0083	Cyanase, E.coli	83.3	77.5	100	90	81.9	75.2	86.8	93.8
T0082	Ribonuclease MC1, Momordicacharantia (Bitter Gourd)	77.4	81.2	75	76.4	67.4	62.3	76.4	68.1
T0081	Methylglyoxal synthase, E. coli	71.7	73.4	64	73	73.5	95.3	72	56.8
T0080	3-methyladenine DNA glycosylase, human	72.6	65.6	75	73.3	65	94.1	67.5	59.8
T0079	MarA protein, E. coli	79.8	92	0	66.7	85.3	100	0	68.8
T0078	Thioesterase, E. coli	67.7	82.8	76.4	58.2	64	84.1	68	56.2

T0077	Ribosomal protein L30, Saccharomyces cerevisiae	76.2	94.3	74.2	61.5	86	98.2	83.9	76.7
T0076	cdc4p, Schizosaccharomycespombe	95.7	96.5	100	94.5	94.7	100	100	87.5
T0075	Ets-1 protein (fragment), mouse	82.7	80	0	87.8	85	79.4	0	95.1
T0074	The second EH domain of EPS15, human	88.8	97.7	100	81.5	85.3	98.5	100	77.6
T0072	CD5 domain 1, human	78.2	63.6	60.5	91.8	79.9	90.9	71.8	82.9
T0071	Alpha adaptin ear domain, rat	75.2	59.4	88.9	75.5	82	67.4	84.3	89.1
T0070	Omp32 protein, Comamonasacidovorans	73.8	0	86.3	65.4	64.1	0	84	53.8
T0069	Recombinant conglutinin, bovine	78.8	91.3	77.1	72	73	100	82.9	58
T0068	Polygalacturonase, Erwinia carotovora subsp. carotovora	78.5	100	83.5	73.7	74.4	39.6	81	70.7
T0067	Phosphatidylethanolamine Binding Protein, Homo sapiens	75.9	100	68.4	77.5	80.6	82.6	76.1	82.4
T0065	B SinIprotein, Bacillus subtilis	87.7	96.3	0	85.7	85.2	100	0	77.1
T0064	A SinRprotein, Bacillus subtilis	77.5	92.7	0	79.5	82.6	100	0	83.4
T0063	Translation initiation factor 5A, Pyrobaculumaerophilum	75.4	88.9	90.3	59.7	72.5	100	84.6	59.2
T0062	Flavin reductase, E. coli	83.2	80.6	93.8	78.1	90.3	86.2	97.1	89
T0061	Protein HDEA, E. coli	66.3	78.3	16.7	59.5	59.7	60.8	8.3	66.2
T0060	D-dopachrometautomerase, human	80.3	93.5	81.6	70.8	90.9	100	92.8	83.6
T0059	Sm D3 protein (The N-terminal 75 residues)	82.7	100	85.4	79.4	76.2	100	70.2	85.3
T0058	Uracil-DNA glycosylase, E.coli	79.9	95.7	59.6	78.8	78.3	99.1	70.2	70.7
T0057	Glyceraldehyde 3-phosphate dehydrogenase, S. solfataricus	67.1	64	67	69.6	72.2	66	70.6	79.4
T0056	DnaB helicase N-terminal domain, E.coli	86.8	98.6	0	73.2	79.5	98.4	0	61
T0055	lectin, Polyandrocarpamisakiensis	67.2	70.6	65.9	67.2	59.8	82.4	75.4	50.7
T0054	VanX, Enterococcus faecium	75.7	76.3	48	82.2	70.4	79.6	56	67.3
T0053	CbiK protein, S. typhimurium	72.7	80.5	65.6	61.4	71.1	84.7	67.7	53.2
T0052	Cyanovirin-N, Nostocellipsosporum	64.4	50	53.8	77.6	69.4	71.4	68.3	68.9
T0051	Glutamate mutase component E - Clostridium cochlearium	74.7	84.1	53.8	70.7	72.8	91.8	64	58.1
T0050	Glutamate mutase component S - Clostridium cochlearium	69.3	95.6	47.6	64	75.8	90.2	63.1	73.6
T0049	EstB, Pseudomonas marginata	71.7	79.2	46.2	73.9	51.1	91	38.4	43.4
T0048	Pterin-4-alpha-carbinolamine dehydratase,	62.7	54.8	80	73.3	70.6	65.4	73.3	81.7

	Pseudomonas aeruginosa								
T0047	Alpha(2u)-Globulin	87.7	100	98.5	75.3	86.6	100	100	74.5
T0046	Gamma-Adaptin Ear Domain	79	33.3	92	75	82.7	44.4	94.9	78.8
T0045	HI1434	77.2	61.8	78.6	98.1	82.9	80.3	81	87.5
T0044	RNA-3'terminal phosphate cyclase	72	94.9	66.1	64.9	76.7	86.4	68.2	78.4
T0043	7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK)	66.5	62	81	66.7	55	69.1	73.6	41.4
T0042	NK-lysin from pig, 78a.a.	80.8	94.1	0	83.3	71.6	73	0	100

Table 1 shows the performance of the NN-GORV-II method predicting the three secondary structures states: helices (Q_H), strands (Q_E), and coils (Q_C); and the overall accuracies (Q_3) of the 42 CASP targets. The observed secondary structure predictions of the 42 targets are referenced to the PHD predications of these target sequences as mentioned in the methodology. This independent test portrays a real view about the NN-GORV II algorithm predictions of data that has not been used in its training or testing procedures.

Table 1 also shows the SOV measures of the NN-GORV-II method predicting the three secondary structures states: helices (SOV_H), strands (SOV_E), and coils (SOV_C); and the overall accuracies (SOV_3) of the 42 CASP targets. It is important to note that the SOV measures had been estimated using the same data used in estimating the performance accuracy (Q), and the same program as discussed in the methodology. Since the predicted secondary structures of the 42 targets of the PHD program are used here as observed structures, care should be taken when comparing the performances (Q_3) and qualities (SOV_3) of NN-GORV-II method with other prediction methods or classifiers(Table 1). An Ideal blind test should refer to sequence or targets that are predicted in molecular biology laboratories. However, the PHD program is a stringent and well established classifier that can be taken as reference.

Table 2 shows the mean performance (Q), the SOV measure, and the Mathew's Correlation Coefficients (MCC) of the NN-GORV-II method on the 42 CASP target sequences with the corresponding standard deviations. The values in the table confirmed what has been discussed previously in the above figures and tables. Since they exhibit higher standard deviations, the strand states predictions have a higher variability and less homogeneity followed by the helices states. On the other hand the coils states exhibit less standard deviation and hence predicted in a continuous and homogenous pattern or distribution as seen in Figure 1.

Table 2: The mean of Q_3 and SOV with the standard deviations, and Mathew's Correlation Coefficients (MCC) of CASP targets

Measure	**ALL**	**H**	**E**	**C**
Q	76.87 ± 7.52	79.69 ± 20.75	62.45 ± 31.10	74.58 ± 09.80
SOV	75.44 ± 9.75	81.87 ± 20.62	63.81 ± 31.03	72.33 ± 12.83
MCC	-	0.68	0.63	0.62

As shown in Table 2 which summarizes tables 8.1 and 8.2, the performance of the NN-GORV-II method on the 42 CASP targets (Q3) is 76.87% with a small standard deviation of 7.52% while the quality and usefulness (SOV3) of the method reached 75.44% with relatively small standard deviation of 9.75%. The Mathew's Correlation Coefficients (MCC) is 0.68, 0.63, and 0.62 for helices, strands, and coils, respectively, indicating strong relationship between predicted and observed secondary structures states [23, 24]

The results of this work reflect a practical test of the NN-GORV-II method performance and quality on an independent test set. The values and the results confirmed what has been discussed in a previous work that the NN-GORV-II method is a classifier with high accuracy and quality of prediction [14, 25]

4.0 Conclusion

This work assesses the performance and quality of the prediction of the NN-GORV-II classifier by using an independent test set of protein data that has not been used in training and testing the algorithm. CASP3 protein targets had been used for this purpose. The result of the test gives an empiricalresult of the prediction performance and quality of the NN-GORV-II method despite the limitation of the data set. The blind test proved that it's practical and reliable.The observed secondary structures states of these target sequences are determined by the PHD method and not laboratory methods; so a straightforward comparison with other methods might not be an accurate comparison. The NN-GORV-II method performance accuracy in predicting protein secondary structure and the quality of prediction are far better than many results reported by many researchers.

5.0 Acknowledgments

The author would like to appreciate the assistance rendered by Princess Nourah bint Abdulrahman University (PNU), Riyadh, KSA regarding this research work. The author would like also to thank all the colleagues in the college of Computer and information science at PNU.

6.0 References

[1] Branden, Candtooze, J. (1991). *Introduction To Protein Structure*. Garland Publishing, Inc.: New York.

[2] Pauling, L. and Corey, R. B. (1951). Configurations Of Polypeptide Chains With Favoured Orientations Around Single Bonds: Two New Pleated Sheets. *Proc. Natl. Acad. Sci. USA*. 37: 729-740.

[3] Kendrew., J. C. Dickerson RE, Strandberg BE, Hart RG, and Davies D.R. (1960). Structure Of Myoglobin. Nature. 185: 422-427.

[4] Garnier, J., Osguthorpe, D. J. and Robson, B. (1978). Analysis of The Accuracy and Implications of Simple Methods For Predicting The Secondary Structure Of Globular Proteins. *J. Mol. Biol.* 120: 97-120.

[5] Garnier, J. and Robson, B. (1989). The GOR Method for Predicting Secondary Structures in Proteins. in: Fasman GD, ed. Prediction Of Protein Structure and The Principles Of Protein Conformation. New York: Plenum Press. 417-465.

[6] Garnier, J. Gibrat, J. and Robson, B. (1996). GOR Method For Predicting Protein Secondary Structure From Amino Acid Sequence. Meth. Enz. 266: 540-553.

[7] Frishman, D. and Argos, P. (1995). Knowledge-Based Protein Secondary Structure Assignment. Proteins: Structure, Function, and Genetics. 23:566-579.

[8] Rost., B. (2001). Review: Protein Secondary Structure Prediction Continues To Rise. J. Struct. Biol. 134: 204–218.

[9] Rost, B. (2003). Neural Networks Predict Protein Structure: Hype Or Hit?. Paolo Frasconi ed. in: Artificial Intelligence and Heuristic Models For Bioinformatics, CITY:ISO Press.

[10] Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. Prentice Hall, Upper Saddle River, NJ.

[11] Wang, Zhiyong, Zhao, Feng, Peng, Jian, and Xu, Jinbo. (2011). Protein 8class secondary structure prediction using conditional neural fields. Proteomics, 11(19):3786–3792, ISSN 1615-9861.

[12] Jian Zhou Olga G. Troyanskaya (2014). Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

[13] Granger C. W. (1989). Combining Forcasts: Twenty years laters. Journal of forcasting. 8:167-173.

[14] Subair, S O A (2012). Protein Secondary Structure Prediction: Using Artificial Neural Networks and Information Theory. Lambert Publishing, Germany. Publication Date: February 7, 2012.ISBN-10: 3847330667 | ISBN-13: 978-3847330660. http://www.amazon.com/Protein-Secondary-Structure-Prediction-Information/dp/3847330667

[15] Moult, J., Hubbard, T., Fidelis, K. and Pedersen, J. (1999). Critical Assessment of Methods of Protein Structure Prediction (CASP): Round II. Proteins: Structure, Function, and Genetics. Supplement. 3(1): 2-6.
[16] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. (2014). Critical assessment of methods of protein structure prediction (CASP) - round X. Proteins 82 (S2), 1-6. Published online: DOI 10.1002/prot.24452.
[17] Kryshtafovych A, Fidelis K, Moult J. (2014). CASP10 results compared to those of previous CASP experiments. Proteins 82 (S2), 164-174. Published online: DOI 10.1002/prot.24448.
[18] Rost, B. and Sander, C. (1993). Prediction Of Protein Secondary Structure At Better Than 70% Accuracy. J. Mol. Biol., 232., 584-599.
[19] Rost, B. R., Sander, C. and Schneider, R. "Redefining the Goals of Protein Secondary Structure Prediction". Journal of Molecular Biology. 235: 13-26. 1994.
[20] 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.
[21] http://mvirdb.llnl.gov/
[22] Matthews, B. B. (1975). Comparison Of The Predicted and Observed Secondary Structure Of T4 Phage Lysozyme. BiochimBiophysActa. 405: 442–451.
[23] Baldi, P., Brunak, S., Chauvin, Y., andersen, C. A. F. and Nielsen, H. (2000). Assessing The Accuracy Of Prediction Algorithms For Classification: An Overview. Bioinformatics. 16: 412-424
[24] Crooks, G. E., Jason, W. and Steven, E. B. (2004). Measurements Of Protein Sequence Structure Correlations. Proteins: Structure, Function, and Bioinformatics. 57:804–810.
[25] Subair, S (2014). A Data Oriented Approach to Assess the Accuracy of a Protein Secondary Structure Predictor. SUST Journal of Engineering and Computer Science (JECS), Vol. 16, No. 2, 2015. ISSN 1858-6783