Indian Journal of Advanced Zoology (IJAZ)
ISSN: 2582-9521 (Online), Volume-1 Issue-1, April 2021

Protein Secondary Structure Prediction using Bioinformatics Tools for Hemoglobin

Roma Chandra, PratibhaTiwari

Abstract: ***Protein structure prediction is one of the important goals in the area of bioinformatics and biotechnology. Prediction methods include structure prediction of both secondary and tertiary structures of protein. Protein secondary structure prediction infers knowledge related to presence of helixes, sheets and coils in a polypeptide chain whereas protein tertiary structure prediction infers knowledge related to three dimensional structures of proteins.Protein secondary structures represent the possible motifs or regular expressions represented as patterns that are predicted from primary protein sequence in the form of alpha helix, betastrand and coils. The secondary structure prediction is useful as it infers information related to the structure and function of unknown protein sequence. There are various secondary structure prediction methods used to predict about helixes, sheets and coils. Based on these methods there are various prediction tools under study. This study includes prediction of hemoglobin using various tools. The results produced inferred knowledge with reference to percentage of amino acids participating to produce helices, sheets and coils. PHD and DSC produced the best of the results out of all the tools used.***

Keywords: ***There are various secondary structure prediction methods used to predict about helixes, sheets and coils.***

I. INTRODUCTION

Protein is a biomolecule which is an important dietary source which we as humans consume. Protein is a polypeptide made up of amino acids and is present in its three dimensional arrangement in nature. Protein is studied in four different levels as primary, secondary, tertiary and quaternary which are mentioned as following:

Primary structure: It is the first level of protein structure which contains amino acid sequence in the form of polypeptide chain. During protein biosynthesis amino acids are bound together with peptide bonds. Based on the nature of free groups at the extremities of the sequence the protein hastwo ends: carboxyl terminal (C-terminus) end and the amino terminal (N- terminus) end. Primary structure of any protein is determined from the gene from which it is translated. As we know as per central dogma DNA transcribes to produce mRNA which further translates to produce protein. Primary structure of any protein can be studied from protein databases.The annotated information for any protein sequence can be retrieved from various protein databases like UNIPROT, PDB, etc. Primary sequences of proteins can be extracted from these databases in fasta format and can be used for secondary as well as tertiary structure prediction.

Manuscript received on 07 April 2021 | Revised Manuscript received on 03 April 2021 | Manuscript Accepted on 15 April 2021 | Manuscript published on 30 April 2021.

***** Correspondence Author

Roma Chandra, Department of Biotechnology,IILM College of Engineering & Technology, Greater Noida roma.chandra@iilmcet.ac.in

PratibhaTiwari, Department of Biotechnology,IILM College of Engineering & Technology, Greater Noida

Secondary structure: It is the second level ofprotein structure that is represented in the form of alpha helixes, beta sheets and turns or coils. Basically the backbone of any protein consists of structures that are in the form of helixes or sheets which are connected by the help of turns or coils. Thus, connections producing structures such as helix-helix, sheet-sheet and helix-sheet are seen in the backbone of protein structure. Secondary structure prediction of proteins provides information regarding presence of helix ,sheet and coil that is which amino acid participates in specific type of secondary structure represented as H ,S, C (helix, sheet and coil). Secondary structure prediction of any protein can be done using any secondary structure prediction methods like Chou Fasman, GOR, Artificial Neural Network, etc.

Tertiary structure :Itis the third level of protein structure which represents the three dimensional conformation of protein that includes arrangement of amino acids into helixes, sheets and turns with backbone structure arrangement based on psi, phi and omega angles. Ramachandran plot explains the possible allowed and disallowed regions for the amino acids that further participates to form the three dimensional structure of protein. The polypeptide chain is a folded structure produced due to interactions between the R groups of participating amino acids. The possible interactions seen in tertiary structure includes hydrogen bonds, hydrophobic interactions, Wander wall interaction, disulphide bonds etc. Tertiary structure databases such as PDB provides annotated information regarding three dimensional structures of proteins. X-ray crystallography and NMR spectroscopy are the techniques used to predict the three dimensional of proteins. There are in silico methods that can also predict the three dimensional protein structures. The methods includes Ab-initio method, Homology modeling and Threading

Quaternary structure: It is the fourth level of protein structure which represents multiple polypeptide chains connected to produce a single protein structure. Basically quaternary level represents number of polypeptide chains that are connected to each other like for example hemoglobin is a protein that is made up of four subunits with two alpha and two beta types.

II. PROTEIN SECONDARY STRUCTURE PREDICTION

Secondary structure prediction is an important method used in the field of bioinformatics. Its main motive is to predict secondary structures of proteins based on their amino acid sequences. It provides the complete information of the amino acid sequence like alpha helices, beta strands or turns along with their parameters.The prediction process to search for helices, sheets and coils includes the following six methods:-

Chou FasmanMethod**:**ChouFasman algorithm is extensively used to predict secondary structure of proteins. This method is based on an algorithm that calculates prediction values of each participant amino acid. Conformational parameter for each amino acid is calculated on the basis of specific position frequency of every amino acid present in given polypeptide chain. The conformational parameters are calculated for the 20 amino acids based on information collected from standard proteins and are represented as P(α) P (β) and P(turn) for helixes, sheets and coils. The algorithm includes various steps initialized by assigning relevant parameters to all the amino acid residues of the protein for which prediction needs to be done.In further steps combination of six residues isidentified for helixes, five residues are sheets and four residues for turns. This method shows 50- 60% accuracy for secondary structure prediction.

Nearest neighbor method:Nearestneighbor method is also known as homologous method, memory based method and exemplar based method as it is based on a hypothesis that small length homologous sequences of polypeptide chain will represent similar secondary structures. This method uses structural databases for standard protein information. In this method small fragments are collected to prepare a sliding window. For every window the central amino acid residue is predicted for its secondary structure based on the rest of the residues from the training dataset. The same process is followed for prediction of other residues in the protein to be predicted.

HMM (Hidden Markov model):Hiddenmarkov model is another method used for prediction of protein sequences based on markov model. The output producing probabilities to produce helix, sheet and coil are used while predicting the secondary structure of protein needed.

GOR (Garnier-Osguthorpe-Robson): GOR is another secondary structure prediction method that is based on information theory. It can also predict the helix,β sheets, turn or random coils. The method is better for helix as compared to sheets because sheet depends on interactions with long range between two non-adjacent amino acid residues. In this method sliding window of 17 amino acid residues is used to predict the secondary structure of central residue for the polypeptide chain classifying amino acids into helixes, sheets and coils.The method shows 64% accuracy as being sheet, helix or coil.

Artificial Neural Network:ANN is based on biological neural network and is used to predict secondary structure of proteins based on standard protein training datasets. ANN uses classification method to categorize amino acid residues into helixes sheets and coils. Information is given as primary protein sequence to the ANN tool which is predicted for the presence of helix, sheet and coil based on weight training and updation of output produced to predict the secondary structure of proteins. The method shows 63% accuracy as being sheet, helix or coil.

Self-optimized prediction method (SOPMA): SOPMA is a secondary structure prediction method based on predicting helixes, sheets and coils on multiple alignments using self optimizationmethod .The method shows 63% accuracy as being sheet, helix or coil.

III. PROTEIN SECONDARY STRUCTURE PREDICTION TOOLS

There are various tools based on secondary structure prediction method that includes the following:

- *AGADIR (http://agadir.crg.es/)*
- *APSSP (http://crdd.osdd.net/raghava/apssp/)*
- CFSSP*(http://www.biogem.org/tool/chou-fasman/)*
- GOR *(https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html)*
- HHPRED*(http://toolkit.tuebingen.mpg.de/hhpred)*
- JPRED *(http://www.compbio.dundee.ac.uk/www-jpred/)*
- PROF *(https://www.aber.ac.uk/~phiwww/prof/)*
- PSIPRED *(http://bioinf.cs.ucl.ac.uk/psipred/)*
- SOPMA *(https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html)*
- STRAP *(http://www.bioinformatics.org/strap/Scripting.html)*
- TOPMATCH *(https://bio.tools/topmatch)*
- SPIDER2*(http://sparks-lab.org/yueyang/server/SPIDER2/)*
- SYMPRED *(http://www.ibi.vu.nl/programs/sympredwww/)*
- YASSPP*(http://glaros.dtc.umn.edu/yasspp/)*
- PSSPRED *(http://bioinf.cs.ucl.ac.uk/psipred/)*
- FRAG1D*(http://frag1d.bioshu.se/)*
- SPIDER2*(http://sparks-lab.org/yueyang/server/SPIDER2/)*
- RAPTORX-SS8*(http://raptorx.uchicago.edu/)*

IV. MATERIALS AND METHODOLOGY

The main aim of this research work is the comparative analysis of various secondary structure prediction tools. Primary sequence for hemoglobin used to study the secondary structure results. For this prediction analysis Hemoglobinsubunit gamma-2 protein sequence was retrieved from UNIPROT database in FASTA file format*(https://www.uniprot.org/uniprot/P69892.fasta)*. Thetools used for comparative analysis includes CFSSP, GOR, PHD, SOPMA, DSC, MLRC.

V. RESULTS

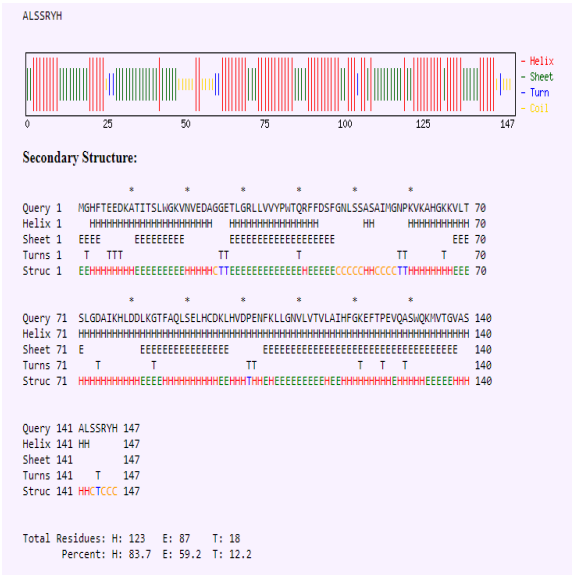
Secondary structure was predicted using hemoglobin sequence taken from the UNIPROT database. Prediction tools produced results that are represented in percentage form for the percentage of amino acids converted into helices, sheets and coils. Comparative analysis of all the following tools revealed that amino acids participating in hemoglobin tend to produce helices greater then sheets or coils.

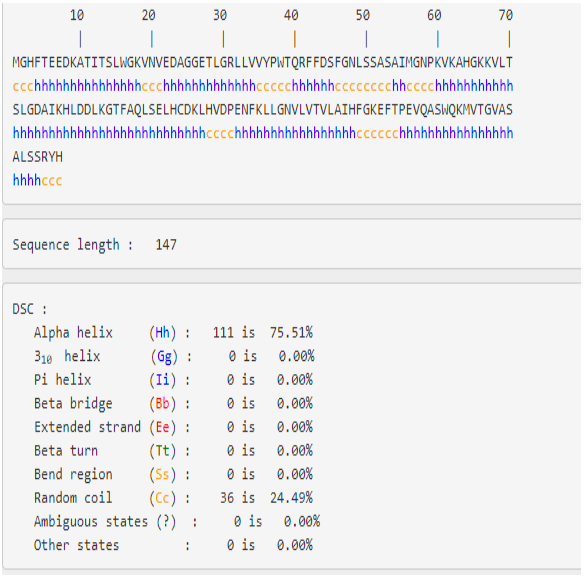
Published By:
Lattice Science Publication (LSP)

Comparative analysis of the results also reveals that these tools tend to produce variant results out of which PHD and DSC tools were close enough to the actual protein secondary structure studied from protein data bank.

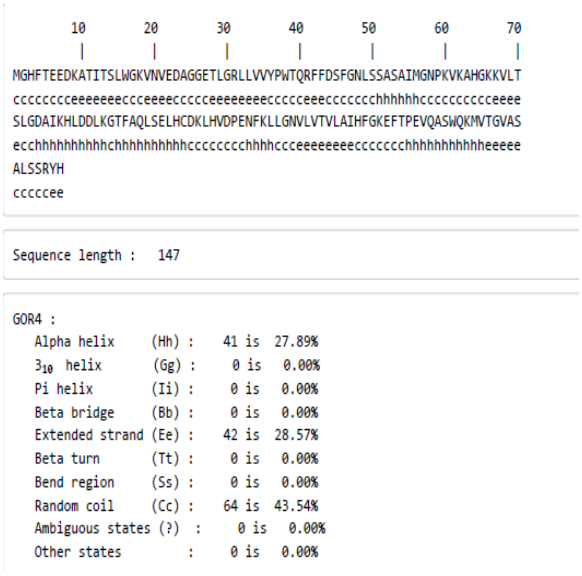
The results of the six tools under study are mentioned as following:

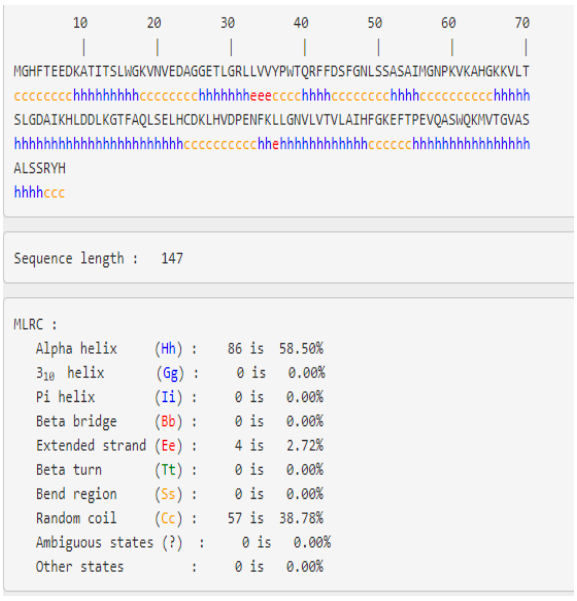
Name	GOR4	PHD	SOPMA	DSC	CFSSP	MLRC
Helix	27.89%	75.51%	66.67%	75.51%	83.7%	58.50%
Sheets	28.57%	2.04%	7.48%	0%	59.2%	2.72%
Coils	43.54%	22.45%	19.73%	24.49%	12.2%	38.78%



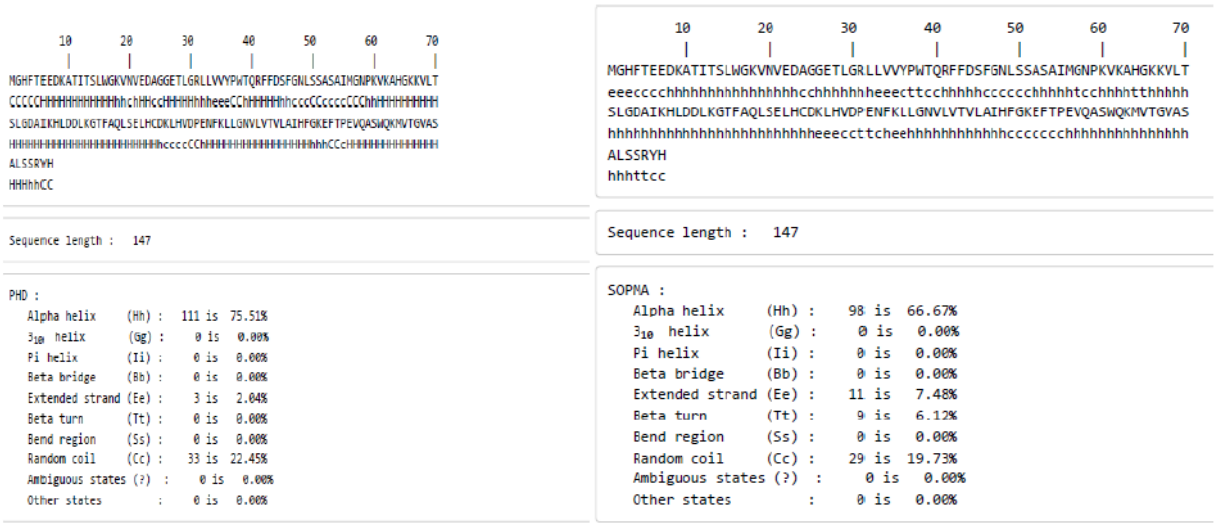


CFSSP **DSC**





GOR4 MLRC



PHD **SOPMA**

VI. CONCLUSION

The results produced using CFSSP, GOR, PHD, SOPMA, DSC, MLRC tools were different and variation in results was seen with respect to helices, strands and coils. Out of all these tools PHD and DSC had predicted helixes to be approximately 75%. Literature studies reveal that almost 75% of all the amino acids are participating in the formation of helical structures in hemoglobin. Thus, we conclude had predicted the best of the results. The same will be verified after predicting the three dimensional structure of hemoglobin and comparing the same from tertiary databases. Future work will include study of three dimensional structures of hemoglobin and possible predicted helices, sheets and coils from the secondary structure.

REFERENCES

1. Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol. 1996;266:540–53.
2. Kloczkowski A, Ting KL, Jernigan RL, Garnier J. Combining the GOR v algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins. 2002;49:154–66. 8.
3. Sen TZ, Jernigan RL, Garnier J, Kloczkowski A, GOR V. server for protein secondary structure prediction. Bioinformatics. 2005;21:2787–8.
4. Lin K, Simossis VA, Taylor WR, HeringaJ.A simple and fast secondary structure prediction method using hidden neural networks.Bioinformatics. 2005;21:152–9. 29.
5. Martin J, Gibrat JF, Rodolphe F. Analysis of an optimal hidden markov model for secondary structure prediction. BMC Struct Biol. 2006;6:25. 30.
6. Won KJ, Hamelryck T, Prügel-Bennett A, Krogh A. An evolutionary method for learning HMM structure: prediction of protein secondary structure. BMC Bioinformatics. 2007;8:357.
7. Cuff, J. A., Clamp, M. E., and Barton, G. J.JPred: A consensus secondary structure prediction server. Bioinformatics .1998;14:892–893.
8. Holley, L.H. and Karplus, M.Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. USA (Biophysics).1989; 86:152–156.
9. Qian, N. and Sejnowski, T.J.Predicting the secondary structure of globular proteins using neural network models. Journal of Molecular Biology.1988; 202:865–884.
10. Salzberg, S. and Cost, S.Predicting Protein Secondary Structure with a Nearest-neighbor Algorithm. Journal of Molecular Biology.1992; 227:371–374.
11. Geourjon C., Deléage G.SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple sequences Comput. Appl. Biosci.1995; 11: 681-684.
12. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol. 1994;235:13–26.