International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 12 (2019) pp. 2806-2811

Multi Agent Network-propelled Data Extraction for Protein Research

S Sulaiha Beevi
Research Scholar,
Bharathiyar University, Coimbatore, Tamil Nadu, India.

Dr. K L Shunmuganathan
Principal,
Aarupadai Veedu Institute of Technology (AVIT)
Vinayaka Mission's Research Foundation (VMRF) – Deemed to be University,
Chennai, Tamil Nadu, India.

Abstract

To comprehend the structure function model, a novel procedure for proteins classification and prediction is proposed. It uses multi agent system technique that represents a new standard for building software systems to predict and classify protein constructs. To categorize the proteins, support vector machine (SVM) has been developed to extract features from the protein sequences. This paper describes a method for predicting and classifying the subordinate structure of proteins. Support vector machine (SVM) modules were developed using multi-agent system principle for predicting the proteins and its functions there-by achieving the goals of accuracy, specificity, sensitivity, of 92%, 94.09%, and 91.59% respectively. The proposed algorithm provides an understanding of the protein structure, which can greatly improve biological science by analyzing the relationships amongst proteins.

Keywords: SVM, Protein Database, RNA, Protein Clusters, Classification and Multi-Agent System

1. INTRODUCTION

Protein structures are experimentally analysed using either x-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy. While both methods are increasingly being applied in a thorough manner, structure determination is not an easily understandable process. The scope for X-ray crystallography is reduced by the problem of making proteins form crystals. NMR can only be applied to relatively tiny protein molecules. Whole-genome sequencing efforts have led to large numbers of known protein sequences and their matching protein structures are found out at a markedly slower speed.

Even after decades of work in this domain, the problem of predicting the complete three- dimensional structure of a protein from its sequence remains a mystery. Nevertheless, computational methods can lead to advances in protein structure determination. Series-based methods are habitually made use of to help describe protein structure.

Proteins can be considered as the major components of living organisms which are the working molecules of cells. Proteins have genetic codes that describe the biological development of living organisms. So, proteins can be defined as polymers of amino acids that contain a main chain of repeating units with variable side chains. A protein is mainly made up of amino acids, which decide its structure.

The work, demonstrated a method for obtaining data related with voice system users. This included a range of conversations with many voice system users. For each conversation, a speech waveform was taken and digitized and the data of at least one audio feature was mined. The features were connected with a minimum of one attribute from the aspects of gender, age, accent, native language, dialect, socioeconomic classification, educational level and emotional state. Attribute data along with at least one identifying indicia were stored for each user in a data warehouse, in a particular form to ease subsequent data mining. The resulting collection of stored data was then mined to offer information for altering the business logic of the voice system. A gadget appropriate for executing the method included a dialog management unit, an audio capture module, an audio end, a processing module and a data warehouse. Appropriate method steps can be implemented by a digital computer running a suitable program stored on a program storage device. We proposed a novel classification method to identify the RNA binding sites in proteins by combining a new interacting feature (interaction propensity) with other sequence and structure-based features.

The study indicated the development of a novel method for predicting membrane protein types by exploiting the discrimination capability of the difference in amino acid composition at the N and C terminus through split amino acid composition (SAAC). In this study, membrane protein types were classified using three feature extractions and several classification strategies. In another work, proteins tertiary structure classification was identified by using agent system with four layers (Data fusion, Feature selection, Model building, and Knowledge discovery) and data mining method to predict a relative solvent accessibility (RSA) of 3D structure of most of the proteins, to extract hidden information from protein sequences.

1.1 The DSSP Code

Dictionary of Protein Secondary Structure (DSSP) represents an algorithm for assigning secondary structure of the protein to the amino acids. The types secondary structure is shown below.

- G = 3-turn helix (310 helix). Min length 3residues.
- H = 4-turn helix (α helix). Min length 4residues.
- I = 5-turn helix (π helix). Min length 5residues.
- T = hydrogen bonded turn (3, 4 or 5turn).
- E = extended strand in parallel and/or anti-parallel β-sheet conformation. Min. length 2 residues.
- B = residue in isolated β-bridge (single pair β-sheet hydrogen bond formation).
- S = bend (the only non-hydrogen-bond based assignment).

2. PROPOSED MULTI AGENT SYSTEM

Various traditional systems were applied to classify and cluster of proteins. The biologists used Nuclear Magnetic Resonance (NMR) and x- ray crystallography techniques to determine the proteins structure, but these methods took years to determine the structure of one protein. Consequently, today, in Protein Data Bank (PDB) there were over 1 million proteins whose amino acid sequences were known; however, only around 50,000 of these protein structures are still known. Therefore, having tools to classify and predict the structure of a protein is very important and necessary. In this paper, a new method for proteins data mining was developed by combining unsupervised clustering, logistic classifier, and multi-agent system to achieve the objectives. The method was based on the multi-system agents' principle to predict protein and all that were related to it, which is based on the segmentation system. For example, there was an agent responsible for the clustering process and another for the receiving process, which aimed at combining the proteins and multiple structures.

The proposed system is composed of several work agents, each one made up of one or several resources. The multi-agent architecture is shown in figure 1. Initially, the system obtains DSSP code from the secondary protein structure, and then generates the feature vector from feature extraction. Next, the classification of DSSP as group is done based on the

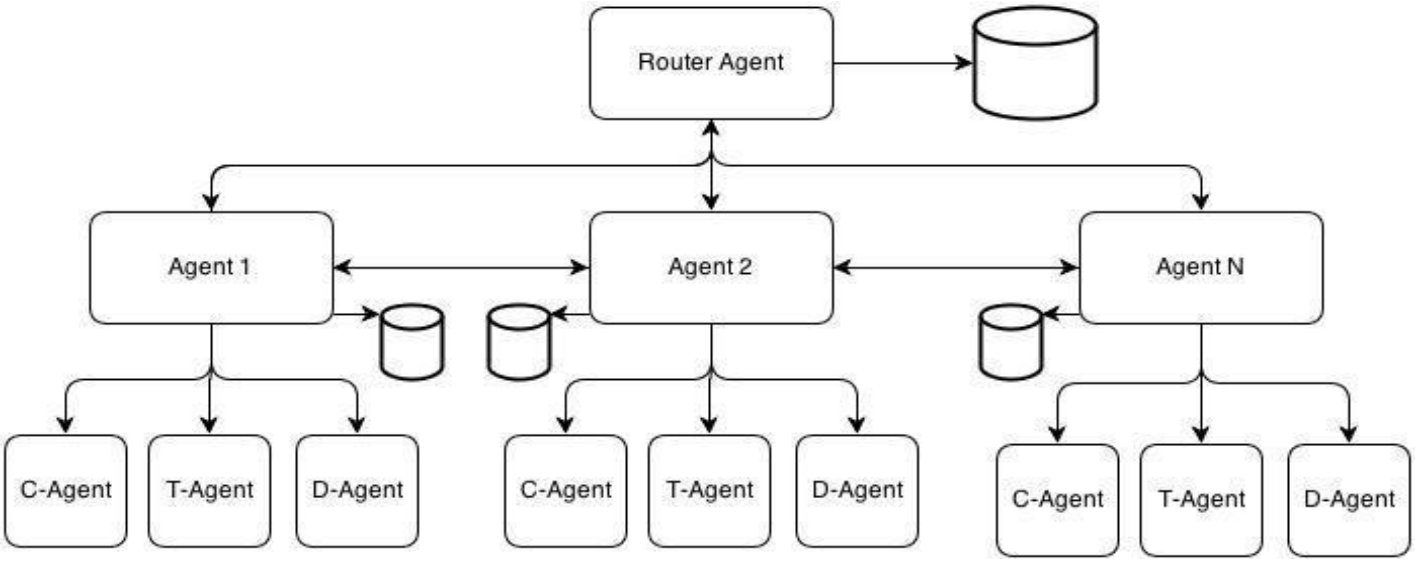

Figure 1: Proposed Agents System Architecture

functionality, structure, or shape. In each group the cluster can be split into several types by using the logistic classifier principle. The system consists of (n) agents that work with together to perform a prediction and classification of protein, and this is done through several stages by getting into the DSSP of secondary protein structure from the dataset by the router agent. This agent will send this dataset to several agents that contains private databases, which in turn are applied to the Principle Support Vector Machine (SVM) using three sub agents (C, D, and T agents). In each one of these sub agents, the descriptor represents a vector. Finally, each agent is responsible for a particular function. The agent is directed to analyze the data and return any of the groups that belong to this sample of the structure of protein if it is present in the original database, otherwise it creates a new group within the specifications continues the work program until the largest amount of different types of proteins are extracted.

Feature Extraction

Feature extraction is important for a variety of reasons: generalization performance, running time requirements, constraints and interpretational issues imposed by the problem itself. Constructing an effective feature vector to represent a protein is the key step for successful SVMs proteins classification.

Feature Vectors

For every secondary sequence, the feature vector was assembled from encoded representations of secondary structure composition. There are three descriptors that are used to predict the protein structure from DSSP sequence.

The Composition Descriptor Vector

This vector represents the percentage of certain amino acids as shown in the equation below.

- *100%

The Transition Descriptor Vector

This vector calculates the frequency of which is followed by amino acids of a different property as shown in the equation below.

- *100%

The Distribution Descriptor Vector:

This vector measures the percentage of sequence length within which the first 25%, 50%, 75%, and 100% of the helix of a certain property, such as polar, is located respectively.

Multi Agents System

Composition Agent (C-Agent)

The C-agent is the agent that is responsible for calculation of the residues in the DSSP sequence for translating the DSSP to feature vector.

Transition Agent (T-Agent)

The T-Agent: is the agent that responsible for transitions process between residues in the DSSP sequence.

Distribution Agent (D-Agent)

The D-Agent is the agent responsible for distribution process, where it calculates the percent of sequence length by splitting the protein sequence into four quarters25%,50%,75%, and 100%, respectively, these processes can be explained using the below example. The process can be explained with a simple example:

Ex.1 Assume the DSSP as the follow:

DSSP	C	T	E	C	E	E	T	T	E	C	T	C	E	E	C	C	E	E	E	T	T	T	C	T	E
Index	1				5					10					15					20						25

DSSP sequence for illustration of derivation of the feature vector, DSSP structure consisting of 3 structures E, β-strand; T; and C.

Composition Agent (C-Agent):

- ✓ **Step 1:** Count the frequency of each residue. The result was as follow:
  - The DSSP according to the above example contains 25 amino acid distributed among (C, T and E) residues.

- The DSSP model sequence contains 7 type of C residue, 10 types of E residue and 8 types of T residue.

✓ **Step 2:** Calculate the frequency ratio of each residue in total of DSSP sequence as the follow

- The C descriptor for Cs = 7/25*100=28%, 10/25*100%=40% for Es, 8/25*100%=32% for Ts, respectively.

Transition Agent (T-Agent):-

✓ **Step 1:** Calculate the transition between residues as the follow:

- There are 3 transitions between C and T, 2 between T and C, 3 between T and E, 3 between C and E, 2 between E and T, and 3 between E and C.

✓ **Step 2:** Calculate the transition ratio for each residue in the DSSP sequence

- The frequency ratios of these transitions are-

3/24=12.5% for transition between C and T,

2/24=8.3% for transition between T and C,

3/24=12.5% for transition between T and E,

3/24=12.5% for transition between C and E,

2/24=8.3% for transition between E and T, and

3//24 =1205% for transition between E and C, respectively.

Distribution Agent (D-Agent):

✓ **Step 1:** Spilt the DSSP sequence into four quarters (25%, 50%, 75%, and100%).

✓ **Step 2:** Determine the locations of each residues as follow:

- Cs residues are located within 1, 4, 12, 16, and 23, respectively.
- Es are located within 3, 6, 13, 18, and 25, respectively.
- Ts are located within 2, 7, 11, 21, and 22, respectively.

✓ **Step 3:** Calculate the distributed descriptor of each residues as follow

- D descriptor in the first, 25%, 50%, 75% and 100% for Cs is 8%, 16%, 48%, 64%, and92%.
- D descriptor in the first, 25%, 50%, 75% and 100% for Es is 12%, 24%, 52%, 72%, and 100%.
- D descriptor in the first, 25%, 50%, 75% and 100% for Ts is 8%, 28%, 44%, 84%, and88%.

✓ **Step 4:** Calculate the sequence descriptors for all residues in the DSSP sequence as package as follows:

C = (28%,40%,32%),

T = (12.5%,12.5%,12.5%, 8.3%, 12.5%, 8.3%), and

D = (8%, 16%, 48%, 64%, 92%, 12%, 24%, 52%, 72%, 100%,8%, 28%, 44%, 84%, 88%).

After the feature vectors were constructed, normalizations were performed for each dimension of vectors in the training data set to adjust the values of all the feature vectors to a standard level. The normalization function is

Where, which is the sample standard deviation of x, and ∑ respectively.

3. RESULT ANALYSIS

The data set consisted of a positive subset and a negative sub set. Protein in the positive sub set was known to have the function that the SVM was trained to recognize. Proteins in the negative subset were known not to have that function. When analyzing the results about prediction of protein secondary structure in data set, a set of concepts and measurements by which performance was measured had been taken into account. This affected the accuracy, specificity, and sensitivity of the system performance.

Table 1: The performance of the SVMs

TP Rate	FP Rate	Precision	Recall	F - Measure	Class
1	1	0.995	1	0.998	No
Weight Avg.	0.995	0.995	0.991	0	Yes

Accuracy	Specificity	Sensitivity	Mean Absolute Error
92%	94.09%	91.59%	0.0098

The results indicate that the SVMs trained by the amino acid physicochemical properties and sequence amino acid composition have a certain level of capability to classify proteins that are distantly related by sequence. SVM can fine the common factor in a diverse set of training data set and use the common factor to fine the optimal classification. Thus, this proposed method may be used as a complementary method to those sequence alignment methods in protein function prediction. Results showed that SVMs are comparable or superior to other existing machine learning methods in handling those problems. Thus, SVMs may be utilized to solve protein classification problems and complement the methods based on sequence similarity.

4. CONCLUSION

The proposed multi agent system for proteins classification and prediction used the DSSP sequence of proteins. The DSSP sequence was parsed and converted into composition, transitive, and distribution vector of features, by the usage of multi-agent systems. The multi agents worked in a parallel way, and each agent was dedicated for a particular task. Hence, in the proposed system there were three types of agents. The C-agent, T-agent, and D-agent were responsible for constructing the composition vector, transition vectors, and distribution vector correspondingly.

REFERENCES

[1] Alex B. and Stephen J. S., (1997). Data Warehousing, Data Mining, and Overlap, 1st edition, McGraw-Hill, Inc. New York, USA.

[2] Alireza, M., Nasser, G. A., Mehadi, S., (2011). Modeling and implementing an agent–based system for prediction of protein relative solvent accessibility, Expert Systems with Applications, vol. 38, pages: 6324-6332.

[3] Altun G., Hu H-J., Brinza D., Harrison R.W., Belkovsky A. and Pan Y. (2006). Hybrid SVM kernels for protein secondary structure prediction, Proc. IEEE Intl Conference on Granular, Computing (GRC 2006), pages 762-765.

[4] Aydin, Z., Altunbasak, Y., and Borodovsky, M. (2006). Protein secondary structure prediction for a single- sequence using hidden semi-Markov model, BMC bioinformatics. Vol. 7: 173-175.

[5] Javier B. P. (Eds), (2012). Highlights on practical Applications of Agents and Multi-Agent Systems, Springer, Spain .

[6] Liu,Z.P., Wu,L.Y., Wang,Y., Zhang,X.S. and Chen,L., (2010). Prediction of protein-RNA binding sites by a random forest method with combined features. Bioinformatics, 26, 1616–1622.

[7] Maqsood H., Asifullah K.,(2011). Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition, Vol. 271, Issue 1, Pages10–17,Pakistan.

[8] Rafael H. B., Mehdi D. and Jürgen D,(2009). Multi- Agent Programming, Springer.

[9] Sherwood, Dennis, Cooper ,and Jon, (2011). Crystals x-rays and proteins : comprehensive protein crystallography,621 p,USA.

[10] Balakrishnan. S and K L Shunmuganathan. Article: A JADE Implementation of Integrated Agent System for E-Mail Coordination (IASEC). International Journal of Computer Applications 58(5): 5-9, November 2012.

[11] P.Arivazhagan, Balakrishnan. S and K L Shunmuganathan. “An Agent Based Centralized Router with Dynamic Connection Management Scheme Using JADE”, International Journal of Applied Engineering Research, ISSN 0973-4562, Volume 11, Number 3 (2016) pp 2036-2041.

[12] Balakrishnan. S and K L Shunmuganathan, R. Sreenevasan, “Amelioration of Artificial Intelligence using Game Techniques for an Imperfect Information Board Game Geister” International Journal of Applied Engineering Research (IJAER). ISSN 0973-4562. Vol 9, Number 22 (2014) pp. 11849-11860.

[13] Balakrishnan. S and K L Shunmuganathan, An Agent Based Collaborative Spam Filtering Assistance Using JADE”, International Journal of Applied Engineering Research, ISSN 0973-4562, Volume 10, Number 21 (2015) pp 42476-42479.

[14] A.Jebaraj Rathnakumar, S.Balakrishnan, Design Of Multi-Agent Based Systems For Entrusted Communication Using JADE”, Taga Journal of Graphic Technology, Vol. 14, pp. 766-774, 2018.

[15] Balakrishnan, S., Janet, J., Sujatha, K., & Rani, S. (2018). An Efficient and Complete Automatic System for Detecting Lung Module. *Indian Journal Of Science And Technology, 11*(26). doi:10.17485/ijst/2018/v11i26/130559

[16] P.Palanikumar, S.Geofrin Shirly, Balakrishnan S, "An Effective Two Way Classification of Breast Cancer Images", International Journal of Applied Engineering Research, ISSN 0973-4562, Volume 10, Number 21 (2015) pp 42472-42475.