BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Web server

Volume 11(8)

NNvPDB: Neural Network based Protein Secondary Structure Prediction with PDB Validation

Seethalakshmi Sakthivel, Habeeb S.K.M*

Department of Bioinformatics, School of Bioengineering, Faculty of Engineering & Technology, Kattankulathur Campus, SRM University, Potheri – 603203, Tamil Nadu, India; S.K.M Habeeb – Email: habeeb_skm@yahoo.co.in; *Corresponding author

Received July 07, 2015; Accepted July 26, 2015; Published August 31, 2015

Abstract:
The predicted secondary structural states are not cross validated by any of the existing servers. Hence, information on the level of accuracy for every sequence is not reported by the existing servers. This was overcome by NNvPDB, which not only reported greater Q_3 but also validates every prediction with the homologous PDB entries. NNvPDB is based on the concept of Neural Network, with a new and different approach of training the network every time with five PDB structures that are similar to query sequence. The average accuracy for helix is 76%, beta sheet is 71% and overall (helix, sheet and coil) is 66%.

Availability: http://bit.srmuniv.ac.in/cgi-bin/bit/cfpdb/nnsecstruct.pl

Keywords: protein secondary structure, neural network, automatic validation, online server

Background:
Protein secondary structure prediction plays a vital role and acts as an intermediate in solving tertiary structures; which provides an insight in to protein function **[1, 2].** Artificial Neural Networks (ANN) based prediction provides accurate results when compared to other methods **[3].** ANN is a simplified computational model that is capable of pattern recognition, feature extraction and image mapping. These are based on neural biology concepts, where signals are passed between individual nodes using weighted connection links; and the activation function used by each node determines the output **[4].**

Methodology:
Among many existing servers PSIPRED uses two feed forward neural network on PSI-BLAST output **[5].** YASPIN based on ANN and HMM uses SCOP1.65 to train and test the network and the datasets are built using PDB **[3].** Cfpred, available at (http://cib.cf.ocha.ac.jp/bitool/MIX/) uses a dataset of 106 protein sequences from PDB to train the network **[1].** These tools have reported higher overall accuracy however, the challenge remains in validating these predictions automatically and competently. In this study, ANN coupled with the homologous sequences were used to predict the secondary structures of proteins and automatically validate the predictions upon comparison to homologous PDBs. NNvPDB predicts three states Helix, Sheet and Coil; and reports percent accuracy by comparing it with similar PDB structures. This server would help scientists to extract validated results efficiently than the existing servers.

Neural Network Algorithm
The network contains 1 input layer, 1 hidden layer with 4 units and 1 output layer with 2 units. The input layer has 17 groups and each group has 21 units (17*21=357), making 357 units which are binary. A local coding scheme is used to prepare these binary inputs **[1].** The network uses bipolar sigmoid activation function **[6, 7]** to the net input as an activation function. Once the network is initiated, the state of each unit is defined by the formulae **[1]**.

Figure 1: A) Query page of NNvPDB; **B)** Result Page of NNvPDB showing the secondary structure states and Validation report of the predicted results; **C)** Comparative view of Predicted states with similar PDB Structures; **D)** a) Secondary structure of lowest reported Q_3 by NNvPDB(1IMX:A) b) IMX:A Validation file.

The main objective of the network is to map the given protein sequence with its corresponding secondary structure. Once mapped, the error rate is calculated using delta rule **[8],** and the weights of the network are adjusted using back propagation learning algorithm by gradient descent method to minimize the error. Each training data is iterated with back propagation learning until the error is minimized **[1].** Once the network is trained, the query/test set is used without back propagation algorithm.

Dataset Preparation

Training Set Preparation

The available tools on machine learning technique use a set of sequences for training the network. But here, we attempted a different approach of training the network with 5 sequences that are homologous to the query sequence, and the effort of combining neural network concept with homologous training set provides significant result in predicting the secondary structure of protein sequence. To implement this concept, we have written a Perl function to subject the query sequence to Blastp against the PDB database and five topmost hits are considered for further study. A sub sequence database is created using the selected hits. Each sequence in this database was converted to local coding scheme **[1]** and the experimentally determined secondary structures of these proteins were obtained from PDB secondary structure local database, and their secondary structure assignments were converted to binary codes as follows: Helix(1,0), Sheet(0,1) and coil (0,0).

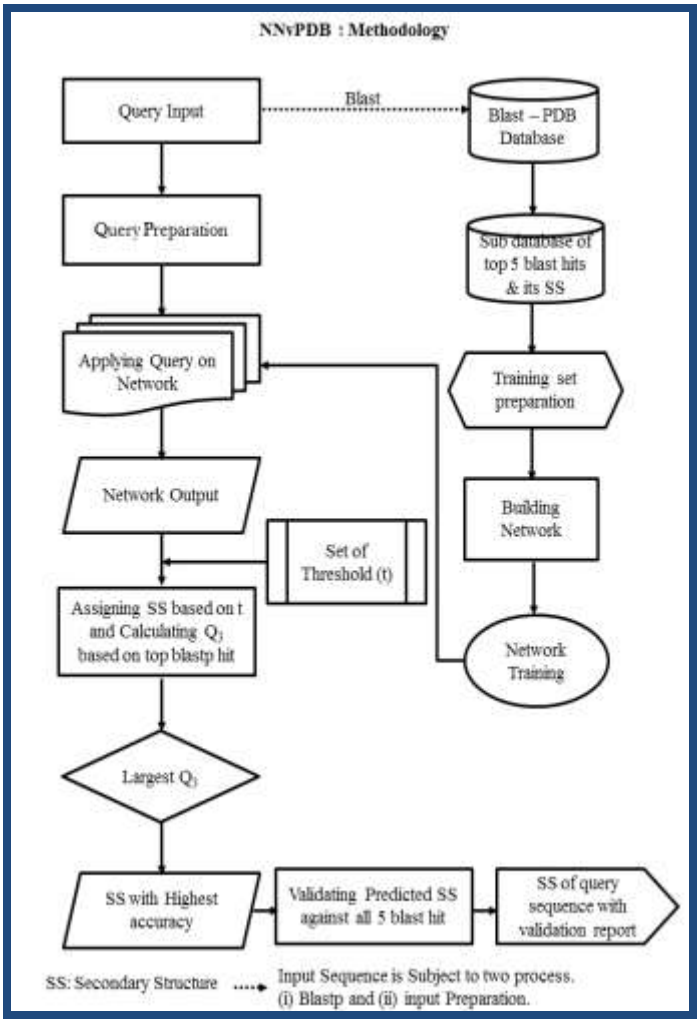

Figure 2: NNvPDB Methodology: The input sequence is subjected to Blastp against PDB database and a sub database is created with top 5 blast hits and its secondary structure. Each entry in the database is prepared and is used to train the network. Once the network is trained, the input sequence is prepared and tested against the trained network. The network outputs are converted to secondary structure states based on a set of threshold value. Structure assigned to each threshold value is validated against PDB and the structure with highest Q_3 is reported to user with validation information

Network Training:

For each user input sequence, the network is trained using supervised learning algorithm **[1]** for 100 iterations using the training set prepared using above said method. Initially the network was assigned with a random weight ranging between -0.5 to 0.5 and the learning rate of the network will get adjusted according to the percentage identity and query coverage of blast hits using perl function. Once the network is trained, it is used to predict the secondary structure of user queried sequence.

Test Set Preparation:

The protein sequence submitted by user act as test set and it is prepared **[1]** and tested against the trained network and the output is a floating point value, which are converted to binaries using threshold value. Here, we used a set of 74 sequences listed in Appendix 1. to validate the network.

Dependency of Q_3 on threshold value

The output of the neural network will be a floating point values (the network has 2 output units) which will be converted to binaries (0 or 1) based on some threshold value (t). If the network output is greater than t, it is set to 1 else it is set to 0 and from the binary values, secondary structures are assigned ([1,0]=helix, [0,1] = sheet, [0,0]=Coils). As the threshold value has higher influence over the assignment of secondary structure to the query sequence and in order to get best secondary structure, we used a set of threshold values ranging from -0.0001 to 0.9 for converting the floating point output from the network to binaries. For each threshold value, secondary structures are assigned and compared with the secondary structure of topmost blast hit and it's Q_3 **[9]** accuracy is calculated. Threshold value affording secondary structures with highest Q_3 is selected as final structure.

Automatic Validation

A local DSSP secondary structure database was created for all the proteins available in PDB. A Perl - MySQL function was written to select the secondary structures of the top 5 blast hits from this database. The predicted secondary structure was validated against the secondary structure of the top 5 Blast hits and its Q_3 accuracy was calculated **[9]** (**Figure 1B)** validation result). The overall working principle of this method is depicted in the flowchart **(Figure 2).**

Testing NNvPDB with Non- Homologous protein

The basic of NNvPDB methodology is training the network dynamically with similar protein sequences. In order to test the accuracy of NNvPDB for proteins with no homologous, we took a set of protein sequences listed in Appendix 2. Each protein in the list is treated as test sequences and for each sequences, we performed blastp against PDB and selected the hit which has less than 35 % sequence identity. The selected hit is used to train the network. Once the network is trained, the test sequences are applied against the network and the Q_3 accuracy was calculated.

Results & Discussion:

The query page of NNvPDB **(Figure 1A)** was designed in a user friendly manner which accepts single FASTA format protein sequence. The results page **(Figure 1B**) provides the predicted secondary structure of the query sequence, where the states of the secondary structures are represented as Helix(H), Sheets (B), Coil (C) and unassigned region(*). The validation table provides the percentage accuracy of the predicted result by comparing it with similar existing PDB structures. This table provides the following information: PDB ID, E-value, query coverage, percentage identity, subject start, subject end, query start, query end, length of the query sequence, number of residues predicted in all 3 states, number of residues correctly predicted in all 3 states and Q_3 accuracy percentage. Since all the protein sequence available in sequence database does not have a 100% identical similar PDB structures, so query start and query end provides information about the segment of the sequence that are matching with PDB structures which are used for validation. Validation file **(Figure 1C)** is provided as part of result page, which contains the details of blast search, query sequence, predicted secondary structure and secondary structures obtained from PDB. (*) is used to represent match and (.) represents mismatch.

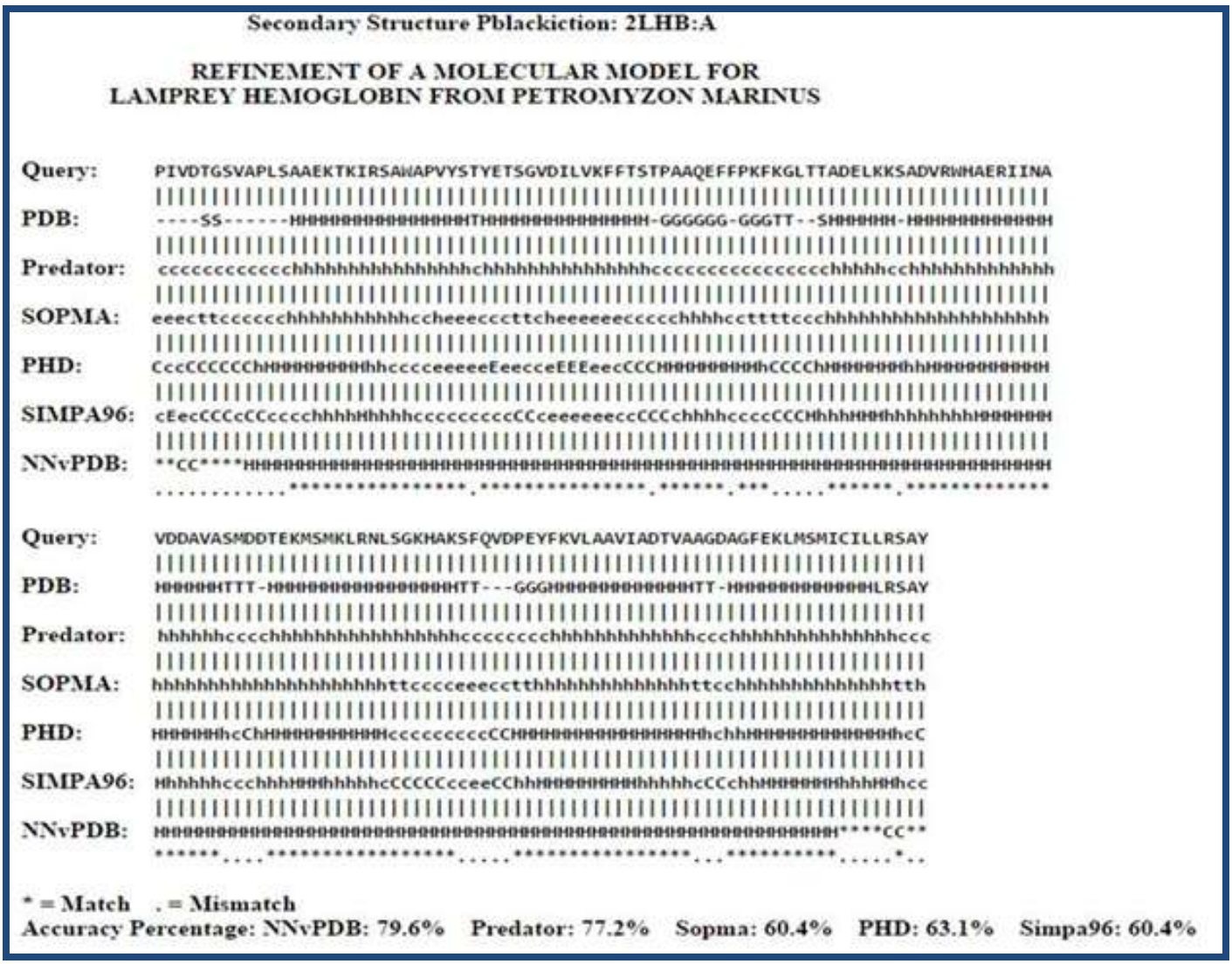

Figure 3: Secondary Structures predicted by Predator, SOPMA, PHD, SIMPA96 and NNvPDB for 2LHB:A and Secondary structures predicted by NNvPDB was compared with PDB. (*) match and (.) mismatch.

Three State Predictions

Accuracy of Helix prediction

The idea behind the NNvPDB was to overcome the shortcomings and to develop an application better than the existing ones in the field of secondary structure prediction. NNvPDB was compared with existing secondary structure prediction tools Predator (based on local pairwise alignments and amino acid interactions) **[10],** SOPMA (multiple sequence alignment & nearest neighbor method) **[11],** PHD (Neural Networks) **[12]** and SIMPA96 (homology based) **[13].** A total of 74 sequences were used for validation (Appendix 1). NNvPDB had the highest Q_3 for 28 sequences followed by Predator, PHD, and SIMPA96 & SOPMA with highest Q_3 for 27, 11, 8 & 1 sequence respectively. The relative accuracy at the residue level was also compared and the results are shown in **Table 1(see supplementary material)**. Off the 3604 helical residues reported by the PDB, 2722 were correctly predicted by NNvPDB at 75.5% accuracy. In the case of other tools, the accuracy percentage was 65.8%, 62.6%, 64.1% & 59.9% for Predator, SOPMA, PHD & SIMPA96 respectively. These results indicate the better performance of NNvPDB over other existing applications.

Accuracy of Sheet and Coil prediction

Similarly, the number of residues experimentally predicted to be sheets in PDB was 2949; 70.7% off these were correctly predicted by NNvPDB. The accuracy percentage reported by NNvPDB was higher than other tools and stood at 50.0%, 54.6%, 61.0% & 49.5% for Predator, SOPMA, PHD & SIMPA96 respectively. This result again shows the better performance of NNvPDB over other tools. However, the accuracy of coil prediction by NNvPDB was lowest among all the tools compared at 48.0% in comparison to 88.0% reported by Predator. One reason for this was the fact; NNvPDB uses threshold value as an important parameter to assign the secondary structure to the query sequence. This threshold value influences the binary codes representing the secondary structural states: Helix (1, 0), Sheet (0, 1) and coil (0, 0) as they are optimized based on the accuracy percentage of query sequence when compared with PDB.

Q_3 percentage accuracy

Table 1 (see supplementary material) shows the overall percent accuracy; highest and lowest Q_3 obtained for a single

individual sequence by the tools compared. The average percent accuracy of NNvPDB was 53.5% followed by 52.4% (Predator) and 48.8% (lowest) by SOPMA. From the pool of 74 sequences, the highest Q_3 observed for an individual sequence was 86.3% (1MBS:A) by Predator, SOPMA (80.0%), and NNvPDB (79.6%) (2LHB:A). **Figure 3** shows the results of the NNvPDB for the PDB entry 2LHB:A which was compared with other tools and validated with the experimental PDB results. The asterisk (*) represent the match and dot (.) represents the mismatch in predictions between NNvPDB and PDB. 109 accurate predictions were made by NNvPDB of 128 residual states. Impetus was also on to find the lowest Q_3 for the individual sequence; and it was witnessed Predator (19.6%) to have the lowest Q_3. In the case of NNvPDB, the minimum Q_3 recorded was 32.3%, well ahead than the SOMPA at 26.1%. Although, NNvPDB reported 32.3 % as lowest accuracy for 1IMX:A as shown in **Figure 1D(a),** it figured out secondary structure states of 21 out of 36 residues correctly as depicted in **Figure 1D(b)**. These results typify the promising performance of NNvPDB in secondary structure prediction when compared to existing tools.

Comparison of NNvPDB with PHD

PHD is one of the knowledge based method which generates profile using multiple sequence alignment and feed the generated profile as input to network. The network model of PHD has 3 levels **[12].** Since Neural Network act as a base for the development of NNvPDB and PHD, the secondary structures predicted by both the applications were compared. Table 2 brings out the number of residues correctly predicted by NNvPDB and PHD. The 74 experimentally determined proteins taken for validation found to comprise 9311 secondary structural states (Helix, Sheet and Coil). It was observed that NNvPDB predicted 65.6% (6132 residues) of the states correctly in comparison to PHD 64.2% (5982 residues) which differs from NNvPDB by 1.6% which marks the betterment of NNvPDB in the field of secondary structure prediction.

Sensitivity and Specificity

Sensitivity and Specificity are the terms to measure the veracity of the tool; and these were calculated to compare the performance of NNvPDB and PHD using the formulae **[14].** Sensitivity is the measure of proportion of actual positives which are correctly identified as positives and Specificity measures the proportion of negatives which are correctly identified as negatives. Appendix 3 lists out the sensitivity and specificity calculated for 68 sequences **[14]. Table 2 (see supplementary material)** summarizes the computed sensitivity and specificity between NNvPDB and PHD. The average sensitivity imparted by NNvPDB for helix was 53.6% and for sheet was 60.3% which was found to be higher than PHD which had an average sensitivity of 48.3% for helix and 51.5 % for sheet. But the average coil sensitivity by NNvPDB (42.7%) was lower than PHD (67.5%) as threshold optimization conquered the total coil prediction which results in increased specificity of Coil (77.5%) than by PHD (74.0%).The average specificity for helix by NNvPDB (61.6%) was lesser than PHD (83.2%) and specificity of sheet by NNvPDB was higher than PHD which was 86.7% and 82.3% respectively. When NNvPDB was compared with PHD with respect to sensitivity and specificity, NNvPDB tried to map secondary structure states to query sequence with higher accuracy.

Conclusion:

This study was intended to develop a secondary structure prediction server which not only predicts the residual states, but also validates the predictions with the structural homologs housed in the PDB database. NNvPDB promises to predict helix and sheet states with higher accuracy than Predator, SOPMA, PHD and SIMPA96 and also validates it simultaneously; a service offered exclusively by NNvPDB. The approach of training the network with robust dynamic homologous dataset ensures higher accuracy than the others. The overall Q_3 reported by NNvPDB was 53.5%. This would ensure the quality of the predicted structure before advanced studies could be taken up. NNvPDB would be a notable advancement in the field of secondary structure prediction with an attempt to validate the predicted result in an efficient and accurate manner than the existing servers.

Acknowledgment:

This study was supported by Department of Bioinformatics, School of Bioengineering, SRM University, India.

References:

[1] Qian N & Sejnowski T*J, J Mol Biol*. 1998 **202**: 865 [PMID: 3172241]
[2] Kloczkowski A *et al. Proteins* 2002 **49**: 154 [PMID: 12210997]
[3] Lin K *et al. Bioinformatics* 2004 **21**: 152 [PMID: 15377504]
[4] Shilpi Rani & Falguni Parekh, *IJIRSET*. 2014 **3**: 7
[5] McGuffin J *et al. Bioinformatics* 2000 **16**: 404 [PMID: 10869041]
[6] Karlik B & Olgac A, *IJAE*. 2011 **1**: 111
[7] Yonaba H *et al. J Hydrol Eng*. 2010 **15**: 275
[8] Huk M, *Int J Appl Math Comput Sci.* 2012 **22**
[9] Cuff JA & Barton GJ, *Proteins* 1999 34: 508 [PMID: 10081963]
[10] Frishman D & Argos P, *Protein Eng*. 1996 **9**: 133 [PMID: 9005434]
[11] Geourjon C & Deléage G, *Comput Appl Biosci.* 1995 **11**: 681 [PMID: 8808585]
[12] Rost B & Sander C, *J Mol Biol.* 1993 **232:** 584 [PMID: 8345525]
[13] Levin JM *et al. FEBS Lett*. 1986 **205:** 303 [PMID: 3743779]
[14] Do C *et al. Bioinformatics* 2006 **22**: e90 [PMID: 16873527]

Edited by P Kangueane

Citation: **Sakthivel & Habeeb**, Bioinformation 11(8): 416-421 (2015)

BIOINFORMATION

open access

Supplementary material:

Table 1: Accuracy Percentage Reported by tools

	PDB	NNvPDB	PREDATOR	SOPMA	PHD	SIMPA96
Residues reported in Helix State	3604	4984	3053	3611	3486	3200
Residues accurately predicted in Helix state#		2722	2371	2256	2310	2158
% Helix accuracy		**75.5[1]***	**65.8**	**62.6**	**64.1**	**59.9**
Residues reported in Sheet State	2949	2974	1886	2777	3256	2174
Residues accurately predicted in Sheet state#		2087	1475	1609	1798	1459
% Sheet accuracy		**70.7[1]***	**50.0**	**54.6**	**61.0**	**49.8**
Residues reported in Coil State	2758	3549	7211	5705	5363	6719
Residues accurately predicted in Coil state#		1323	2425	1985	1874	2247
% Coil accuracy		**48.0**	**88.0***	**72.0**	**68.0**	**81.5**
Average Q_3		**53.5[1]***	**52.4**	**48.8**	**49.85**	**49.2**
% Highest Q_3		79.6 (2LHB:A)	86.2* (1MBS:A)	80 (2ICB:A)	79.1 (1MBD:A)	74.7 (2ICB:A)
% Lowest Q_3		32.3* (1IMX:A)	19.6 (1CRN:A)	26.1 (1CRN:A)	21.7 (1CRN:A)	24.0 (1CRN:A)

Percentage accuracy of helix, sheet and coil with average Q_3, highest and lowest Q_3 obtained for single sequence. n* Highest value (n = rank) and * Highest value. Respective PDB Id of highest and lowest Q_3 was given in brackets. # True Positives with respect to PDB.

Table 2: Sensitivity and Specificity of NNvPDB and PHD

Parameters	Secondary Structures	NNvPDB	PHD
Average Sensitivity (64 Sequences)	Helix	53.6% *	48.3%
	Sheet	60.3%*	51.5%
	Coil	42.7%	67.4%*
Average Specificity (64 Sequences)	Helix	61.6%	83.2%*
	Sheet	86.7%*	82.7%
	Coil	77.5%*	74.0%
H+S+C Correctly Predicted (74 Sequences)		6132 (65.9%)	5982 (64.2%)

The calculations are based on 12232 total residues and 9311 (H+S+C) predictions in PDB for 74 protein sequences. * Highest Value.