Australian Journal of Basic and Applied Sciences, 9(20) June 2015, Pages: 518-526

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Remote-R3d: A Novel Technique to Reducing Dimensionality In Remote Homology Finding Using Predicted Protein Contact Maps

[1]A. Hepsiba and [2]Dr. R. Balasubramanian

[1]*Assistant Professor, MCA, Karpaga Vinayaga College of Engineering &Technology, Madhuranthagam. India*
[1]*Research Scholar, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India*
[2]*Dean, Karpaga Vinayaga College of Engineering &Technology, Madhuranthagam, Tamilnadu, India*

ARTICLE INFO

Article history:
Received 22 May 2015
Accepted 12 July 2015
Available online 18 July 2015

Keywords:
Remote R3-D, Protein Contact Map, Data Mining, MATLAB

ABSTRACT

A long standing problem in structural bioinformatics is to determine the three-dimensional (3-D) structure of a protein when only a sequence of amino acid residues is given. Many computational methodologies and algorithms have been proposed as a solution to the 3-D Protein Structure Prediction (3-D-PSP) problem. Most discriminative methods concatenate the values extracted from physicochemical properties to build a model that separates homolog and non homolog examples. Each discriminative method uses a specific strategy to represent the information extracted from the protein sequence and a different number of indices. After the vector representation is achieved, Support Vector Machines(SVM) are usually used. Most classification techniques are not suitable in remote homology detection because they do not address high dimensional datasets. In this work, proposed a method that reduces the high dimensionality of the vector representation using models that are defined at the 3D level. Next, the models are mapped from the protein primary sequence. This proposed method is called remote-R3D, is presented and tested on the ASTRALSCOP1.53 and ASTRALSCOP1.55 datasets. The Remote-R3D method achieves a higher accuracy than the composition based methods and a comparable performance with profile-based methods. The Proposed Classifier of protein classification shows significant improvement in terms of performance measure metrics: accuracy, sensitivity, specificity, recall, F-measure, and so forth implemented in MATLAB.

To Cite This Article: A. Hepsiba and Dr. R. Balasubramanian, Remote-R3d: A Novel Technique to Reducing Dimensionality In Remote Homology Finding Using Predicted Protein Contact Maps. ***Aust. J. Basic & Appl. Sci.,*** *9(20): 518-526, 2015*

INTRODUCTION

Proteins play a vital role in our life because they perform important tasks such as catalysis of biochemical reactions, transport of nutrients, and transmission of signals. The importance of proteins in our life drives biologists to discover more proteins and study their biological functions. The 3D structure of a protein can be used as a good indicator of its function. The determination of 3D structure by biological methods such as X-ray and NMR is very cumbersome and costly. Despite the improvement in experimental procedures to determine protein structure, the gap between the number of known protein sequences and their structures continues to increase. Therefore, developing new machine learning approaches or improving current approaches to predict protein structure can decrease this gap. Earlier studies indicate that developing an accurate protein contact map predictor will be very helpful in the reconstruction of protein 3D structure. Accordingly, much research is implemented in this problem due to the current low accuracy.

On Proteins, Structure And Representation:

From a structural perspective, a protein is an ordered linear chain of building blocks known as amino acid residues. Each protein is defined by its unique sequence of amino acids. This sequence causes the protein to fold into a particular three-dimensional shape. Predicting the folded structure of a protein only from its amino acid sequence remains a challenging problem in mathematical optimization (Lander and Waterman,1999). The challenge arises due to the combinatorial explosion of plausible shapes each of which represent a local minimum of an intricate non-convex function of which the global minimum is sought. In nature, proteins typically present 50 to 500 amino acid residues. The books by Lesk (Lesk, 2002) and Tramontano (Tramontano, 2006) present elegant, comprehensive overviews of protein structure.

Corresponding Author: A. Hepsiba, Research Scholar, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India

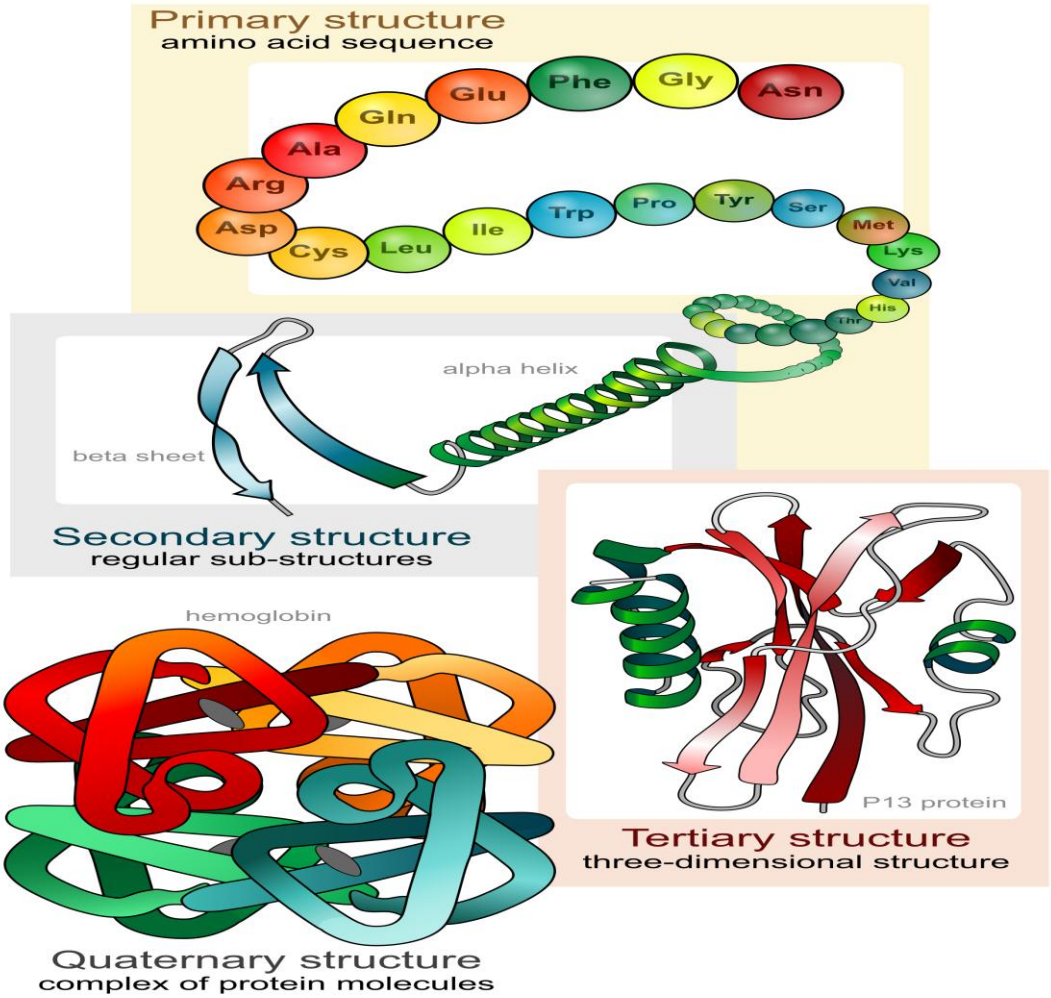

Fig. 1: The protein structure hierarchy.

In nature there are 20 distinct proteinogenic amino acids, each one with its own chemical properties (including size, charge, polarity, hydrophobicity, i.e. the tendency to avoid water packing) (Lodish *et al*., 1990; Lehninger *et al*., 2005). depending on the polarity of the side-chain, amino acids vary in their hydrophilic or hydrophobic character. The importance of the physical properties of the side-chains comes from the influence they have on the amino acid residues interactions in the 3-D structure. The distribution of the hydrophilic and hydrophobic amino acids are important to determine the tertiary structure of the polypeptide. A detailed description of the amino acid properties can be found in Lehninger (Lehninger *et al*., 2005) and Lodish (Lodish *et al*., 1990).

A peptide is a molecule composed of two or more amino acid residues chained by a chemical bond called the peptide bond. This peptide bond is formed when the carboxyl group of one residue reacts with the amino group of the other residue, thereby releasing a water molecule(H2O). Two or more linked amino acid residues are referred to as a peptide, and larger peptides are generally referred to as polypeptides or proteins (Creighton, 1990; Lesk, 2002). The peptide bond (C-N) has a double bond and is not allowed rotation of the molecule around this bond. The rotation is only permitted around the bonds N-Cα and Cα-Cα These bonds are known as PHI and PSI angles, respectively, and are free to rotate.

Literature review:

Jong Park and Dan Bolser established a bioinformatics research group in UK named MRC-DUNN. They stated their research on protein network. They worked on structure of proteins. They also used PSIMAP concept. But the limitation is that they only focused on protein intractability and taxonomic diversity. As a result their concept did not help that much on protein structure analysis using PSIMAP concept.

Wan K. Kim, Dan M. Bolser and Jong H. Park had used PSIMAP for large-scale coevolution analysis of protein structural interlogues. They investigated the degree of co-evolution for more than 900 family pairs in a global protein structure interact map. They have constructed PSIMAP by systematic extraction of all protein domain contacts in the web based Protein Data Bank. Their PSIMAP contained

37387 interacting domain pairs with five or more contacts within 5A. They have first confirmed that correlated evolution is observed extensively throughout the interacting pairs of structural families in PDB, indicating that the observation is a general property of protein evolution. The overall average correlation was 0.73 for a relatively reliable set of 454 family pairs, of which 78% showed significant correlation at 99% confidence. In total, 918 family pairs have been investigated and the correlation was 0.61 on average. But the statistical validity was weak for the family pairs with small N (the number of member domain pairs) of their research. This is the first step in protein classification technique two combine two properties of proteins, namely, structure comparison and interactivity.

Daeui Park, Semin Lee, Dan Bolser, Michael Schroeder some other scientists at beginning of 2005 have developed Comparative interactomics analysis of protein family interaction networks using PSIMAP (protein structural interactome map) They have confirmed that all the predicted protein family interactomes (the full set of protein family interactions within a proteome) of 146 species are scale-free networks, and they share a small core network comprising 36 protein families related to indispensable cellular functions. To construct the protein family interaction network in a particular proteome, they first assigned the known 3D structural families (on which PSIMAP is based) to the protein sequences. 146 completely sequenced species from the European Bioinformatics Institute (EBI) and their 578,625 protein sequences were used (Pruess, *et al*., 2003).

RESULTS AND DISCUSSIONS

Evaluation Criteria:

To evaluate the performance of our approach, the following criteria are chosen in the experiments, which are the accuracy (ACC), sensitivity (SN), precision (PE) and Matthews correlation coefficient (MCC), written as:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Corralation\ Coefficent = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}}$$

where TP, TN, FP and FN denote true positive, true negative, false positive and false negative, respectively. In addition, the receiver operating characteristic (ROC) curve is also adopted to evaluate the prediction performance. The ROC curve plots the true positive rate (TPR) versus the false positive rate (FPR) with the threshold varying. The area under the ROC curve is called the area under the ROC curve (AUC), which falls into (0,1). The larger the AUC, the better prediction performance we can achieve.

Datasets:

In order to produce reliable results, proteins used in a trusted benchmarking set are used in this work. Evaluation of automatic protein structure prediction servers (EVA) supports researchers with huge dataset of proteins under specific criteria. In this work, the dataset was downloaded on June 2, 2015 from the EVA servers. Mainly, no pair in a subset has more than 34% identical residues over more than 100 aligned residues. In addition, the preference is given to high-resolution structures. EVA lists on its servers a huge database of protein chains and their PDB files, FASTA files, PSI-Blast files, and other related useful files formats. A filtering process is applied to eliminate unwanted protein PDB files that have unusable data. These files may give misleading results during training stage. This process was adopted in many previous works. It starts by removing the corrupted PDB files, which may contain erroneous or incomplete data. Then short and long proteins are removed. Short sequences with less than 30 amino acids are removed because most likely they do not have actual structure and may disrupt the system during the training stage, which may result in unreliable outputs for the testing.

Selecting the dataset:

The ASTRALSCOP database 1.53 has been used as a standard for remote homology detection. A total of 5244 sequences are selected by removing similar sequences using an E-value threshold of 10_25. The data are further filtered for classification purposes keeping only ASTRALSCOP families having at least five positive sequences for testing and 10 positives sequences for training, which gives a total of 52 families. Most of the remote homology detection methods use the ASTRALSCOP1.53 dataset, so it is especially easy to perform a comparison analysis. However, PDB-style files with coordinates for each protein are not available for the ASTRALSCOP1.53 dataset. In this work, the PDB files are necessary to calculate the contact maps when the 3D structural models are found, and thus, a ASTRALSCOP database based on stable ASTRALSCOP identifiers with coordinates for each protein is mandatory. Here decided to use both the ASTRALSCOP1.53 and ASTRALSCOP1.55 datasets. The ASTRALSCOP1.55 dataset is a more recent dataset than the ASTRALSCOP1.53 and the PDB files are available. The ASTRALSCOP1.55 dataset is first used to obtain the models and calculate the accuracy of the remote-R3D method. Then, the same 3D models are used to detect remote homologs in the ASTRALSCOP1.53 dataset and to compare the Remote-R3D with some discriminative

methods. The ASTRALSCOP1.55 dataset was first filtered by selecting sequences with less than 42% of identity to each other. In addition, for classification purposes, only ASTRALSCOP families having at least five positive sequences for testing and 10 positives sequences for training were kept, which gave a total of 5244 domains in 52 families. Table1 shows the ASTRALSCOP1.55 database after the filtering process.

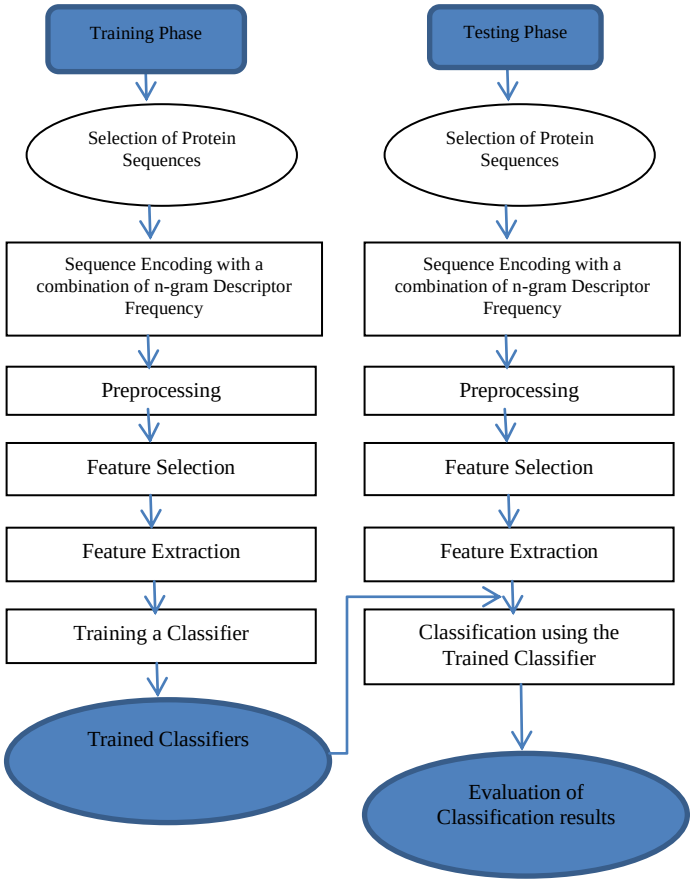

Fig. 2: Flow Diagram of REMOTE-R3D.

Methods:

In this section, we explain in detail every step in the remote-R3D method. The remote-R3D method includes obtaining the 3D models, predicting a contact map from the primary sequence, calculating the count vector, and building a classifier for each ASTRALSCOP family. The general overview of the remote-R3D method.

Obtaining the 3D models:

The first step in the remote-R3D method is obtaining 3D models from the contact map representation of protein in a given dataset. A distance matrix of a protein is a square matrix containing the Euclidean distances between all pairs of Cα atoms in the protein. A contact map is achieved by discretizing the distance matrix. Different thresholds can be used to determine when two residues are in contact. In this research, we assume that two residues are in contact when the Euclidean distance between the corresponding Cα atoms is less than or equal to 8.0Angstroms. A cut off distance of 8.0Angstroms has been considered a standard threshold for most of the current contact maps prediction programs. Although contact maps of any two proteins are different, there are specific are as in the contact maps that can be recognized as patterns even in different domains. In fact, Choietal. found that a set of 100 representative 10X10 sub matrices extracted from the distance matrices are able to discriminate domains in the protein structure comparison problem, which is related to classifying a protein in to the correct ASTRALSCOP classification at the class, fold, super family, and family levels. Choietal. established that the 100 representative sub matrices or models reflecting local structural features, and combinations of these models can be used to reconstruct the original distance matrices. In addition, Choietal. demonstrated that even though there are millions of different sub matrices in all proteins, most of the sub matrices are common, and thus, a finite number of models can be sufficient to represent observed interactions in the distance matrix. If the 3D information is available then solving there mote homology detection problem is very easy; otherwise it becomes a complicated

task. Therefore, the remote-R3D method predicts contact maps from the primary sequence alone. Even though remote homology detection is performed using predicted contact maps, the collection of models are obtained from actual contact maps. A collection of 3D models (i.e.,common mXm submatrices in the contact maps; where m goes from 4 to 12) are used in the remote-R3D method. The 3D models are extracted from a training dataset of 40 proteins selected from the ASTRALSCOP1.55 dataset by taking two proteins randomly selected for each of the 20 ASTRALSCOP super families in the dataset. Distance matrices, and thus, actual contact maps are available for the ASTRALSCOP1.55 dataset. The above section shows how the selection of the proteins affects the clustering process. Obtaining the 3D models starts by calculating a contact map for each protein in the training data set used specifically for the clustering process. Then, discretized submatrices of mXm are extracted as local structural features. The size of the sub matrix determines the size of the models that are used. A large size of the submatrix (i.e.,12X12 submatrix) captures a large portion of a 3D interaction in the contact map. On the other hand, a small size of the submatrix (i.e.,4X4 submatrix) might not been ought to capture are preventative interaction. We propose to use different sizes (i.e.,4X4, 6X6, 8X8, 10X10,12X12) to discover the impact of the size of the submatrix in the remote-R3Dmethod. For the obtained collection of discretized submatrices, the clustering algorithm CLARA (Clustering Large Applications)is used to obtain representative discretized submatrices (i.e.,resulting methods after clustering algorithm) that are taken as 3D models in the remote R3D method. The CLARA algorithm is used because it can work with large datasets.

Traditional clustering algorithms cannot be applied in our dataset. In addition, the CLARA algorithm returns medoids, which means that each model is a typical structure that actually exists in the dataset. It is expected that these discrete models are common and frequently used in contact maps. Following the methodology proposed by Choietal., each protein in training is first clustered into 50 representative submatrices, and then the obtained medoids are clustered again to obtain a final set of k models. This allows having a reasonable amount of submatrices in the clustering process. Otherwise, there would be millions of submatrices to be clustered, affecting the performance of the machine learning algorithm. Here in this paper, attempted different sizes of the submatrices (i.e., 4X4, 6X6, 8X8, 10X10, 12X12) and different number of models (i.e.,10,20,30,40,50). The collection of models M10 10X10 (i.e., 10 discretized 10X10 submatrices). Contacts are represented in black and non-contacts in white. Every 3D model reflects a local structural interaction. m1 represents the helix contacts in the main diagonal; m2 represents the end of a parallel beta-sheet; m3 represents the extreme outer part(right)of the main diagonal; m4 represents then on-contacts between residues and is the most frequent model in any protein; m5 represents the inner part of the main diagonal; m6 represents the diagonal(non-helix); m7 represents the outer part(left)of the main diagonal; m8 represents the extreme outer part (left) of the main diagonal; m9 represents the anti-parallel beta sheets; m10 represents the outer part(right) of the main diagonal. The models obtained in the collection M20 10X10 (i.e., 20 discretized 10X10 submatrices). As observed, more detailed models are obtained for each local structural interaction. For instance, models m13, m16, m17, and m18 in the collection M20 10X10 reflect different positions for an anti-parallel beta sheet, whereas only the model m9 was obtained in the collection M10 10X10 to represent the same local structural interaction.

Predicting the contact map from the primary sequence:

Then extstepintheremote-R3D method is to predict the contact map from the primary sequence of every protein in the ASTRALSCOP1.55 dataset. Here it is used the NNcon1.0 program [24] to predict contact maps from the primary sequence. NNcon1.0 is a program that predicts contact maps using artificial neural networks. NNcon1.0 requires the amino acid sequence and returns the predicted contact map at two different threshold so f8.0 and 12.0 Angstroms.NNcon1.0 uses two 2D-Recursive Neural Networks, one to predict a general residue contact map and another to predict the special beta-sheet residue pairing map.NNcon1.0 focuses on the problem of predicting beta-sheet residue contacts, which has been demonstrated to be a difficult task. The output of the NNcon1.0 program includes a predicted probability for each pair of residues to be in contact. Here in this work, attempted two different predicted probabilities in NNcon1.0 (i.e.,0.1 and 0.2). Even though the predictions in the NNcon1.0 program are faster than in other programs, there is an overhead when the contact maps are predicted for every protein in the dataset. However, representing a protein as a predicted contact map is a key aspect when the dimensionality of the remote homology detection problem is being reduced. The predicted 3D information can be used in a natural way to detect proteins that are functionally related. Contact maps have been used in some other problems in Bioinformatics such as protein structure classification. For instance, Suvarnaetal extract rules from contact maps to represent all-alpha structural classes in the ASTRALSCOP database. However, predicted contact maps have never been used in remote homology detection. In this paper, present a first attempt to use predicted contact maps as the protein representation in the remote homology detection problem. Even though the accuracy of the

contact maps prediction programs is still low(i.e.,55% in the NNcon1.0 program), want to discover if the information that is predicted is sufficient to determine the structure of the protein, and thus, allows distinguishing remote homologs.

Protein contact map prediction using decision tree algorithm:

Application of basic statistical methods is used to study the Protein Contact Map. Various Protein Contact Map are provided to carry out a more detailed and separate the attributes which have the maximum effect, judgment trees are used.

Decision trees are a highly flexible modeling technique. For instance, to build regression models and neural network models, the missing values have to be inserted into training data while decision trees can be built even with missing values. Decision trees are intended for the classification of attributes regarding the given target variable (Panian and Klepac, 2003). Decision trees are attractive as they offer, in judgment to neural linkages, data models in readable, comprehensible form – in fact, in the form of rules. They are used not only for classification, but also for the prediction (Gamberger and Šmuc, 2001).

Logistic regression is an analysis of asymmetric relations between two variable sets of which one has the predictor status and the other criterion status (Halmi, 2003). The dependent variable is dichotomous and marked by values 0 and 1, while the independent variables in logistic regression may be categorical or continuous (Hair, Anderson and Babin, 2009).

Decision Tree Algorithm:

Application of basic statistical methods is used to study the Protein Contact Map. Various Protein Contact Map are provided to carry out a more detailed and separate the attributes which have the highest effect, decision trees are used. A Adaboost classifier having the form $H(x) = \text{sign} \sum_t a t h t(x)$ can be trained minimizing a minimizing a loss function L; i.e. by scalar rating scalar at and we a (learner ht(x) at each iteration t.

Step 1: compute classification entropy.
Step 2: for each attribute, calculate information gain using classification attribute.
Step 3: select attributes with highest information gain.
Step 4: remove node attribute, for future calculation.
Step 5: repeat steps 2-4 until all attribute has been used.

Proposed Modified Boost Decision Tree:

Step 1. Initialize weights (sorted in decreasing order).
Step 2. Train decision tree ht (one node at a time) using the Quick Stump Training method.
Step 3. Perform standard Boosting steps:
(a) Determine optimal at (i.e. using line-search).
(b) Update sample weights given the misclassification error of ht and the variant of Boosting used.
(c) If more Boosting iterations are needed, sort sample weights in decreasing order, increment iteration number t, and go to step 2.

Discussions:

Since precisely predicting the function or structure of a novel protein from a protein sequence is a significant problem in machine learning and bioinformatics research community, the exact knowledge about the structure and function of proteins provides a way to analyze and model protein sequences and is also helpful in the treatment of numerous diseases. This will also be useful in the design and detection of new drugs for many diseases. In this study, nine functionally important super families of protein sequences were taken into account; these are very essential to perform various critical functions. The variable-length protein sequences were chosen randomly from the benchmark ASTRALSCOP database. The protein sequences in the chosen superfamilies are relatively long and have very low sequence similarity among them; therefore, it was very difficult to classify them into existing superfamilies with high accuracy using previous approaches. We have utilized a combination of α-gram protein descriptor's frequency-based features encoding to represent a protein in the form of a fixed-length feature vector. The statistical metric was employed for the selection of informative features which significantly reduced the size of the feature vector. The classification results using popular classification algorithms on each dataset have been shown in the next section. Figures 3, 4, and 5 showed the graphical representation of the performance measure metrics obtained on each dataset using different classification algorithms. In all of the graphs, the x-axis contains the performance measure metrics that we have measured in the experiments and on the y-axis; the values of the metrics obtained with each classifier are shown. The analysis of the graphs in Figures 3, 4, and 5 indicates that the neural network classifier on all three datasets shows improved classification accuracy, specificity, sensitivity, precision, recall, F-measure, and MCC. The trends of the improvements found in the performance measure metrics were very similar on the three datasets. The architecture of neural network used in our experiments was comprised of default parameters with 10 hidden neurons and 3 output layers.

Table 1: Performance Analysis of Various Classifiers.

Classifier	Accuracy(%)	True Positive Rate	False Positive Rate	Precision(%)	Recall(%)	Classification Error (%)	Kappa Statistics	RMS Error
Decision Tree	50.68	0.507	0.230	0.478	0.507	49.32	0.211	0.404
Random Forest	63.34	0.633	0.254	0.570	0.633	36.66	0.354	0.313
J48	64.45	0.500	0.544	0.521	0.500	35.55	0.344	0.310
PRISM	63.45	0.750	0.350	0.825	0.750	36.55	0.635	0.718
IBK	54.50	0.871	0.594	0.571	0.871	45.50	0.484	0.539
Naïve Bayes	53.75	0.571	0.594	0.528	0.571	46.25	0.484	0.539
Simple Cart	60.16	0.600	0.490	0.682	0.600	39.84	0.470	0.654
Multi Layer Perception	61.58	0.546	0.500	0.855	0.546	38.42	0.495	0.356
Proposed	68.80	0.650	0.545	0.420	0.650	31.20	0.301	0.212

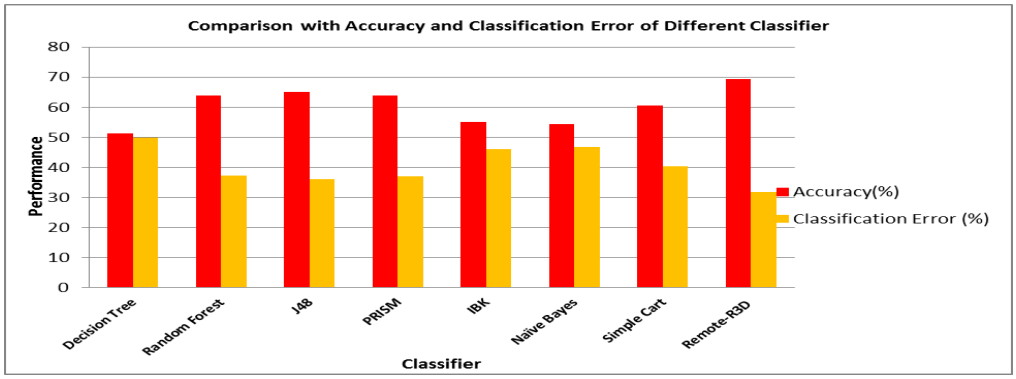

Fig. 3: Comparison of the performance measure metrics using dataset 1.

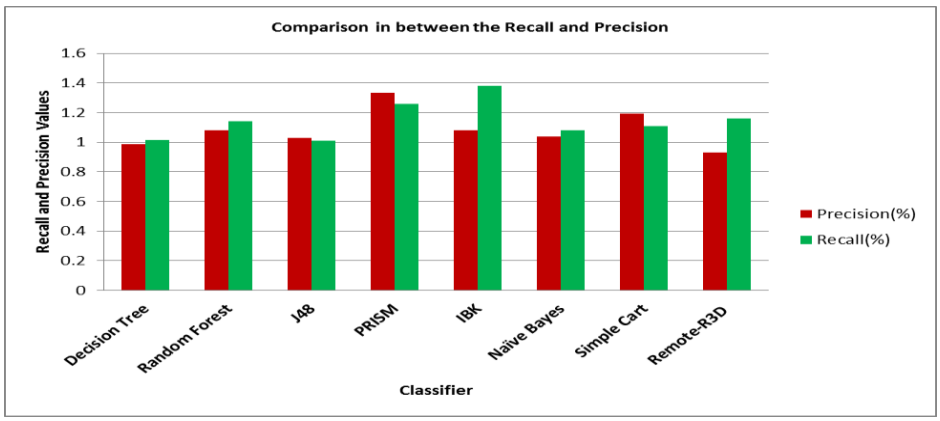

Fig. 4: Comparison of the performance classifier in terms of Recall and precision.

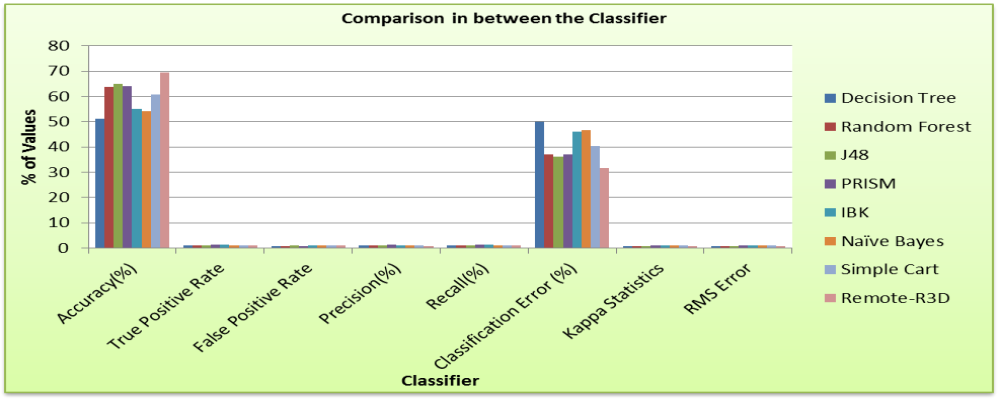

Fig. 5: Comparison of the performance classifier.

Figures 3, 4, and 5 show the comparison of existing and proposed classification accuracy results obtained on different datasets using the proposed sequence encoding and feature subset selection techniques. Without using the feature subset selection technique, the feature size would be bigger and this would ultimately decrease the classification accuracy and more computational cost would be required. Figure 5 shows the comparison of the accuracy of the proposed method with the previously available classification methods.

Conclusion:

The proposed feature subset selection technique uses a threshold to select the highly informative and important features. The results of the technique were validated through the well-recognized classification/learning algorithms. The protein sequences of three different datasets have been effectively classified into relevant super families with substantially high classification accuracy. The introduced classification method is alignment-free, simple, fast, and reliable. This technique of feature selection and classification would be useful in machine learning and bioinformatics in reducing the high dimensionality of data during the prediction of the structure or function of unknown protein sequences. In the future, the proposed technique can be extended to other areas of pattern recognition like the classification of different kinds of proteomics and genetic diseases.

REFERENCES

Best, R.B., X. Zhu, J. Shim, P. Lopes, J. Mittal, M. Feig, A. MacKerell, 2012. Optimization of the additive charmmall-atom protein force field targeting improved sampling of the backbone_, and side-chain_1 and_2 dihedral angles.J.Chem. Theory Comput, 8(9): 3257–3273.

Biasini, M., S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. Cassarino, M. Bertoni, L. Bordoli, T. Schwede, 2014. Swiss-model:modelling protein tertiary and quaternary structure using evolutionary information. Nucleic AcidsRes.,12: 252–258.

Bibby, J., R.M. Keegan, O. Mayans, M.D. Winn, D.J. Rigde n, 2012. Ample:acluster-and-truncate approach to solve the crystal structures of small proteins using rapidly computed abinitio models. Acta Crystallogr. Sect.D: Biol. Crystallogr, 68(12): 1622–1631.

Bin, Liu., Xu. Jinghao, Quan Zou, Xu. Ruifeng, Xiaolong Wang, Qingcai Chen, 2014. Using distances between Top-n-gram and residue pairs for protein remote homology detection, BMC Bio inf., 15(2): S3.

Bramucci, E., A. Paiardini, F. Bossa, S. Pascarella, 2012. Pymod: sequence similar- ity searches, multiple sequence-structure alignments, and homology modeling within pymol. BMC Bioinf., 13(4): S2–S7.

Cao, M., L.J. Cowen, 2012. Remote homology detection on alpha-structural proteins using simulated evolution.in: Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine,ACM, pp: 353–360.

Chitraranjan, C., L. Alnemer, O. Al-Azzam, S. Salem, A. Denton, M. Iqbal, S. Kianian, 2011. Frequent substring-based sequence classification with an ensemble of support vector machines trained using reduced amino acid alphabets, in: 2011 10th International Conference on Machine Learning and Applications.

Chou, K.C., 2011. "Some remarks on protein attribute prediction and pseudo amino acid composition,"Journal of Theoretical Biology, 273(1): 236–247.

de Brevern, A.G., 2005. "New assessment of a structural alphabet," In Silico Biology, 5(3): 283–289.

de Brevern, A.G., C. Etchebest and S. Hazout, 2000. "Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks," Proteins, 41(3): 271–287.

Eickholt, J., J. Cheng, 2013. A study and benchmark of DN con: a method for protein residue-residue contact prediction using deep networks, BMC Bioinf, 14(14): S12. In-Geol Choi, J. Kwon,

Fujitsuka, Y., S. Takada, Z. Luthey-Schulten, P. Wolynes, 2004. Optimizing physical energy functions for protein folding. Proteins: Struct. Funct. Gen., 54(1): 88.

Garcez, A., L. Lamb, 2006. A connectionist computational model for epistemic and temporal reasoning. Neural Comput, 18(7): 1711.

Garcez, A., L. Lamb, D. Gabbay, 2007. Connectionist modal logic: representing modalities in neural networks. Theor. Comput. Sci., 371(1–2): 34.

Ko, J., H. Park, C. Seok, 2012. Galaxytbm: template-based modeling by building a reliable core and refining unreliable local regions. BMC Bioinf, 13(1): 198–207.

Kumar, A., L. Cowen, 2010. Recognition of beta-structural motifs using hidden Markov model strained with simulated evolution, Bioinformatics, 26(12): i287–i293.

Li, W. and A. Godzik, 2006. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," Bioinformatics, 22(13): 1658–1659.

Li, Y., Y. Zhang, 2011. Atomic-level protein structure refinement using frag- ment guided molecular dynamics conformation sampling. Structure, 19(12): 1784.

Li, Z., Y. Yang, J. Zhan, L. Dai, Y. Zhou, 2013. Energy functions in de novo protein design: current challenges and future prospects. Annu. Rev. Biophys, 42(1): 315–335.

Lise, S., D. Buchan, M. Pontil, D.T. Jones, 2011.

Predictions of hot spot residues at protein–protein interfaces using support vector machines. PLoS ONE, 6(2): e16774.

Liu, B., X. Wang, Q. Chen, Q. Dong, X. Lan, 2012. Using amino acid physicochemical distance transformation for fast protein remote homology detection, PLoS One ,7(9): e46633.

Liu, B., X.Wang, Q. Chen, Q. Dong, X. Lan, 2012. Using amino acid physicochemical distance transformation for fast protein remote homology detection, PLoS One, 7(9): e46633.

Liwo, A., C. Czaplewski, D. Kleinerman, P. Blood, H. Scheraga, 2010. Implementation of molecular dynamics and its extensions with the coarse-grained unres force field on massively parallel systems; towards millisecond-scale simulations of protein structure, dynamics, and thermodynamics. J. Chem. Theory Comput, 6(3): 890.

Maciej, B., J. Michal, K. Sebastian, K. Andrzej, 2013. Cabs-fold: server for the de novo and consensus-based prediction of protein structure. Nucleic Acids Res., 41: W406–W411.

Mirny, L.A. and E.I. Shakhnovich, 1999. "Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function," Journal of Molecular Biology, 291(1): 177–196.

Moult, J., K. Fidelis, A. Kryshtafovych, A. Tramontano, 2011. Critical assessment of methods of protein structure prediction (casp) – round ix. Proteins: Struct. Funct. Bioinf, 79 (S10), 1.

Moult, J., K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, 2014. Critical assessment of methods of protein structure prediction (casp) – round x. Proteins: Struct. Funct. Bioinf, 82: 1–6.

Muda, H., P. Saad, R. Othman, 2011. Remote protein homology detection and fold recognition using two-layer support vector machine classifiers, Comput.Biol. Med., 41(1): 687–699.

Ratheesh, R.K., S.N. Nagarajan, P.A. Arunraj, 2012. "HSPIR: a manually annotated heat shock protein information resource," Bioinformatics, 28(21): 2853–2855.

Solis, A.D. and S. Rackovsky, 2000. "Optimized representations and maximal information in proteins,"Proteins, 38(2): 49–164.

SuvarnaVani, K., M. OmSwaroopa, T.D. Sravani, K. Praveen Kumar, 2014. Frequent substructures and fold classification from protein contact maps, in: 2014IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, 1(8): 21–24.

Tai, C., H. Bai, T.J. Taylor, B. Lee, 2014. Assessment of template-free modeling in casp and roll. Proteins: Struct. Funct. Bioinf, 82: 57–83.

Thomas, P.D. and K.A. Dill, 1996. "An iterative method for extracting energy-like quantities from protein structures," Proceedings of the National Academy of Sciences of the United States of America, 93(21): 11628–11633.

Webb-Robertson, B., K. Ratuiste, C. Oehmen, 2010. Physico chemical property distributions for accurate and rapid pairwise protein homology detection, BMC Bioinf, 11(1): 145–183.

Xu, D., J. Zhang, A. Roy, A. Zhang, 2011. Automated protein structure modeling in casp9 by i-tasser pipeline combined with quark-based ab initio folding and fg-md-based structure refinement. Proteins: Struct. Funct. Bioinf, 79(10): 147.