An Efficient Way of Finding Optimal Path using Protein Data Set: Ant Colony Optimization with Rough Set Theory for Feature Selection

A.Revathi[#1], S. Dhanakotteeswaran[*2]

[1]*PGT Computer Science, SRM Nightingale Matriculation HSS, West mambalam, Chennai-33,*

[2]*Assistant Prof., Dept of Computer Science, New Prince Shri Bhavani Arts & Science College, Medavakkam, Chennai-100,*

Abstract**—** Bioinformatics is one of the emerging technologies which is played an important role in the field of biology. The molecular biology and Bioinformatics information are extracted from the protein data set which is used for analysing the different kind of biological information. The major challenges in the protein data set are larger in size, which increases the complexity during the further experimental process. The complexity of the system is reduced by hybridized Soft Computing techniques and Evolutionary Methods. Thus, in this paper proposed that optimal feature selection method for reducing the dimensionality of the protein feature set to improve the performance of the proposed system. Initially, the biological data are grouped into the clusters which is fall into the pre-processing step for removing the missing and unwanted data's. The cluster formation is done by Ant Colony with Rough Set Theory (ACRST) based feature selection process. The performance of the system is evaluated with the help of the existing algorithms such as wrapper method, Greedy Forward Selection, Particle Swarm Optimization, Scatter Search and the comparison is analyzed with the help of the accuracy, sensitivity and specificity..

Keywords— *Bioinformatics; Protein Dataset, Soft Computing; Feature Selection; Ant Colony based Rough Set Theory; Greedy Forward Selection; Protein Data Bank*

1. Introduction

Proteins are the huge volume of macromolecules or biomolecules which consist of long chain of amino acids [1]. The proteins are performed several functions in the living organisms such as DNA replications, catalyzing metabolic reactions, product manufacture, routine maintenance, transporting molecules, waste cleanup, shape and inner organization [2]. The protein structure differs from one another by means of amino acid sequence, which is the direction of the nucleotide sequence and determines the activities of the cell organisms. So, the activities of the cell organisms are determined by the protein function which is analyzed based on the genomic sequence of data [3]. The information is gathered from the nucleic acid homology sequence, gene profiles, protein domain structure, phenotypic profiles, phylogenetic profiles and protein-protein interaction. These gathered information's are used to analyze the cellular pathway because the single protein may play several roles in the organisms. Thus the function of the protein is predicted by several prediction methods such as Homology-based methods, Genomic context-based methods, Network-based methods, Structure-based methods and Sequence motif-based methods [4]. But the protein data set has several feature vectors it is difficult to process during the function prediction and other processing. So, the dimensionality of the feature vector is reduced by using the data mining and soft computing approaches.Then the evaluation of the protein data set sequence prediction process is enhanced by the enormous amount of dataset because it has various potential information[5].

Thus the Bigdata approach based huge volume of protein dataset is used for predictingthe protein function because it has thousands of RNA and DNA informatics in terms of volume, velocity veracity, variety,variability and complexity information's [6]. This Big data based analysis overcomes issues in various applications, especially in Bioinformatics, protein sequences, protein secondary structure analysis and sub cellular localization prediction. There are two reasons [7] how the Bigdata approach helps during the protein sequence and function prediction process. First, it provides important information about the protein structure,details during the analyzing process which means it just explains that which and how the proteins works in the cellular organisms. Second, the bigdata method provides additional information such as RNA, DNA, cell structure which is used to analyze the particular person complete growth progress records.So, the big data approach is played on the essential role during the protein function prediction and sequence analysis and related research.

This paper handling the data mining and soft computing techniques to obtain the knowledge based system for analyzing the protein sequence from the huge volume of unorganized dataset. Thus the main contribution of the paper is as per following: initially the feature set is minimized by applying the Ant Colony based Rough Set Theory (ACRST)approach which reduces the data set for selecting the optimized features. This optimal features are used to further classification or analyzing process. The performance evaluation is made with the several existing methods such as wrapper method, Greedy Forward Selection, Particle Swarm Optimization, Scatter Search. The rest of the paper is organized as follows: Section 2 provides the related discussion and researches about the protein sequence analysis methods, section 3 delinates and discuss about the protein sequence informatics, section 4 explains that the feature selection process, section 5 discuss the experimental results of the proposed system and finally, conclusion.

2. Related Works

Proteins are the most necessary and versatile biomolecules of life and building the blocks and functional unit of a cell. This section provides the various discussions and research experience related to the protein sequence and prediction function.

Bagyamathi et al., [8] proposed that Harmony Search algorithm with Rough Set theory for analyzing protein sequence in the big data. This paper uses the protein sequence data set and the amino acid based features are extracted those the tuples are selected based on the improved optimized technique which reduces the complexity in the big data based analyze. Then the performance of the system is compared with the Rough Set based PSO Quick Reduct and Set Quick Reduct. *Muhammad Javed Iqbal et al., [9]* focuses on the bioinformatics based data analyzing and managing method for managing the protein data set and their sequences. This paper uses the three different data set for improving the protein data analyzing and storage methods. Initially the features are considered as the descriptor which size is limited up to 3 amino acids and the irrelevant features are eliminated. Then the features are selected based on the threshold value and the selected features are used to further classification and analyzing process. Thus the paper implements the simplest and easy feature selection and classification approach which reduces the dimensionality of the protein data set.

Wen-Yun Yang et al., [10] proposed that sequence based approach for localizing the protein sub cellular and predicting the protein functions. The amino sequences are extracted by using the term frequency, discriminate analyzing and statistical based approaches. The proposed extracted features compare with the amino acid composition method, voting system, amino acid tuple approach and then the performance is evaluated by 7579 eukaryotic protein sequences of 12 sub cellular locations which is obtained by 5 fold cross validation based five predictors.

Kocbek et al., [11] uses different feature selection techniques which are used to analyze the relation between the protein and their stability. Thus the features are placed an important role during the protein sequence analyzing process which is compared with different feature selection methods.

Eun-Mi Kim et al [12] focuses on the DNA patterns or protein data set for analyzing and classifying the bioinformatics because it has huge information. The raw DNA or protein data set is difficult to process so, the Support Vector Machine and the noise whitening method are used for better bioinformatics classification problem. Then the performance of the system is evaluated with the abnormal protein sequence ovarian cancer data set.

Ben Blum et al [13] focuses on the protein structure prediction system with minimum energy consumption, which is obtained by applying the Rosetta method that uses the Monte Carlo energy minimization technique. Initially the features are selected by L1 regularized linear regression and the decision tree method and resampling technique is used to reduce the energy during the structure identification process. Then the performance of the system is estimated with the help of the nine small protein benchmark data and the prediction of the protein structure is demonstrated.

Elena Marchiori et al[14] proposed that the selecting and ranking based method for identifying the relationship between the protein alignments because which plays the major role during the function identification. This paper uses the SMAD and MH2 domain factors based data set and the proposed method selecting the nearest neighbor relevant features for protein factor selection. From the selected features the subtype specific sites are identified and selected which is compared with existing approaches.

Chandran et al [15] focuses on the protein feature extraction and feature selection process because it is used to improve the further analyzing process. This paper uses the single sequence PDS based retrieved data set and the amino acid or k-tuple features are extracted. The extracted features are selected by applying the enhanced fuzzy based rough set theory which eliminates the noise and unwanted data from the single sequence data set. The performance of the system compared with the existing methods. Thus the contribution of the paper is, huge volume of data set is used for analyzing the protein sequence by optimized feature selection approach. During the feature selection process the noisy features are removed also dimensionality of the features also reduced, which improves the further process that is discussed in the following section.

3. Protein Sequence Informatics

Proteins are the most important and versatile biomolecules of living cell which is used for creating, building blocks and functional components of organisms. This protein plays a second major role in the living organisms and it has several polypeptides in terms of amino acids [16]. In protein twenty different amino acids are available which are denoted as the various alphabets that is represented as the linear sequence which is called as the primary structure. The protein structure is shown in following figure 1.

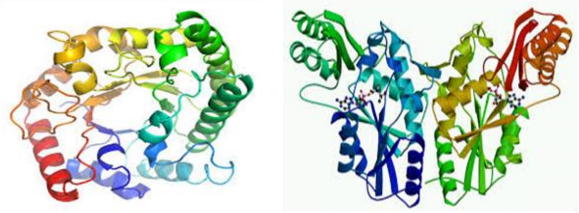

Fig. 1: Sample Protein Structure

This protein is having three different structures, namely primary, secondary and Tertiary structure which is used during the protein stability structure analyzing process. Then the stability of the protein structure is analyzed with the help of the spectroscopic methods such as, UV, infrared, CD and fluorescence [17]. This analyzed the protein structure is used to identify the protein functions, encoded proteins, missing functionalities, peculiarities and novelities are predicted in the human organisms. The Function is nothing but to do all the natural activity which is done with the help of the protein. These functions are analyzed in terms of computational analysis, structure analysis, mutation and functional analysis. This functional analysis is used to analyze the RNA structure, Gene profile, family classification and so on. So, the protein function prediction process is the important concept in the data mining process to classify the several problems [18]. The classification process is represented in terms of the hierarchy of classes which has some of the challenges such as it classifies the protein data into different levels and it requires various characteristics of protein data. This paper uses the big data based protein data set that uses the sequence analyzing and classification problems. But the major issues in the big data based data set is, it has high dimensional data's which consumes more time to process the protein data. Thus the paper overcome the above issues by using the feature selection process which reduces the dimensionality of the data set also removes the unwanted data. This selected data is used for further analyzing process.

4. Protein Feature Selection using Data Mining Techniques

The main goal of this paper is to analyze the protein data set with the help of the reduced feature dataset. This reduced or selected features are used to analyze the structure, behavior of the RNA, DNA characteristics, family classification and other classify problems. In this paper three different big data based protein data sets such as VariBench [19], BioGrid [20]and PDB data set[21] are used. This data set have various number of features and noise data which reduces the functionality and performance of the further system so, it has to be reduced by applying the feature selection method. Thus the paper proposed optimized Ant Colony with Rough Set Theory feature selection approach for reducing the dimensionality of the huge volume of protein data set. The basic architecture of the proposed system depicted in the figure 2. In the below figure 2 select the optimized features such as peptide residue, polarizability, amino acid composition (C), predicted secondary structure (S), polarity (P), normalized van der Waals volume (V), hydrophobicity (H) by using the Ant Colony with Rough Set Theory approach. This selected feature used for further sequence analyzing, classification problems which is explained in the following section.

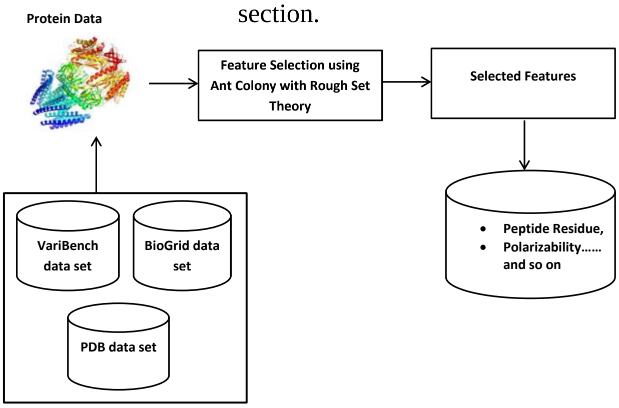

Fig.2: Proposed Architecture for Protein Feature Selection Process

4.1 Ant Colony Optimization

Ant Colony Optimization (ACO) is one of the probabilistic and optimized technique that is used to identify the best path for searching the food [22]. Generally ant finds the food by randomly wandering and return the same path, while wandering ant's laying the chemical pheromone trail on the ground. This chemical pheromone trail plays an important role during the other ant's searching the best path in the food ground. But these chemical pheromone trails have some issue such as it is evaporating faster and it reduces the attractive length of the path. Even though the shortest path evaporates its pheromone very faster, its density should high that is used to avoid the convergence of the local optimal solution. Thus the basic idea behind the ACO is biological ant' mimic behavior with simulated ant's that is guidance to find the shortest path from nest to food [23]. Thus the biological ant's create the indirect communication during the food searching, which is shown in following figure 3.

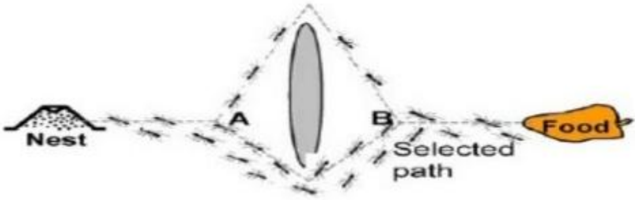

Fig.3: Sample Indirect Communication based Shortest Path food searching

During the indirect communication, the ant's have following characteristics such as, it creates positive feedback, demonstrate distributed computational architecture, exploit global data structures which change dynamically of each ant traversing route, distribution of the computation is depends on the population of the ant's and the probability of the transition always depends on both pheromone also problems specific local heuristics. This biological ant's characteristics are used to select the feature subset(s) form the whole feature set (D). The main problem with the feature subset selection is a representation of the graph because the graph needed to identify the shortest path from one feature to another feature. In the graph, each feature considered as the node and the edges between the feature is next choice feature and the optimal feature set is selected by calculating the minimum number of node visits in the graph and that should satisfy the searching criteria. After representing the graph the transition and the pheromone update rules have to be generated because each feature having own pheromone and heuristic values [24]. These heuristic value is used to analyze the optimal solution from the sequence of finite set of solution. Initially the optimal solution is identified by empty partial solution and then the solution set has enhanced further solution identification process. During this process, probabilistic transition rule is updated as follows,

$$P_i^k(t) = \begin{cases} \frac{|t_i(t)|^{\alpha}|n_j|^{\beta}}{\sum_{\mu}|t_i(t)|^{\alpha}|n_j|^{\beta}} & if\ i\ \epsilon j^k \end{cases} \quad (1)$$

Where $j^k$is the set of feasible features t and n is the pheromone values and α, βis the heuristic information. This calculated transition probability is used by ACO which is helping to balance between the pheromone intensity measure. Then the pheromone evaporates is controlled by applying the pheromone update rule which is performed as follows,

$$\Delta t_i^k = \begin{cases} \phi.\gamma\left(s^k(t)\right) + \frac{\phi.(n-|s^k(t)|)}{n} & if\ i\ \epsilon\ s^k(t) \\ 0 \quad otherwise & \end{cases} \quad (2)$$

Where $s^k(t)$is the selected feature subset

This process is repeated to find the optimal solution from the group feature set and the related steps are defined as follows,

Steps for Ant Colony Optimization
Step 1: Initialize the pheromone trails and parameters
Step 2: Generate the population of m solutions
Step3: For each feature (ant) calculate the transition probability and pheromone value
Step 4: For each ant determine its best position based on the values
Step 5: Determine the best global feature (ant)
Step 6: Update the pheromone trail values.
Step 7: Check the termination condition (termination=true)

4.2 Rough Set Theory

Rough Set Theory (RST) is one of the mathematical model which is used to dealing with the uncertainty of the data and also reduce the available information in terms of feature selection process [26]. The main purpose of the RST is to reduce the attributes and generating the decision rules that consist of several advantages such as establishing efficient algorithm for analyzing the hidden patterns, evaluating significant data, finding minimum data set and it does not require the membership function during the data set evaluation. These RST advantages are used to identify the optimal feature set from the collection of data set in terms of the indiscernibility relationship between lower and upper approximation data [27]. The relation can be defined as follows,

$$IND(P) = \{\{(x,y)U^2 | a\ P, a(x) = a(y)\}\} \quad (3)$$

Where *x,y* is the indiscernibility value of P, P is the relationship between the lower and upper approximation. From the calculated value the degree of the attribute estimation is used to identify the optimal features which are defined as follows,

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \quad (4)$$

Where POS_p is the positive boundary region |U| denotes the cardinality of set U features, which is used during the feature selection process. In this proposed paper both ACO and RST methods are used to reduce the big protein data set which reduces the dimensionality of the data and remove noisy and unwanted data that improves the further protein sequence analyze processing.

4.3 Proposed Ant Colony with Rough Set Theory (ACRST) for Feature Selection – Implementation Process

In this paper, huge volume of the protein data set is processed by several approaches, but the major challenges in the protein data set are high dimensionality of the data. This problem is overcome by Ant Colony with Rough Set Theory (ACRST) based feature selection method which reduces the dimensionality of the features as well as removes the unwanted data set. The collection of protein data set is initialized and the features (ant) are placed in the graph which is used to construct the optimized path. From the initialized feature position the transition probability of the data is calculated by applying the equation 1 which is performed until to reach the stopping condition. If the features have not satisfied the transition probability condition, then the heuristic value is calculated and the updates those values by using the equation 2. The selected features and the related pheromone values are gathered which applies to the Rough Set Theory approach to identifying and choosing the optimized feature. In the RST, the positive and negative region boundary value is evaluated and those values are used to calculate the degree of the features by using the equation 3 and 4. From the degree of the value, the decision rule is generated which means, whether the selected features are optimized or not based on the condition the optimized features are selected. Then the proposed algorithm steps are explained as follows,

Proposed Feature Selection Steps
Step 1: Initialize the population (feature data set)
Step 2: Calculate the transition probability of the features
Step 3: Repeat the step 2 until to reach the stop condition (probability=true)
Step 4: If not satisfy the condition then calculate the heuristic and pheromone values
Step 5: Pass the selected values to the RST and select the POS region value
Step 6: Estimate the value of the degree.
Step 7: Generate the decision rule to select the optimized features
Step 8: Continue this step to reach the stop condition
Step 9: Update all the values to repeat the above process.

This process forms the clusters with an optimized feature set that is used to further classification process. In this approach, ACO method reduces the dimensionality of the feature and RST removes the noise and missing data from the protein data set.The performance of the proposed system is explained in the following section.

5. Experimental Result and Discussions

The Protein Sequence and structure is plays an important role in several living organisms applications. During the protein sequence analyzing process huge volume of information is the major problem which is overcome by applying the Ant Colony with Rough Set Theory based feature selection approach. Then the proposed system demonstration is performed on the following data sets such as VariBench, BioGrid and PDB data set which is explained as follows,

5.1 Dataset Varities

5.1.1 VariBench Data set

VariBench Data set is one of the larges Biological data set which consists of four important protein details such as protein tolerance, protein stability, splice sites and the transcription for binding sites. In addition the varibench dataset provides the residue and residue mapping in the position of RNA,DNA sequence [27]. Thus the varibench data set has both testing and training benchmark data sets which provide following data sets, MMR missense variants, DBASS3&5, protein disorder, protein solubility and cancer variation dataset.

5.1.2 BioGrid Data set

Biological General Repository for Interaction Datasets (BioGRID) is one of the publicly available data set which consists of 830,000 genetic and protein interaction data from model organisms [28]. This data set focus on the binary protein-protein and genetic interaction which is used to analyze the different genetic actions in the living organisms.

5.1.3 PDB Data set

Protein Data Bank (PDB) has the information about the 3D protein shapes, nucleic acid, complex assemblies which are used to understand the importance of the proteins in terms of health and disease [29]. This data set is used in several applications in molecular biology, structural biology, computational biology and beyond.

5.2 Discussions

The performance of the proposed Ant Colony with Rough Set Theory (ACRST) feature selection approach is with the help of accuracy, sensitivity and specificity,mean square error and so on. The feature selection process improves the above protein sequence analyzing and function prediction process which is explained in the following results which was compared with the existing methods such as wrapper method [30], Greedy Forward Selection [31], Particle Swarm Optimization [32], Scatter Search[33] and Quick Reduct [34]. The following figure 1 shows that how the proposed feature selection approach selects the feature from the different data set.

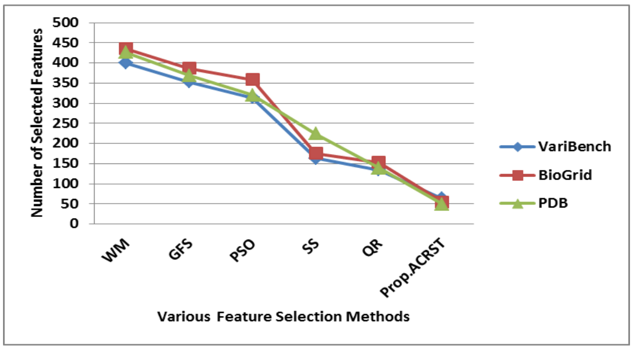

Fig.4: Performance of Various Feature Selection Methods

The above figure 4 clearly explains that the proposed ACRST feature selection method selects the optimized and minimized feature set with three different dataset than compared to the other selection methods. Then the feature selection methods Mean square error analyzes is shown in following figure 5 that helps to identify how the proposed system exactly selects the optimized features from the huge amount of feature set.

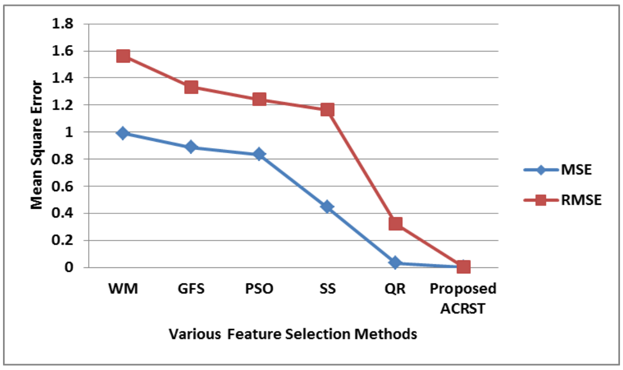

Fig.5: Error Value of the Various Feature Selection Methods

The above figure 5 explains that the proposed selection method has the minimum mean square error and the root mean square error value which means it correctly identify and select the optimized features from the set of population and the sensitivity, specificity [35] of the proposed system is evaluated as follows,

Sensitivity=TP/((TP+FN)) (5)

Specificity=TN/((TN+FP)) (6)

Where, TP = True Positive, TN = True Negative
FP = False Positive, FN = False Negative.

The following Figure 6 and 7 shows that the Sensitivity and Specificity value of the proposed system which is compared to the several classification methods such as wrapper method, Greedy Forward Selection, Particle Swarm Optimization, Scatter Search and Quick Reduct.

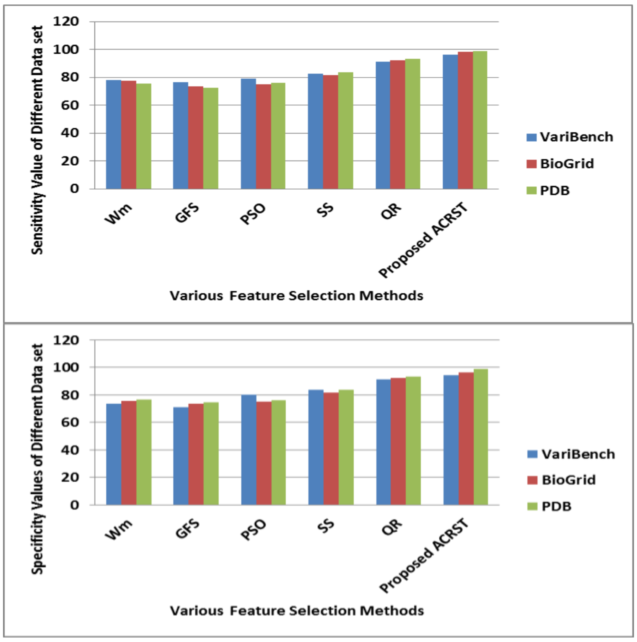

Fig. 6, 7: Sensitivity Value of the VariousFeature Selection Methods

From the figure, it is easy to justify that the proposed system produces the best features which is described by using the sensitivity and specificity. So, that selection accuracy of the proposed system in shown in following Table 1.

Table 1: Selection Accuracy for Different Feature Selection Techniques

S.No	Classification Techniques	Classification Accuracy (%)
1	Wrapper Method	71
2	Greedy Forward Selectoion	85
3	Particle Swarm Optimization	75.6
4	Scatter Search	87.3
5	Quick Reduct	93.21
6	Proposed ACRST	98.73

Thus the proposed ACRSTselection method selects the optimized protein data from the various data set such as VariBench, BioGrid and PDB data set. Then the proposed system produces the optimized features which is justified by using the mean square error, sensitivity, specificity and accuracy measures.

6. Conclusion

Bioinformatics is one of the interesting area which is used to analyze the protein sequence, protein functions and so on. Thus the protein functionality has been identified by using the large volume of dataset, but the main problem is the high dimensionality of the data which lead to reduce the performance of further processing. Then the paper proposed that Ant Colony with Rough Set Theory (ACRST) based feature selection method in terms of using the transition probability, heuristic values, pheromone value and the degree of features. The feature selection method reduces the dimensionality of the dataset also minimize the error rate and the selected features are clustered. The clustered features are preprocessed for improving the rest of analyzing process. Thus the performance of proposed selection method is evaluated with the help of the experimental results.

7. Future Enhancements

- Ant colony optimization algorithm for the problem of partitioning graphs with supply and demand could be processed in the subset of protein data with enhancemernts.
- Very effective method can be used to find the optimal solutions in more that 50% of the test instances in th future.

- Average relative error of less than 0.5% when compared to known optimal solutions will be analyzed with the help of using general graphs, Halin graphs, series–parallel graphs and trees techniques.

References

[1] Dr.S.Vijayaraniand Ms. S.Deepa, “Protein Sequence Classification InData Mining– A Study”, International Journal of Information Technology, Modeling and Computing (IJITMC) Vol. 2, No.2, May 2014.

[2] Suprativ Saha and Rituparna Chaki, “Application Of Data Mining In ProteinSequence Classification”, International Journal of Database Management Systems (IJDMS) Vol.4, No.5, October 2012.

[3] Ananya Basu and Suprativ Saha, “Delineation Of Techniques ToImplement On The Enhanced ProposedModel Using Data Mining For ProteinSequence Classification”International Journal of Database Management Systems (IJDMS) Vol.6, No.1, February 2014.

[4] Gaurav Pandey, Vipin Kumar and Michael Steinbach, “Computational Approaches for Protein FunctionPrediction: A Survey”, available at., https://www.dtc.umn.edu/ publications/ reports/2007_04.pdf.

[5] Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya, “Big Data Analytics in Bioinformatics: A MachineLearning Perspective”, Journal Of Latex Class Files, Vol. 13, No. 9, September 2014.

[6] Kara Dolinskia, and Olga G. Troyanskayaa, “Implications of Big Data for cell biology”, Journal of Molecular Biology of the Cell, Volume 26, Issue 14, 2015.

[7] Jia-Feng Yu,Yue-Dong Yang, Xiao Sun, and Ji-Hua Wang, “Sequence and Structure Analysis of Biological Molecules Based on Computational Methods”, BioMed Research International, Volume 2015, 2015.

[8] M.Bagyamathi, H.Hannah Inbarani, “A Novel Hybridized Rough Set and Improved Harmony Search Based Feature Selection for Protein Sequence Classification” Big Data in Complex System, Volume 9, pp 173-204, 2015.

[9] Muhammad Javed Iqbal, Ibrahima Faye, Brahim Belhaouari Samir, and Abas Md Said, “Efficient Feature Selection and Classification of Protein Sequence Data in Bioinformatics”, The Scientific World JournalVolume 2014 (2014), Article ID 173869, 12 pageshttp://dx.doi.org/10.1155/2014/173869.

[10] Wen-Yun Yang , Bao-Liang Lu , and Yang Yang, “A Comparative Study on Feature Extraction fromProtein Sequences for Subcellular Localization Prediction”, at., http://www.cs.ucr.edu/~yangyang/Site/Welcome_files/CIBCB06-YwYang.pdf

[11] Kocbek, Stiglic, Pernek, Kokol, “Stability of different feature selection methods for selecting protein sequence descriptors in protein solubility classification problem”, IEEE 23rd International Symposium onComputer-Based Medical Systems (CBMS), 2010.

[12] Eun-Mi Kim, Jong-Cheol Jeong, Ho-Young Pae, Bae-Ho Lee, “A New Feature Selection Method for Improving the Precision of Diagnosing Abnormal Protein Sequences by Support Vector Machine and Vectorization Method”, Adaptive and Natural Computing Algorithms, Volume 4432, 2007.

[13] Ben Blum, Michael Jordan, David E. Kim, Rhiju Das, Philip Bradley, David Baker, “Feature Selection Methods for Improving ProteinStructure Prediction with Rosetta”, at., http://www.cs.berkeley.edu/~jordan/papers/blum-etal-nips07.pdf

[14] Bing Xue, Mengjie Zhang, and Will N. Browne, “Particle Swarm Optimization for Feature Selection inClassification: A Multi-Objective Approach, “Ieee Transactions On Cybernetics, 2012.

[15] Ahmed T. Sadiq Al-Obaidi, “Improved Scatter Search Using Cuckoo Search”, International Journal of Advanced Research in Artificial Intelligence,Vol. 2, No. 2, 2013

16] Anitha , Dr.P.Venkatesan, “Feature Selection By Rough –Quick Reduct Algorithm”, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 2, Issue 8, Aug. 2013.

[17] Aneeshkumar A.S and Jothi Venkateswaran C, “A novel approach for Liver disorder Classification using Data Mining Techniques”, Engineering and Scientific International Journal, Vol. 2, Issue 1, March 2015, pp.15-18,

[18] Durairaj, Sivagowry, “Feature Diminution by using Particle Swarm Optimization for Envisaging the Heart Syndrome”, International Journal of Information Technology and Computer Science, 2015, 02, 35-43.