Towards A Multi Agent System Based Data Mining For Proteins Prediction And Classification

Mohammad Khaled Awwad Al-Maghasbeh, Habes Mahmoud Masoud Al-Khraisat

Abstract: To understand the structure function paradigm, in this paper, a new algorithm for proteins classification and prediction is proposed. It uses multi agent system technique that represents a new paradigm for conceptualizing, designing and implementing software systems to predict and classify the protein structures. For classifying the proteins support vector machine (SVM) has been developed to extract feature from the proteins sequences. This paper describes a method for predicting and classifying secondary structure of proteins. Support vector machine (SVM) modules were developed using multi-agent system principle for predicting the proteins and its function, and achieved maximum accuracy, specificity, sensitivity, of 92%, 94.09%,and 91.59% respectively. The proposed algorithm provide a good understanding for proteins structure, which affect positively on biological science specially on understanding the behavior, and the relationships between proteins.

Keywords: Protein Structure, Protein Database, DNA, RNA, Protein Clustering and Classification, Multi- Agent System.

Introduction

Protein structures are determined experimentally using either x-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy. While both methods are increasingly being applied in a high-throughput manner, structure determination is not yet a straight-forward process. X-ray crystallography is limited by the difficulty of getting some proteins to form crystals, and NMR can only be applied to relatively small protein molecules [12]. As a result, whereas whole-genome sequencing efforts have led to large numbers of known protein sequences, their corresponding protein structures are being determined at a significantly slower pace. On the other hand, despite decades of work, the problem of predicting the full three-dimensional structure of a protein from its sequence remains unsolved [9]. Nevertheless, computational methods can provide a first step in protein structure determination, and sequence-based methods are routinely used to help characterize protein structure [5]. The proteins can be considers as the major components of living organisms and are considered to be the working and structural molecules of cells. Proteins contain genetic instructions that describe the biological evolution of living organisms. So the proteins can be simply defined as polymers of amino acids containing a constant main chain or backbone of repeating units with a variable side chain attached to each. A protein is primarily made up of amino acids, which determine its structure [5].

Related Work

Alex, et al., in their work, showed a method for collecting data associated with the voice of a voice system user includes conducting a plurality of conversations with a plurality of voice system users.

- *Mohammad Khaled Awwad Al-Maghasbeh, masters degree in computer science Al Balqa' Applied University, As salt- Jordan.*
  E-mail: m.maghasbeh@gmail.com
- *Habes Mohmoud Masoud Al-Khraisat, PhD of computer science, Assistant professor at Al-Balqa' Applied University, Email: h.alkhraisat@gmail.com*

For each conversation, a speech waveform is captured and digitized, the data least one acoustic feature is extracted. The features are correlated with at least one attribute such as gender, age, accent, native language, dialect, socioeconomic classification, educational level and emotional state [1][2]. Attribute data and at least one identifying indicia are stored for each user in a data warehouse, in a form to facilitate subsequent data mining thereon. The resulting collection of stored data is then mined to provide information for modifying underlying business logic of the voice system. An apparatus suitable for carrying out the method includes a dialog management unit, an audio capture module, an acoustic from end, a processing module and a data warehouse[1][7]. Appropriate method steps can be implemented by a digital computer running a suitable program stored on a program storage device. Present the agent and multi agent system that has been matured during the last decade and many effective applications of this technology within an international forum to present and discuss the latest scientific developments and their effective applications, to assess the impact of the approach, and to facilitate technology transfer, [8]. 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,[9][14]. While other study indicated about developing 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 are classified using three feature extraction and several classification strategies, [10]. In other work, proteins tertiary structure classification has been 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 or tertiary structure of most proteins, to extract hidden knowledge and information from protein sequences [6].

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. There several types of the secondary structure as follow below:

- G = 3-turn helix (3_{10} helix). Min length 3 residues.

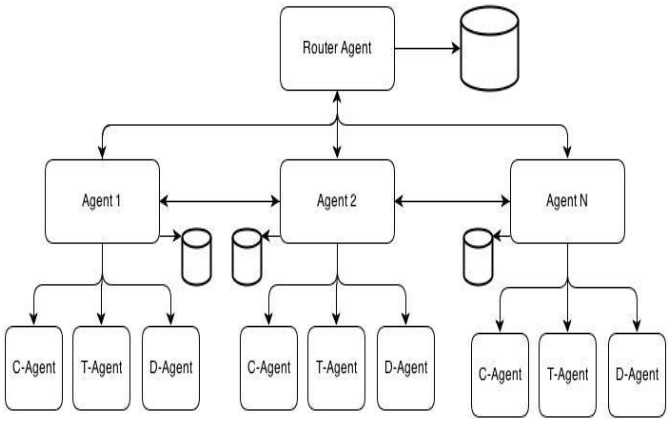
- H = 4-turn helix (α helix). Min length 4 residues.
- I = 5-turn helix (π helix). Min length 5 residues.
- T = hydrogen bonded turn (3, 4 or 5 turn).
- E = extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues.
- B = residue in isolated β-bridge (single pair β-sheet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

Proposed Multi Agent System

Different traditional systems or techniques have been applied to classify and cluster of proteins. The experimental biologists use Nuclear Magnetic Resonance (NMR) and x-ray crystallography techniques to determine the proteins structure, but these methods take years to determine the structure of one protein [12]. Consequently, today, in Protein Data Bank (PDB) there are over 1 million proteins whose amino acid sequences are known; however, only around 50,000 of these protein structures are known. Therefore, having tools to classify and predict the structure of a protein is very important and necessary [8]. In this paper, new method for proteins data mining has been developed by combining of strengthen of unsupervised clustering, logistic classifier, and multi-agent system to achieve the method objectives[11]. The method based on the multi-system agents principle to predict protein and all that is related to it, which is based on the segmentation system with large size into several subsystems have been the representation of each small system or process represent an agent has a specific function, for example, there is agent responsible for the clustering process and another on the receive process, which aims combined to predict the expected proteins, different types, and multiple structures.

Proposed Agents System Architecture

The proposed system is composed of several work agents, each one made up of one or several resources. The multi-agent architecture is shown below.



Proposed a Multi-Agent System Architecture.

Firstly, the system gets DSSP code from secondary protein structure, and then generates the feature vector from feature extraction. Secondly, after this process makes the classification of DSSP as group based the functionality, structure, or shape. In each group can be cluster into several type of cluster 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 of going through the system starting to get on the DSSP of secondary protein structure from the dataset by router agent that designed for this purpose. where this agent shall to send this sample of dataset to several agents and that contain on private databases, which in turn is applied to the principle Support Vector Machine(SVM) using three sub agents (C,D, and T agents). In each one of these sub agents called the descriptor that represents a vector [3]. Finally, in each agent it's responsible for a particular function, and then does the agent directed 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 and certain properties, and so continue the work program until you get to know the largest amount of different types of proteins, which have positive impact on the biological sciences in general and in particular human.

Feature Extraction

In many supervised learning problems feature extraction is important for a variety of reasons: generalization performance, running time requirements, and constraints and interpretational issues imposed by the problem itself. Constructing an effective feature vectors to represent a protein is the key step for successful SVMs proteins classification [3][5].

Feature Vectors

In this paper, 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 as the follow:-

The Composition Descriptor Vector

This vector represents the percent of amino acids of certain properties as shown in the equation that is mentioned below.

- $\text{The composition descriptor} = \frac{residue frequency}{total\ of\ sequence} * 100\%$

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 that is mentioned below.

- $\text{The composition descriptor} = \frac{residue\ transition\ frequency}{total\ of\ sequence-1} * \ 100\%$

The Distribution Descriptor Vector:

This vector measures the percent of sequence length within which the first 25%, 50%, 75%, and 100% of 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 the residues frequency in the DSSP sequence in order to translate the DSSP to vector of feature.

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 calculate 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, Tand E) residues.
  - The DSSP model sequence contains 7 type of C residue, 10 types of E residue and 8 type 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 frequencies ratio 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%, and 100%).
- **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.
- **Step3**.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%, and 92%.
  - 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%, and 88%.
- **Step 4**. Calculate the sequence descriptors for all residues in the DSSP sequence as package as follow :
  - 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 among 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

$$\chi_{ni} = \frac{\chi_i - \bar{\chi}}{S_x}$$

Where $S_x = \sqrt{\sum_{i=1}^{n} \frac{(\chi_i - \bar{\chi})^2}{n-1}}$, which is the sample standard deviation of *x*, and $\bar{\chi} = \sum_{i=1}^{n} \frac{\chi_i}{n}$.

Result Analysis
The data set consisted of a positive subset and a negative sub set. Protein in the positive sub set were known to have the function that the SVM were trained to recognize. Proteins in the negative subset were known not to have that function. When begin analyze the results about prediction of protein secondary structure in data set was taking into account a set of concepts and measurements by which performance is measured such as accuracy, specificity, and sensitivity of the system performance.

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

Table1. *The performance of the SVMs*

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.

Conclusion

The proposed multi agent system for proteins classification and prediction uses the DSSP sequence of proteins. The DSSP sequence is parsed and converted in to composition, transitive, and distribution vector of features, by the use of strengthen of multi-agent system. The multi agents are working in a parallel way, and each agent is dedicated for a specific task, therefore in the proposed system there are three types of agent. The C-agent, T-agent, and D-agent are responsible for construct the composition vector, the transition vectors, and distribution vector respectively.

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., Zelikovsky 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] Cai,C.Z., Han,L.Y., Ji,Z.L., Chen,X. and Chen,Y.Z.,(2003). SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. Vol.31: 3692–3697.

[6] Ian H.W., Eibe F., (2005). Data mining : practical machine learning tools and techniques, 2nd edition, Diane Cerra, France .

[7] Ian, H. W., Eibe, F., and Mark A. H., (2011). Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann.

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

[9] 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.

[10] Maqsood H., Asifullah K.,(2011). Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition, Vol. 271, Issue 1, Pages 10–17,Pakistan.

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

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

[13] Steinwart and A.,(2008). Christmann. Support vector machines**.** Springer, New York.

[14] Xiaojing Y., Jianping C., *, (2006). Predicting rRNA-, RNA, and DNA-binding proteins from primary structure with support vector machines, pp.175-184,jornal of theoretical biology, China.