Motif Structure Prediction in Distributed Framework using Machine Learning Algorithms

D. Shine Babu, Latha Parthiban, Sivagama Sundari. G

ABSTRACT--- It is a challenging work for researchers to design and develop new techniques for processing of data and development of new drugs. A distributed approach, which will work for huge amount of protein data and for predicting the motif structures in a large scale is proposed in this paper. ANNs has been used as classifier to estimate the motif structure of proteins. It will be helpful for the researchers and aids in understanding the relation between protein sequence and structure using which new drugs and novel enzymes can be designed after analyzing the protein structures.

Keywords— Bioinformatics, Big data, Map Reduce, Machine learning, Apache Hadoop, protein structure prediction

I. INTRODUCTION

Proteins are important for living organisms with 20 different amino acids that are present in them. These amino acid names are written in a three letter code but with vast diiferences. These amino acids that make up proteins can be grouped according to many criteria, including *hydrophobicity, size, aromaticity, or charge*.

There are four different structure types of proteins. *Primary structure* is the base structure for all types. Secondary structures combine to form Motifs like β-hairpin, α-α hairpin, helix turn helix, β- α- β. *Tertiary structure* is a combination of secondary structures which becomes polypeptide chains . *Quaternary structure* is created by tertiary structure[1].

Protein structure prediction is defined as “inference of the three-dimensional structure of a protein from its amino acid sequence; that is, the prediction of its folding and its secondary and tertiary structure from its primary structure”. In this work, a neural network based approach is proposed to identify the primary structures and also protein motif structure prediction. The paper is organized as follows: section 2 explains the related work, section 3 the proposed methodology and section 4 the results and section 5 the conclusions.

II. RELATED WORK

Many studies in literature showed the correlation that exists between sequence, structure and function [1-3]. Proteins that have related sequences and structures have related functions. An important goal in structural bioinformatics is prediction of 3D protein structure from its 1D sequence[4].

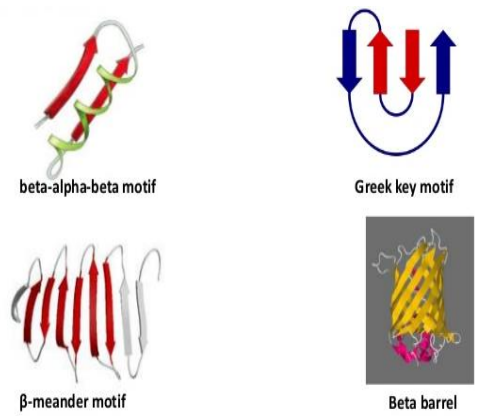
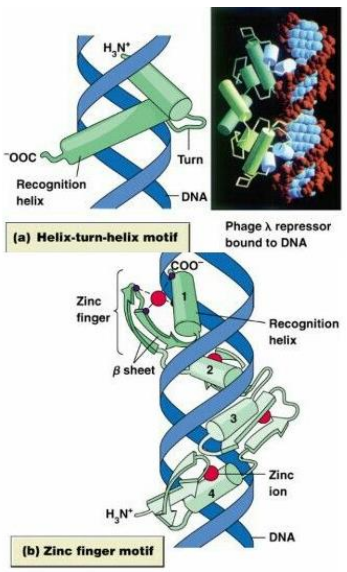


Fig 1 Super secondary structures (Motifs)

Figure 1 shows the super secondary structure which are motifs. Table 1 shows the primary and super secondary structures of proteins.

Table 1 :Primary and Secondary protein structures

Primary structure	APKDNTWYTGAKLGW…VSYRFG
Super secondary structure	LLLLLEEEEELHHHH…EEEELL



Revised Manuscript Received on February 11 , 2019.

D. Shine Babu, Research Scholar, Department of Computer Science and Engineering, Sathayabama Institute of Science and Technology, Chennai, India.

Dr. Latha Parthiban, Research Supervisor, Department of Computer Science, Pondicherry University CC, Puducherry, India. (E-mail: lathaparthiban@yahoo.com)

Sivagama Sundari. G, Associate Professor, M.V.Jayaraman College of Engineering, Bangalore, India. (E-mail: itsmevasigas@gmail.com)

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

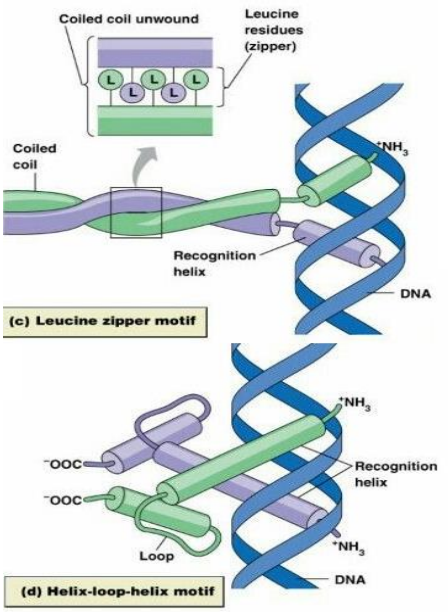


Fig 2 Protein folding motifs

In 3D structure along with α helix, beta sheet and beta turns and other non covalent interactions, the protein folds to form motif as shown in figure 2. Amino acid sequence in protein has information which help protein to fold to specific shape[5]. The Dictionary of Protein Secondary Structure (DSSP) is the term that describes protein secondary structure. Literature on PSSP (protein secondary structure prediction) shows that many research challenges has to be addressed[6]. Data mining and bioinformatics has many research issues for scalable and effective analysis. Tertiary structure prediction (TSP) gives structure and function of viral proteins that helps in drug design. [7].

Apache Hadoop is an open source innovation for handling and dispersing dependable information. The product developed from it helps in appropriated handling of tremendous informational collections. They are needed to scale from single servers to a great many machines offering neighborhood calculation and capacity. Its library is intended to discover and handle deficiencies at application layer. Figure 3 shows the map reduce framework

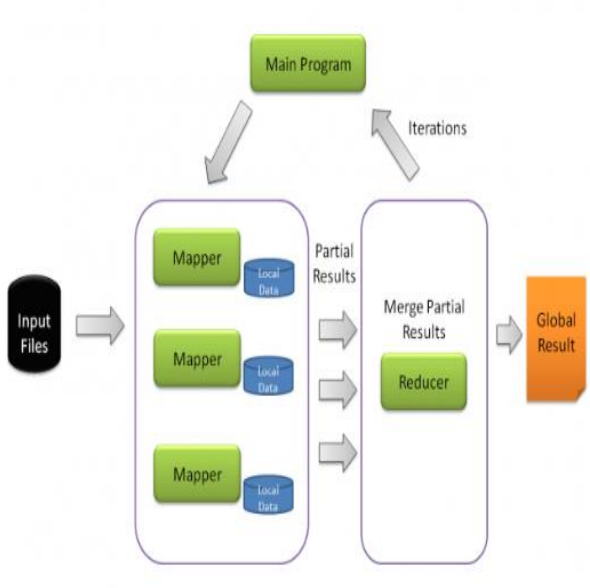


Fig. 3. A typical map reduce framework

Artificial neural network is a supervised machine learning technique for protein structure prediction. ANN produces output for specific input pattern. Many neural network structure like multilayer perceptron, radial basis function, back propagation network are available in literature[8].

III. PROPOSED METHODOLOGY

This section explains the proposed architecture and the system model

A. Proposed Architecture

Computers are always a very effective way of processing of huge amount of data. Big data framework provides such a platform which provides an efficient computing capacity, bandwidth, storage, security, and reliability of the system. PSSP is the process of identifying and predicting the structures among the Protein sequence and we are implementing the Machine Learning approach using Hadoop and Map Reduce concepts. Proposed architecture is shown in figure 4.In this paper, cloud era is used [7] and a framework is designed with distributed approach that enhances the accuracy of motif prediction .

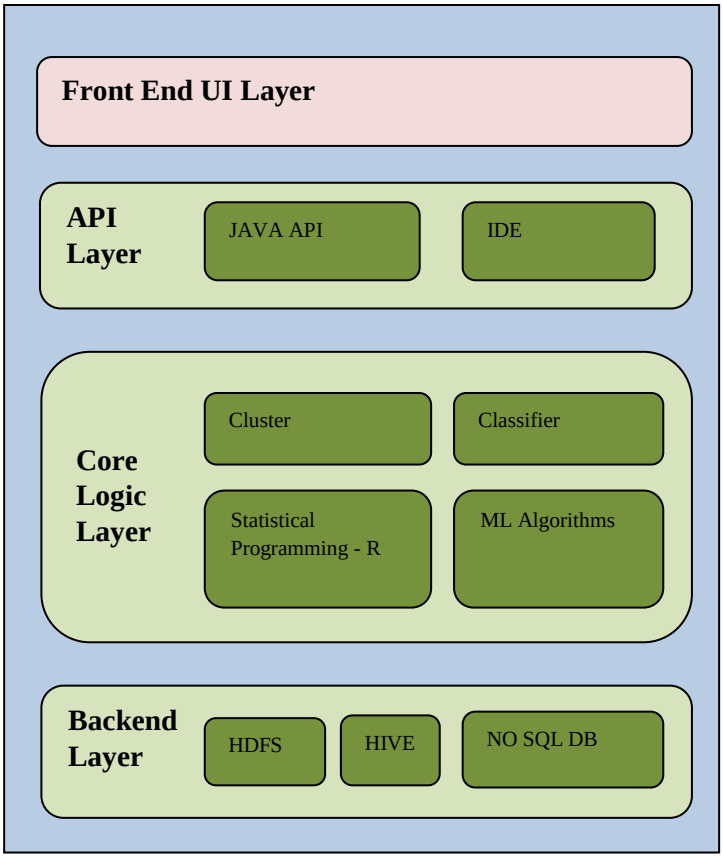


Fig. 3. Proposed Architecture

B. System Model

The flowchart of entire system model is shown in Fig 4. The system model for predicting the secondary structure of proteins is presented below.

1) Input Proteins Dataset from PDB

Five types of proteins are considered as input data. The proteins considered are Myoglobin and Hemoglobin which are transport proteins, insulin and adrenalines which are hormonal proteins and Lactase which is a Catalytic Protein. These proteins belong to different secondary categories and hence considered.

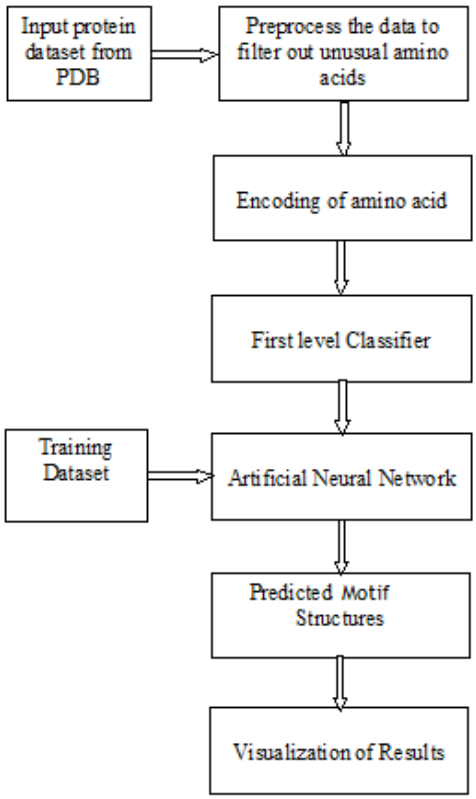


Fig. 4. System model

2) Dataset Collection

Amino acid codes for input proteins are taken from Protein Data Bank (PDB) along with DSSP codes.

3) Protein Encoding

For encoding protein sequence, unique alphanumeric coding scheme is used.

4) Training - Testing of Neural Network

The network is trained with the coded protein structures. After a model has been built, the network is validated by testing it.

5) Extraction of super secondary structures

The secondary structure are evaluated for the proteins based on inclinations of amino acids to form different secondary structures.

Figure 5 shows a predicted motif structure with its visualization

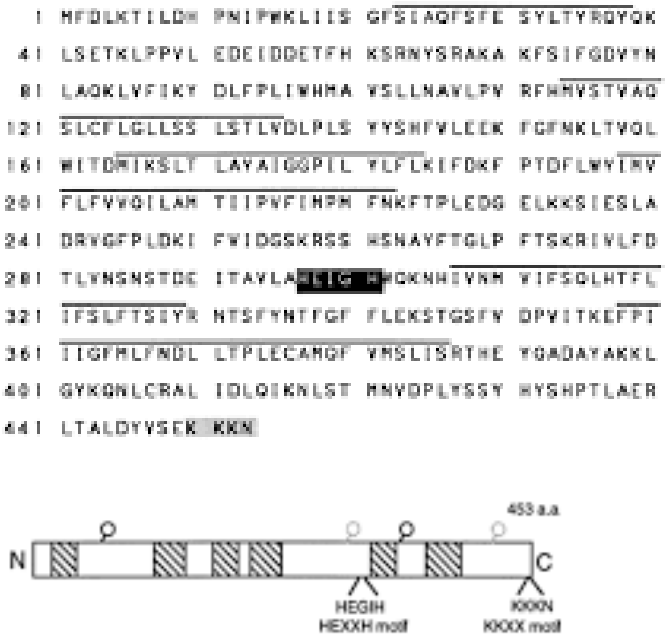


Fig. 5 Predicted motif structure and its visualization

IV. EXPERIMENTAL RESULTS

Table 2 shows the Hadoop cluster setup for two nodes The RAM capacity and processor speed may vary for the nodes. Figure 5 shows the structure prediction visualization and fig 7 shows the solvent accessibility prediction visualization. The methodology followed is presented in four steps.

- *Load Files part*

JAVA Map Reduce program is used to dynamically parse PDB [8] files based on the query protein pairs. Every coordination file is split into chunks and stored on HDFS.

Table 2: Hadoop Cluster set up

	Base			VMWare Player		
Node	**OS**	**RAM**	**Processor**	**OS**	**RAM**	**Hard Disk**
Node1	Windows Server 2008 R2 Enterprise	GB	GHz	Centos 6.2	6GB	500GB
Node2	Windows Server 2008 R2 Enterprise	GB	GHz	Centos 6.2	6GB	500GB

Fig 6 Structure prediction visualization

Fig 7 solvent accessibility prediction visualization

- *Mapper*

A mapper that has a primary key and a value runs its algorithm during mapping phase. It runs the partition data and produces the intermediate result. We can have 5 mappers for each of the proteins chosen. Each Mapper can incorporate the Mahout package to classify the data individually before passing it to the Reducer.

- *Algorithm part of Reducer*

A Reducer program runs during reducing phase and is aimed to collect intermediate results and output the result to HDFS. The Reducer will receive inputs from each of the 5 Mappers after proper Shuffle and Sort phase.

- *Statistical Programming (R) to visualize the results*

The R language is used for data analysis and help researchers efficiently to visualize the results.

V. CONCLUSIONS

In this paper, a distributed computing environment setup is proposed using Hadoop and Machine Learning Algorithms which are leveraged to efficiently parallelize the existing protein super secondary structure prediction algorithms. Big Data distributed computing technologies like Apache Hadoop will be helpful to develop a dedicated, error free, query table database for the research. Machine learning algorithms developed using Apache Hadoop Technology improves accuracy and speed creating more challenges and opportunities. Our future work will focus towards other machine learning algorithms like SOM to predict protein super secondary structure.

REFERENCES

1. H. Bordoloi and K. K. Sarma, ``Protein Structure Prediction Using Multiple Artificial Neural Network Classifier'''', as a Chapter of a volume titled Soft Computing Techniques in Vision Science, Studies in Computational Intelligence, 2012, Volume 395/2012, pp. 137-146, DOI: 10.1007/978-3-642-25507-6_12 , 2012.
2. H. Bordoloi and K. K. Sarma, ``Protein Structure Prediction using Artificial Neural Network'''', IJCA Special Issue on Electronics, Information and Communication Engineering ICEICE (3), pp. 24-26, December 2011. Published by Foundation of Computer Science, New York, USA.

3. Li, J. , Wu, J. and Chen, K. (2013) PFP-RFSM: Protein fold prediction by using random forests and sequence motifs. *Journal of Biomedical Science and Engineering*, **6**, 1161-1170.
4. A. Deka and K. K. Sarma,"Soft Computational Framework for Tertiary Protein Structure Prediction", International Journal of Electronics Signals and Systems (IJESS), ISSN:2231-5969, Vol.1, Issue 3
5. Chou, K.C. and Shen, H.B. (2009) Review: Recent advances in developing web-servers for predicting protein attributes. Natural Science, 2, 63-92.
6. Nanni, L. (2006) A novel ensemble of classifiers for protein fold recognition. Neurocomputing, 69, 2434-2437.
7. D. cutting, “Apache Hadoop is a new way for enterprises to store and analyze data.,” Cloud era, 2010.
8. Liu, L., Hu, X.Z., Liu, X.X., Wang, Y. and Li, S.B. (2012) Predicting protein fold types by the general form of chou’s pseudo amino acid composition: Approached from optimal feature extractions. Protein & Peptide Letters, 19, 439-449. .
9. Yang, T., Kecman, V., Cao, L., Zhang, C. and Huang, J.Z. (2011) Margin-based ensemble classifier for protein fold recognition. Expert Systems, 38, 12348-12355.