[Loganathan *, 5(11): November 2018]
DOI- 10.5281/zenodo.1568848

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

A STUDY ON TWILIGHT ZONE PROTEINS CONSENSUS SECONDARY STRUCTURE PREDICTION

E.Loganathan[1,2], K.Dinakaran[3], S.Gnanendra[4] & P.Valarmathie[5]
[1]Department of M.C.A., Mahendra Engineering College, TamilNadu, India
[2]Research and Development center, Bharathiyar University, TamilNadu, India
[3]Department of CSE, PMR Engineering College, TamilNadu, India
[4]Department of Biotechnology, Mahendra Arts & Science College, TamilNadu, India
[5]Department of CSE, Savetha Engineering College, TamilNadu, India

ABSTRACT

Proteins are sequence of amino acids linked by means of peptide bonds to form a primary structure. The formation of Hbond within chain and between chain of these aminoacids tends to for secondary structures. The prediction of secondary structure plays a vitalrole while finding its similarity to determine the 3D structure of other proteins. Hence the problem addressed in this research is to evaluate the protein secondary structure prediction methods from NPS@ server, i.e. GOR, HNN, DPM, DSC, PHD, PREDATOR, SOPMA, STRIDE etc and to make a comparison among the methods using the same data sets. In this study we have proposed a consensus secondary structure prediction method, in which the four secondary structure prediction methods. PHD, PREDATOR, HNN and SOPMA were combined. We have observed anaverage *Q*3 prediction accuracy of 71.2% which is an improvement of 0.9%over PHD and Segment Overlap Accuracy (SOV) of 72.4%. In general, DSSP defines 8 states for secondary structure prediction, but these are reduced to 3 prediction by many researchers. Theses reduction in states ate proves to be less expensive in computational and has the advantage of not requiring the calibration of involved parameters.

Keywords: *Secondary structure, consensus methods, protein, 3D structures.*

I. INTRODUCTION

The growth in the genomic sequences and its accumulation in repositories has led to the exponential growth of predicted protein sequences while leaving a gap in determination of protein 3D structures.However, the gap between protein structure knowledge and its sequence is rapidly increasing. In this scenario, the understanding protein biology and itsstructure are very essential. Computationalstructure prediction methods have sought to meet the challenge of bridging the sequence-structure gap inproviding expensive information for the determination of protein structures [1,2].

The knowledge on protein-folding is a long-term goal in which the researchers can understand the depth of three-dimensional structure of a protein that are derived based on its aminoacids. Secondary structure prediction is often regarded as the initial starting point in predicting the three-dimensional structure of a protein [3]. Fundamentally, it attempts to classify amino acids in protein sequence according to their predicted local structure, which can be subdivided into three states: a-helix, b-sheets, or loops. While the number of states may vary depending on the algorithm employed, we will simplify our analysis to a three-state problem, Q3, where turns, coils, or other helices will collectively be called “loops”.

The fundamental assumption of all secondary structure prediction methods are based on theassociationof amino acid sequence and its secondary structure [4]. Because the entire information for forming secondary structure is contained in the primary sequence, any short stretch of amino acid sequence will preferentially adopt one kind of secondary structure over another. Thus, many algorithms that checks 13-17 residues window andassume the central amino acid will adopt a confirmation determined by the side groups of all the amino acids in the window. For a-helices, this window is typically 5-40 residues long, and for b-sheets, this window ranges from 5-10 residues in length. While earlier algorithms assumed that each amino acid within the sequence window was unaffected by other neighboring amino acids, later methods recognized the oversimplification, and accounted for the possibility that more distant interactions within the primary amino acid chain may influence local secondary structure [5].

Secondary structure prediction algorithms:

The three widely used secondary structure prediction methods:

- Chou-Fasman and GOR methods
- Neural networks
- Nearest-neighbor

Chou-Fasman and GOR methods:

Programs that uses this algorithms : DPM, DSC, GOR IV

In 1974, Chou and Fasman [3] developed a statistical method based on the amino acids propensities.They proposed that adopt secondary structures are based on the observation of their location in 15 protein structures determined by X-ray diffraction. These statistics derive from the particular stereochemical and physicochemical properties of the amino acids.Over the years, these statistics have been refined using a larger set of proteins. Unlike Chou-Fasman, GOR (Garnier, Osguthorpe, and Robson)[6]revealedthat the flanking aminoacids determine the secondary structure of the central amino acid residue. In GOR IV, the amino acids pairwise combinations of flanking region and the central residue can are considered for the conformation of the central amino acid. For instance, a particular amino acid that is surrounded with aminoacids with helix propensity, then that amino acid would be likely to be in a helix, even if its propensity to helix is low..

Neural network models

Programs that uses this algorithms: PHD, PSIPRED, NNPREDICT

A neural network is comprised of a machine learning approach, providing computational processes the ability to "learn" in an attempt to simulate the complex patterns of synaptic connections formed among neurons in the brain during learning. Computers are trained to recognize patterns in known secondary structures using training sets of non-homologous structures, and tested with proteins of known structure [7]. Neural networks have been able to achieve a level of 73% overall three-state per-residue accuracy. The reasons for improved prediction accuracy is attributed to its ability to align the query sequence with other related proteins of the same family and find protein members with known structures to aid its assignment of secondary structures. While neural networks can detect interactions between amino acids within a window of amino acids.

Nearest-neighbor methods

*Programs that uses this algorithms:*SOPM, SOPMA, NNSSP, PREDATOR

In this approach, the secondary structure of the central residue is performed by finding some number of the closest sequences (from a database of proteins with known structure) to a subsequence defined by a window around the amino acid of interest. Using the known secondary structures of the aligned sequences (generally weighted by their similarity to the target sequence) a secondary structure prediction is made. For instance, a large list of short sequence fragments is made by sliding a window of defined sequence length along a set of ~400 training sequences of known structure that are non-homologous to each other, and recording the secondary structure of the central amino acid of each window [8-10]. Subsequently, a window of the same size is then selected from the query sequence and compared to the list of short sequence fragments to identify the 50 best matches. The frequency that the central amino acid in each of the 50 matching fragments will form a particular secondary structure is then used to predict the secondary structure of the central amino acid in the query sequence. The variability in nearest neighbor methods arises from the selection of subsequences closest to a window around the amino acid whose structure isbeing predicted. Each program uses a different set of parameters, like how similarity is defined, or what sequence window size should be examined.

II. METHODS

Seven secondary structure prediction methods were run on the protein primary sequences and each method is briefly described:

DPM (Double Prediction Method)
The DPM (Double Prediction Method) algorithm uses two approaches to produce the final result by predicting the protein structural class followedby the secondary structure for the sequence. It takes 4 steps, first, from AA composition the structural class is predicted, secondly, estimation of preliminary secondary structure, two independent predictions comparison and parameters optimization.

DSC (Discrimination of protein Secondary structure Class)
Discrimination of protein Secondary structure Class (DSC)method uses linear statistical methods for prediction. In this comprehensible prediction method, the relative information from different sources isused to measure. The prediction accuracy of 70.1% on a standard set of 126 proteinswas recorded for DSC [11]. This was not significantly different from PHD, a popular prediction method. For medium length sequences DSC was more accurate than PHD, and combining DSC and PHD produced a prediction method more accurate than either.

GOR (*G*arnier, *O*sguthorpe, and *R*obson)
Theinformation theory based method is named after*G*arnier, *O*sguthorpe, and *R*obson. The prediction algorithm of this method considers probability of every amino acid having a particular secondary structure and its conditional probability by assuming its neighbors with the same structure. This method is more sensitive and accurate method as the the structural propensities of amino acids are strong for only proline and glycine. This GOR method more successful in predicting alpha helices than beta sheets, which it frequently mispredicts as loops or disorganized regions

HNN (Hierarchical Neural Network)
The Hierarchical Neural Network (HNN) prediction method is seen as an improvement on the famous classifier [12], and derived from the system NETtalk (Guermeur). As its predecessor, it is made up of two networks: a sequence-to-structure network and a structure-to-structure network. The prediction is thus only based on local information. The improvements mainly deal with two points:

- Technical tricks (recurrent connections, shared weights etc.) have been used to increase the context on which the prediction is made and concomitantly decrease by two orders of magnitude the number of parameters (weights).
- Physico-chemical data have been explicitly incorporated in the predictors used by the structure-to-structure network.

These modifications have significantly improved the error in generalization.

PHD
PHD are neural network systems (a sequence-to-structure level and a structure-structure level) to predict secondary structure (PHDsec), relative solvent accessibility (PHDacc) and transmembrane helices (PHDhtm) [13]. The NPS@ server only uses PHDsec. PHDsec focuses on predicting hydrogen bonds. The procedure essentially involves executing a BLASTP search of your sequence, filtering these results and aligning them with CLUSTALW, then using the multiple alignment as the input of the neural network. The PHD prediction done with NPS@ is better than the PHD prediction on the single sequence. But it's not exactly the same and could be a little bit less accurate than the PredictProtein one.

PREDATOR
This method is based on detection of hydrogen-bonded residues in single amino acid sequence. This method predicts by taking a single protein sequence as input and can optimally use a set of unaligned sequences to predict the query

sequence. PREDATOR relies on careful pairwise local alignments of the sequences in the set with the query sequence to be predicted [14].

SOPMA(Self-Optimized Prediction Method with Alignment)
Self-Optimized Prediction Method with Alignment (SOPMA) is based on the homolog method. The improvement takes place in the fact that SOPMA takes into account information from an alignment of sequences belonging to the same family [15].

Consensus prediction method
For consensus method, the Q3 accuracy of DPM, DSP and GOR methods were observed as lower than the other methods. Thus, a consensus was calculated from PHD, PREDATOR,HNN, and SOPMA. According to the NPS@ web server's consensus prediction algorithm, the standard consensus was calculated from the predictions of each method by taking the most popular state. (for example if aamino acid residue was predited to be helices by HNN,PHD, PREDATOR, and strand by SOPMA, then the consensus prediction will be Helix. However, if no consensus for a particular residue was predicted, then the PHD method prediction will be assigned [16].

In this scenario, we have investigated the various combinations of the prediction methods, in an attempt to raise the average Q3.All possible combinations of methods were tried to calculate the consensus, but no combination of methods improved upon the average Q3 of the consensus of HNN, PHD, PREDATOR and SOPMA.

III. ACCURACY CALCULATION

The accuracy of the predictions are calculated by two methods such as average Q3 and Segment Overlap.

Q3, measure the overall percentage of predicted residues,

$$Q_3 = \sum_{(i=H,E,C)} \frac{predicted_i}{observed_i} \times 100.$$

Segment overlap values capture the segment prediction, and vary from an ignorance level of 37% (random protein pairs) to an average 90% level for homologous protein pairs. Segment overlap is calculated by:

$$Sov = \frac{1}{N}\sum_{s} \frac{minov(s_{obs}; s_{pred}) + \delta}{maxov(s_{obs}; s_{pred})} \times len(s_1).$$

Where *N* is the total number of residues, ***minov*** is the actual overlap, with ***maxov*** is the extent of the segment. ***δ*** is the accepted variation which assures a ratio of 1.0 where there are only minor deviations at the ends of segments.

IV. RESULTS AND DISCUSSION

In this study, we have combined four prediction methods such as HNN, SOPMAPHD and PREDATOR as a simple majority-wins method and the average Q3 and SOV for three-state was predicted and given in Table 1.

Fig : 1 Predicted secondary consensus method

```
                   10        20        30        40        50        60
70
                    |         |         |         |         |         |
|
UNK_116760
RTDCYGNVNRIDTTGASCKTAKPEGLSYCGVSASKKIAERDLQAMDRYKTIIKKVGEKLCVEPAVIAGII
HNNC
cccccccccccccccccccccccccceeeehccchhhhhhhhhhhhhhhhhhhhhccceeecchhhhhhh
PHD
cccccceeeeeeccccccccccccccccccchhhhhhhhhhhhhhhhhhhhhhhhhccccccccceeeee
Predator
ccccccccccccccccccccccccccccchhchhhhhhhhhhhccccchhhhhhhhhhhccchhhhhhhh
SOPMA
ccctttceeeeecccccccccccccccetccchhhhhhhhhhhhhhhhhhhhhhhhhhttcchhhhhhhhh
Sec.Cons.
ccccccc?????ccccccccccccccc?cc?chhhhhhhhhhhhhhhhhhhhhhhh???ccc?hhhhhhh

                   80        90       100       110       120       130
140
                    |         |         |         |         |         |
|
UNK_116760
RESHAGKVLKNGWGDRGNGFGLMQVDKRSHKPQGTWNGEVHITQGTTILINFIKTIQKKFPSWTKDQQLK
HNNC
hhccccceeeccccccccccceeeeeccccccccccccceeeeecchhhhhhhhhhhhhhccccccchccc
PHD
ecccccccccccccccccceeeeeeeccccccccccccchhhhhhhhhhhhhhhhhhhhccccchhhhhhc
Predator
hhcccccccccccccccccccccccccccccccccccchhhhhhhhhhhhhhhhhhhhhccccchhhhhh
SOPMA
hcctttteeeccccccccccceeeeeeccccccccccccchhhhhhhhhhhhhhhhhhhtcccccchhhhhh
Sec.Cons.
h?ccccc??cccccccccc?eeeeecccccccccccccchhhhhhhhhhhhhhhhhhhh?ccccchhhhh?

                  150       160       170       180
                    |         |         |         |
UNK_116760 GGISAYNAGAGNVRSYARMDIGTTHDDYANDVVARAQYYKQHGY
HNNC       ccceecccccccccehheeeccccchhhhhhhhhhhhhhhhhccc
PHD        ccceeeecccceeeeeecccccccccchhhhhhhhhhhhhhhccc
Predator   hhhhhhhhccccccccccccccccccchhhhhhhhhhhhhcccc
SOPMA      hheeeeetttccceeeeeeecccccccccchhhhhhhhhhhhtttc
Sec.Cons.  ??ceeeeccccc?eee??cccccccc?hhhhhhhhhhhhhhccc

Sequence length :   184
```

Fig 2: predictions in different methods and consensus prediction

```
HNNC :
   Alpha helix      (Hh) :    65 is  35.33%
3_10 helix       (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand (Ee) :    26 is  14.13%
   Beta turn       (Tt) :     0 is   0.00%
   Bend region     (Ss) :     0 is   0.00%
   Random coil     (Cc) :    93 is  50.54%
Ambigous states (?)   :     0 is   0.00%
   Other states         :     0 is   0.00%


PHD :
   Alpha helix     (Hh) :    66 is  35.87%
3_10  helix        (Gg) :     0 is   0.00%
   Pi helix        (Ii) :     0 is   0.00%
   Beta bridge     (Bb) :     0 is   0.00%
   Extended strand (Ee) :    29 is  15.76%
   Beta turn       (Tt) :     0 is   0.00%
   Bend region     (Ss) :     0 is   0.00%
   Random coil     (Cc) :    89 is  48.37%
Ambigous states (?)   :     0 is   0.00%
   Other states         :     0 is   0.00%

Predator :
   Alpha helix     (Hh) :    82 is  44.57%
3_10  helix        (Gg) :     0 is   0.00%
   Pi helix        (Ii) :     0 is   0.00%
   Beta bridge     (Bb) :     0 is   0.00%
   Extended strand (Ee) :     0 is   0.00%
   Beta turn       (Tt) :     0 is   0.00%
   Bend region     (Ss) :     0 is   0.00%
   Random coil     (Cc) :   102 is  55.43%
Ambigous states (?)   :     0 is   0.00%
   Other states         :     0 is   0.00%


SOPMA :
   Alpha helix     (Hh) :    74 is  40.22%
3_10  helix        (Gg) :     0 is   0.00%
   Pi helix        (Ii) :     0 is   0.00%
   Beta bridge     (Bb) :     0 is   0.00%
   Extended strand (Ee) :    26 is  14.13%
   Beta turn       (Tt) :    15 is   8.15%
   Bend region     (Ss) :     0 is   0.00%
   Random coil     (Cc) :    69 is  37.50%
Ambigous states (?)   :     0 is   0.00%
   Other states         :     0 is   0.00%

Sec.Cons. :
   Alpha helix     (Hh) :    71 is  38.59%
3_10  helix        (Gg) :     0 is   0.00%
   Pi helix        (Ii) :     0 is   0.00%
   Beta bridge     (Bb) :     0 is   0.00%
   Extended strand (Ee) :    12 is   6.52%
   Beta turn       (Tt) :     0 is   0.00%
   Bend region     (Ss) :     0 is   0.00%
   Random coil      (Cc) :    78 is  42.39%
Ambigous states (?)   :    23 is  12.50%
   Other states          :     0 is   0.00%
```

Table 1: Difference between Q3 and SOV accuracies for each method

Sl.No.	Method	Accuracy	
		Q3	SOV
1.	PHD	71.8	71.3
2.	HNN	70.2	68.4
3.	PREDATOR	66.8	68.7
4.	SOPMA	69.2	69.9
5.	CONSENSUS	72.8	71.6

With the incorporation of evolutionary information, the prediction accuracy have been accomplished and with the combination of secondary structure prediction methods. The development of consensus methods has led to the increase in the prediction accuracy. While forming the consensus sequence per amino acid, the three secondary states such as a-helix (H), b-strand (E) and other/ loop (L) are distinguished. We have compared the predicted consensus sequence with the true three-state sequence derived from the DSSP secondary structure assignment of known 3D structures to determine the prediction accuracy. Each of the three possible states H, E and L are resulted from the collapse transformation of the eight DSSP states such as {G, H, I} a-helix (H), {B, E} b-strand (E), {S, T, `.'} other (L).

Even though there are different standards for reducing DSSP 8-state (H,C,B,E,T,S,G,I) assignments to 3 states (H,C,E), the changein the reduction method can apparently alter the prediction accuracy. However, we have not trained the methods by using different 8 to 3 state reductions, the testing of all methods with different reduction methods have resulted in the higher accuracy of the proposed consensus prediction method.

The new combination of PHD, PREDATOR, HNN and SPOMA presented here shows an improvement of 72.8%.The proposed simple consensus approach that is based on the majority of four prediction methods can be superior when compared to each of the sevensingle methods. Also these method has shown better results compared to complex combinations of more than three single prediction methods that are employed in Jpred[18]. However, this method is yet is to be proven to work on large benchmark data sets with different combinations.

V. CONCLUSION

In this study, we have proposed a secondary structure prediction method, which combines the four secondary structure prediction methods such as PHD, PREDATOR, HNN and SOPMA as a simple majority wins method. The predicted results of four methods were taken for the consensus secondary structure prediction. We claim that, this method is mainly succeeded with the combination four best single methods and the noise-filtering properties of a consensus approach that helped to ignore the training errors of every single method.

REFERENCES

1. **Benner, S. A. (1989).** *Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. Adv. Enzyme Regul.,**28**, 219-236.*
2. **Blout, E. R. (1962).** *The dependence of the conformation of polypeptides and proteins upon amino acid composition. In Polyamino Acids, Polypeptides, and Proteins (Stahman, M., eds.), pp. 275-279, Univ. of Wisconsin Press, Madison.*
3. **Chou, P. Y. &Fasman, U. D. (1974).** *Prediction of protein conformation. Biochem.,**13**, 211-215.*
4. **Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998).***JPred: a consensus secondary structure prediction server. Bioinformatics,**14**, 892-893.*

5. **Defay, T. & Cohen, F. E. (1995).** *Evaluation of current techniques for ab initio protein structure prediction. Proteins,**23**, 431-445.*

6. **Garnier, J., Gibrat, J.-F. & Robson, B. (1996).** *GOR method for predicting protein secondary structure from amino acid sequence. Meth. Enzymol.,**266**, 540-553.*

7. **Garnier, J., Osguthorpe, D. J. & 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.*

8. **Gibrat, J.-F., Garnier, J. & Robson, B. (1987).** *Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J. Mol. Biol.,**198**, 425-443.*

9. **Kanehisa, M. (1988).** *A multivariate analysis method for discriminating protein secondary structural segments. Prot. Engin.,**2**, 87-92.*

10. **Levin, J. M., Pascarella, S., Argos, P. &G arnier, J. (1993).** *Quantification of secondary structure prediction improvement using multiple alignment. Prot. Engin.,**6**, 849-854.*

11. **Maclin, R. & Shavlik, J. W. (1993).** *Using knowledge-based neural networks to improve algorithms: refining the Chou-Fasman algorithm for protein folding. Machine Learning,**11**, 195-215.*

12. **Pauling, L. & Corey, R. B. (1951).** *Configurations of Polypeptide Chains with Favored Orientations Around Single Bonds: Two New Pleated Sheets. Proc. Natl. Acad. Sci. U.S.A.,**37**, 729-740.*

13. **[13]Petersen, T. N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J. et al. (2000).** *Prediction of protein secondary structure at 80% accuracy. Proteins,**41**, 17-20.*

14. **Rost, B. (1999).** *Twilight zone of protein sequence alignments. Prot. Engin.,**12**, 85-94.*

15. **Schulz GE, Schirmer RH.(1979)** *Principles of Proteins Structure. New York: Springer-Verlag, p 1–314.*

16. **Stolorz, P., Lapedes, A. & Xia, Y. (1992).** *Predicting protein secondary structure using neural net and statistical methods. J. Mol. Biol.,**225**, 363-377.*

17. **Szent-Györgyi, A. G. & Cohen, C. (1957).** *Role of proline in polypeptide chain configuration of proteins. Science,**126**, 697.*

18. **Zhou, X., Alber, F., Folkers, G., Gonnet, G. H. & Chelvanayagam, G. (2000).** *An analysis of the helix-to-strand transition between peptides with identical sequence. Proteins,**41**, 248-256.*