International Journal of Bio-Science and Bio-Technology
Vol.7, No.6 (2015), pp.141-150
http://dx.doi.org/10.14257/ijbsbt.2015.7.6.15

Classification of Protein Structure (RMSD <= 6A°) using Physicochemical Properties

[1]Sonal Mishra, [2]Yadunath Pathak and [3]Anamika Ahirwar

[1,2]Maharana Pratap College of Technology TechnologyGwalior - 474006, India
[3]ABV-IIITM, Gwalior - 474015, India Gwalior - 474006, India
sonalmish06@gmail.com, yadupathak86@gmail.com, aanamika77@gmail.com

Abstract

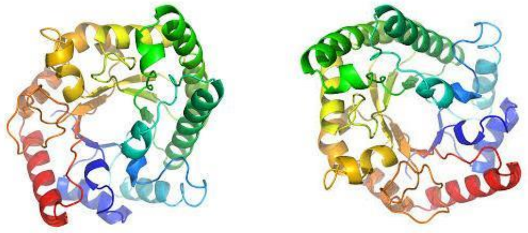
The quality of the protein structure can be determined by physical and chemical properties, therefore it has been used to distinguish native or native like structure from other predicted structures. In this study, the machine learning classification models are explored with six physical and chemical properties to classify the root mean square deviation (RMSD) of the protein structure in absence of its true native state and each protein structure lies between 0A° to 6A° RMSD space. Physical and chemical properties used in this paper are total surface area, Euclidean distance, total empirical energy, secondary structure penalty, residue length, and pair number. There are total 24294 decoys, having 4919 native structures. Artificial bee colony algorithm is used to determine the feature importance. The K-fold cross validation is used to measure the robustness of the best classification model. The results show that random forest method outperforms other machine learning models in the classification of protein structure prediction with sensitivity of 0.72 and accuracy of 70.33% on testing data set. The data set used in the study is available at http://bit.ly/RMSD-Classification-DS.

Keywords—Protein structure prediction, Machine learning, Random forest, Artificial bee colony algorithm

1. Introduction

Protein sequences are translated into 3D tertiary forms to carry out several biological functions. Prediction of high resolution protein structure has become one of the “grand challenge problems” in computational biology. Physical and chemical properties of amino acids and their solvent environment are the key determinants in folding a protein sequence into its unique tertiary structure. These factors essentially generate various types of energy contributors such as electrostatic, Vander Waals, salvation/desolvation which create folding pathways. Ab initio approaches for structure deter-mination employ these physical and chemical factors to generate a structure or an ensemble of structures from the sequence as possible candidates for the native. In the alternative approach, called homology modeling, one uses experimentally known protein structures as templates based on sequence similarity. Due to lack of a clear understanding of the true folding pathway of proteins to the native and insufficient experimental data, several prediction methods end up with low quality structures. These low quality structures may look similar to any high resolution structure passing all the quality assessment criteria but in reality they could be 10-15 A° away from their true native states (refer, Figure 1). It would be highly desirable to have a predictive model which can tell how far a structure is from the native in the absence of its experimental structure. Machine learning classification models have been widely used in protein structure prediction such as 2D and 3D structure prediction [1, 2], fold recognition [3-5], solvent accessibility prediction, disordered region prediction [6-8], binding site prediction [9], transmembrane helix prediction [10], protein domain boundary prediction [11], contact map [12-14], functional site prediction, model generation [15], and model evaluation [16, 17].

This work explores the machine learning classification models to predict native or native like structure in the absence of its true native state using six physical and chemical properties and reports how far a structure is from its true native. Total surface area, Euclidean distance, total empirical energy, secondary structure penalty, residue length, and pair number are the physical and chemical properties used for predicting the native structure. There are total 24294 decoys, having 4919 native structures. Protein sequences are taken from protein structure prediction center (CASP) and protein data bank (RCSB). The root mean square deviation (RMSD) of each structure lies between 0A° to 6A° space. Since some of the considered features may have higher importance than others in predicting the native structures, artificial bee colony (ABC) algorithm is used to determine the feature importance. The features are used by four machine learning models namely decision tree, random forest, support vector machine, and linear model for the prediction of protein structure in absence of its true native state. The K-fold cross validation is used to measure the accuracy of the best predictive model. Rest of the paper is organized as follows. A brief overview of the considered features, data set, ABC algorithm, and machine learning models are presented in Section II. The proposed protein structure prediction methodology is described in Section III. Model evaluation is presented in Section IV. Section V describes experiments, results and discussion. Finally, conclusion is presented in Section VI.



(a) Native Structure (b) Predictive Structure

Figure 1. The RMSD of Predicted Structure from its Native is 10.3 °A (PDB ID:1IF4)

2. Materials and Methods

Data set and its features: There are total 24294 modeled structures having 4919 native structures. The modeled structures are taken from protein structure prediction center (CASP-5 to CASP-10 experiments), public decoys structures database [18] and native structure from protein data bank (RCSB). Table I describes the physical and the chemical properties used in this study. A sample of the data set is shown in Table II. Table III shows the correlation between each feature. There is negative correlation of energy with Euclidean distance, pair number, residue length and area. There is high correlation between (i) Euclidean distance and pair number, (ii) residue length and pair number, and (iii) residue length and area.

Table I. Description of the Features

Feature	Information
Area	Total surface area.
ED	Euclidean distance.
Energy	Total empirical energy.
SS	Secondary structure penalty.
RL	Residue length
PN	Pair number

Table II. Sample Dataset

RMSD	Area	ED	Energy	SS	RL	PN
0	8243.0	4939.6	-3391.1	86	75.00	165
3	7918.2	11984.2	-2273.2	29	153.00	102
4	9354.8	11535.1	-2422.5	66	67.00	186
2	15664.1	129761.0	-5820.4	146	104.00	368
0	8836.1	12198.8	-2926.1	80	66.00	101
5	12629.3	41461.0	-6206.8	146	61.00	116

Table III. Correlation between Each Feature

	Energy	SS	ED	PN	RL	Area
Energy	1.000	0.003	-0.001	-0.001	-0.002	-0.002
SS	0.003	1.000	0.514	0.572	0.670	0.656
ED	-0.001	0.514	1.000	0.953	0.838	0.803
PN	-0.001	0.572	0.953	1.000	0.913	0.837
RL	-0.002	0.670	0.838	0.913	1.000	0.942
Area	-0.002	0.656	0.803	0.837	0.942	1.000

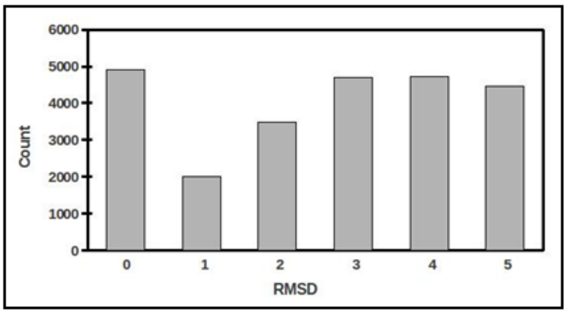


Figure 2. Distribution of RMSD in the Dataset

A. Data Transformation:

Here, RMSD of protein structure lies between 0A$^{\circ}$ and 6A$^{\circ}$. For classification purpose, RMSD is transformed into discrete value using eq. (1), keeping in mind that closer RMSD of protein structures having similar features. There may be more transformation rule that can be used for transformation. The Figure 2 show the distributionof the RMSD in the dataset. The data count for RMSD=0 is highest and least for RMSD=1.

$$\text{Class}= \begin{cases} 0 \;\; if\; 0 \le RMSD \le 1.0 \\ 1 \;\; if\; 1.0 \le RMSD \le 2.0 \\ 2 \;\; if\; 2.0 \le RMSD \le 3.0 \\ 3 \;\; if\; 3.0 \le RMSD \le 4.0 (1) \\ 4 \;\; if\; 4.0 \le RMSD \le 5.0 \\ 5 \;\; if\; 5.0 \le RMSD \le 6.0 \end{cases}$$

B. Feature Measurement

Here, we present a brief discussion of the physical and the chemical properties used in this study.

1. Root Mean Square Deviation (RMSD): The RMSD iscalculated using the superposition between matched pairs of C between two protein sequences. This superposition is computed using the Kabsch rotation matrix [19, 20] as shown below:

$$\text{RMSD} = \sqrt{\sum_i^N \frac{(d_i*d_i)}{N}}$$

where, d_i is the distance between matched pair i, N is the number of matched pairs. RMSD is calculated using the freely available program at [21].

2. Total surface area (Area): Protein folding is ruled byvarious driving forces, which seek towards minimization of its total surface area. Degree of these external forces depends on the surface of protein exposed to the solvent, which convey the strong dependency of free energy on solvent accessible surface area (SASA) [22]. SASA has been widely used as one of the important properties to assess the quality of protein structures. Hydrophobic collapse is considered as a major factor in protein folding and this can be estimated as a loss of SASA of non-polar residues. Each amino acid shows a different affinity to be found on the surface of the protein based on the functional groups present in its side chain [23]. Some questions arise with regard to the usage of SASA: (i) should it be the total area or is it the area of the non-polar residues, (ii) what is the standard fixed value of SASA for a native structure and (iii) isthe rule of minimum area applicable to non-globular proteins. Here, total SASA have been calculated using Lee & Richards [23] method.

3. Euclideandistance (ED): Spatial positioning of$C_\alpha$atomsdecides the overall conformation of a protein. Recently, neighbor-hood profiles of C_α atoms for each pair of residues have been characterized and observed to be invariant in 3618 native proteins suggesting certain geometrical constraints in their positioning [24]. The authors consider four aliphatic non polar residues Alanine (ALA), Valine (VAL), Leucine (LEU) and Isoleucine (ILE); collectively they formed 6 unique pairs among each other. Cumulative inter-atomic distance of their respective C_β atoms were calculated for each residue pair. Euclidean distance is calculated by taking the cumulative difference of $C_\alpha$and C_β . Euclidean distance between two protein sequences p and q is given as:

$$E_d = \sqrt{\sum_{i=0}^{n}(q_i - p_i)^2}(2)$$

Where n is sequence length.

4. Total empirical energy (Energy): The total empirical energy is the absolute sum of electrostatic force, Vander Waals force and hydrophobic force [25, 26]. Molecular dynamics simulation package AMBER12 [12] is used to compute total empirical energy. It is computed as given below:

$$E_{elec}^{ij} = \frac{332*qi*qj}{rij}$$

$$E_{vdW}^{ij} = \frac{C_{12}^{ij}}{r_{ij}^{12}} - \frac{C_6^{ij}}{r_{ij}^{6}}$$

$$E_{hyd}^{ij} = \frac{M_{12}^{ij}}{r_{ij}^{12}} - \frac{M_6^{ij}}{r_{ij}^{6}}$$

Where,r_{ij} is the distance between pair of atoms i and j,

$C_{12}^{ij} = \in \sigma^{12}$,$C_6^{ij} = 2 \in \sigma^6$,σ is be van der waals

radii ,∈ is well depth , $M_{12}^{ij} = \in R^{12}$,$M_6^{ij} = \in R^6$,R is the distance variable and ϵ is set to 1. Finally total empirical energy is given below:

$$E_{total} = \sum_{i}^{n-1} \sum_{j=j+1}^{n} \left(E_{elec}^{ij} + E_{vdw}^{ij} + E_{hyd}^{ij}\right)$$

5. Secondary Structure penalty (SS): Secondary Structure Prediction has reached to 82% accuracy [27] over the last fewyears. Therefore deviation from ideal predicted secondary structures can be used as a measure to quantify the quality of a structure. Secondary structure penalty is measured from the secondary structure sequence. It is computed as the absolute difference of the STRIDE and the PSIPRED scores. STRIDE is used to

assign three secondary structure classes, *i.e.*, helix, sheet and coil to each residue in the protein models based on coordinates. PSIPRED is used to predict the probability for the same secondary structure classes.

$S_{stride}(P)=S_{helix}(P)+S_{sheet}(P)+S_{coil}(P)$
$S_{psipred}(\mathrm{P})=F_1(P)+F_2(P)+F_3(\mathrm{P})$
$\mathrm{SS} = \mathrm{abs}\,(S_{stride}\,(P)-S_{psipred}\,(\mathrm{P})\,)$ (3)

Where, P is the protein sequence ;S_{stride}(P) and $S_{psipred}$(P)Are the STRIDE and PSIPRED scores respectively ;S_{helix}(P),S_{sheet} (P) ,S_{coil}(P) are the STRIDE scores of helix ,sheet andCoil of protein sequence P respectively ;F_1(P) is the predictedProbability from PSIPRED for the secondary structure of the

Central residue in the sequence window ;F_2(P) is the corres-Pondence between predicted and actual secondary structureOver a 21- residue window ;F_3(P) is the secondary structureAssigned by STRIDE, binary encoded into three classes over a 5-residue window.

6. Pair Number (PN): Pair Number is the Total Number of Aliphatic hydrophobic residue pairs in the protein structure and it is calculated by counting the total number of pairs between the Cβ carbons in the protein structure.

7. Residue Length (RL): Residue length is the total number of Cβ carbons in the protein structure.

3. Methodology

The methodology is described in Figure 3. In the first step, The modeled protein structures are taken from protein structure prediction center (CASP-5 to CASP-10 experiments), public decoys database [18] and native structure from protein data bank (RCSB). The feature measurement, as discussed in section, of protein structures is carried out in second step. The removal of duplicates and missing value entries from dataset were carried out along with the transformation in the third step. There are total 24294 decoys structures having 4919 native structures. In the forth step, the ABC algorithm [28] is used to measure the importance of each feature. Feature selection makes the prediction of model efficient and accurate. In the fifth step, the four machine learning approaches (refer, Table V) were trained and tested on the data set with their default parameters. Figure 4 describes the prediction model. Finally, the evolution of the model is done on accuracy and sensitivity and K-fold cross validation is used to measure robustness of the best predictive model.

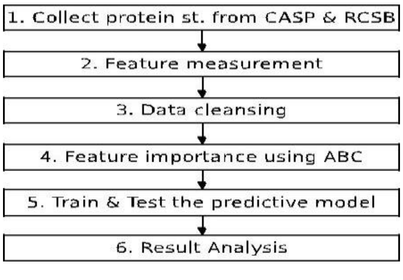

1. Collect protein st. from CASP & RCSB
2. Feature measurement
3. Data cleansing
4. Feature importance using ABC
5. Train & Test the predictive model
6. Result Analysis

Figure 3. Methodology Used

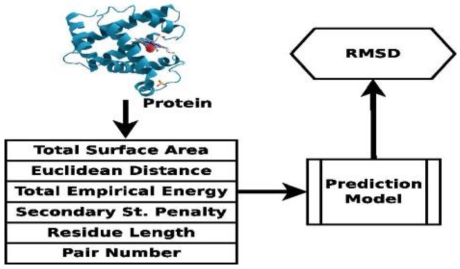


Figure 4. Prediction Model

A. Artificial Bee Colony (ABC)

The ABC algorithm [28] is a relatively recent swarm intelligence based algorithm. The algorithm is inspired by the intelligent food foraging behavior of the honey bees. Each solution of the problem is called food source of honey bees. The fitness is determined in terms of the quality of the food source. The honey bees are classified into three groups namely employed bees, onlooker bees, and scout bees. The number of employed bees are equal to the onlooker bees. The employed bees are the bees which search the food source and gather the information about the quality of the food source. Onlooker bees, which stay in the hive, search the food sources on the basis of the information gathered by the employed bees. The scout bees search new food sources randomly in places of the abandoned foods sources. Similar to the other population-based algorithms, ABC is also an iterative search algorithms. After, initialization of the ABC parameters and swarm, it requires the repetitive iterations of the three phases namely employed bee phase, onlooker bee phase and scout bee phase. There are three control parameters in ABC search process: the number of food sources SN (equal to number of onlooker or employed bees), the value of limit and the maximum number of iterations [29]. The pseudo-code of the ABC is shown in Algorithm 1.

B. Feature Importance using ABC

The ABC is used to find the importance of each features. It defines the weight to each feature according to the objective function defined in eq. (4).

Algorithm 1 Artificial Bee Colony Algorithm:

```
Initialize the parameters;
While Termination criteria is not satisfied do
Step 1: Employed bee phase for generating new food sources;
Step 2: Onlooker bees phase for updating the food sources
depending on their nectar amounts;
Step 3: Scout bee phase for discovering the new food sources in place of abandoned food
sources;
Step 4: Memorize the best food source found so far;
end while
Output the best solution found so far.
```

Table IV. Importance of Each Feature using ABC

Runs	Energy	RL	PN	SS	ED	Area
1	0.256	0.184	0.172	0.150	0.123	0.115
2	0.250	0.190	0.169	0.153	0.120	0.118
3	0.253	0.187	0.172	0.150	0.123	0.115
4	0.249	0.182	0.174	0.148	0.125	0.122
5	0.251	0.184	0.177	0.156	0.117	0.115
Avg.	0.252	0.185	0.173	0.151	0.122	0.117

The parameters for the ABC are the colony size (NP=50; [30, 31]), number of food sources (SN=NP/2), dimension of the problem (D=6), limit (number of trials after which a food source is considered to be abandoned; D*SN [32, 33]) and the termination criteria (number of iterations = 2000).

After five different runs, the weight obtained for each feature is described in Table IV. The average weight of energy is highest and area is lowest that also signifies the importance of each feature in the dataset. As the weight given to each feature is significant so all the features are selected for the experiment.

$$\text{Objfun} = \min\left(\sum_{i=1}^{T}\sqrt{\left(R_i - \sum_{j=1}^{n} w_j * p_{i,j}\right)^2}\right) \qquad (4)$$

where, T is the total number of instances in training data set, R is the RMSD, P is physical and chemical properties, n is the number of properties (6 in this case) and w is the weight given to each feature defined in the range of [0,1].

C. Machine Learning Methods

In this work, we used four machine learning models (refer,Table V) for prediction of RMSD of protein structure. The models are available in R open source software. R is licensed under GNUGPL. A brief of the models is presented below:

1) Decision Trees: This model is an extension of C5.0 classification algorithms described by Quinlan [34].

2) Random forest: It is based on a forest of trees using random inputs [35].

3) Support Vector Machine: SVM is a powerful method for general (nonlinear) classification and outliers detection with an intuitive model representation [36].

4) Linear Models: It uses linear models to carry out regression,single stratum analysis of variance and analysis of covariance [37].

Table V. Machine Learning Classification Model Used

Model	Package	Tuning Parameter(s)	Ref.
Decision Trees	C50	window, model, trials	[34]
Random Forest	randomForest	mtry	[35]
SVM	e1071	nu, epsilon	[36]
LM	stats	None	[37]

Table VI. Parameter Setting for Machine Learning Models

Model	Parameter Setting
Decision Trees	Min Split = 20, Max Depth = 30, Min Bucket = 7
Random Forest	Number of tree = 500
SVM	Kernel Radial Basis
LM	Multinomial

4. Model Evaluation

There are various ways to measure performance of the classifiers; some are more suitable than others depending on the considered application. The formula used for all the machine learning models is given by

$$\text{RMSD} \sim \text{Area} + \text{ED} + \text{Energy} + \text{SS} + \text{RL} + \text{PN}$$

This paper uses sensitivity and accuracy as a pair of (S, C) for measuring the performance of machine learning models. To determine the (S, C), a confusion or error matrix is formedshowing the information about actual and predicted classification done by a classifier. The diagonal elements of confusion or error matrix represent the number of objects for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the error matrix, better the accuracy. If there are n number of classes then the value$C_{ij}$of the confusion matrix of size n × n represents thenumber of patterns of class i predicted in class j.

The classifier accuracy can be calculated as:

$$\text{Accuracy}=\frac{\sum_{i=1}^{n} c_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}} \qquad (5)$$

However, the classification accuracy may show inaccurateresults in a case when there is a high variance in the numberof objects in the classes. Hence, this paper represents the accuracy of the classifier as a pair of values (S;C) where S is the minimum of sensitivities among all classes and C is the overall accuracy [38], [39]. Sensitivity Si for the class i can be defined as the number of patterns correctlypredicted to be in class i with respect to the total number of patterns in class i which is shown below [38].

$$S_i=\frac{c_{ij}}{\sum_{j=1}^{n} c_{ij}}(6)$$

Therefore, the sensitivity (S) of the classifier will be theminimum value of the sensitivities as shown in eq. (7) [38].

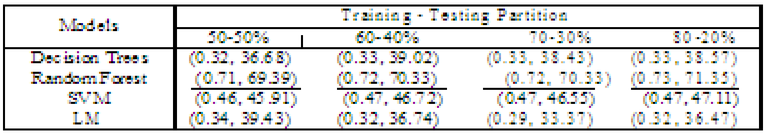
S = min(Si; i = 1; :::; n) (7)

The correct classification rate or accuracy (C) for the classifier

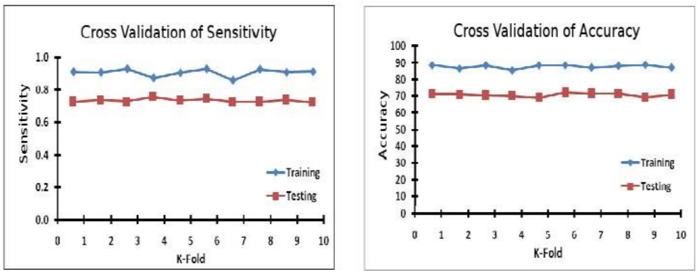
is defined as in eq. (8) that is, the rate of all the correct predictions[38].

$$\text{C}=\frac{1}{n}\sum_{i=1}^{n} S_i \text{ (8)}$$

Table VII. Performance Comparison of all Four Models on Different Training-testing Partitions using Random Forest in Sensitivity and Accuracy Pair



Models	Training - Testing Partition			
	50-50%	60-40%	70-30%	80 -20%
Decision Trees	(0.32, 36.68)	(0.33, 39.02)	(0.33, 38.43)	(0.33, 38.57)
Random Forest	(0.71, 69.39)	(0.72, 70.33)	(0.72, 70.33)	(0.73, 71.35)
SVM	(0.46, 45.91)	(0.47, 46.72)	(0.47, 46.55)	(0.47, 47.11)
LM	(0.34, 39.43)	(0.32, 36.74)	(0.29, 33.37)	(0.32, 36.47)



(a) Sensitivity **(b) Accuracy**

Figure 5. 10-fold Cross Validation of Sensitivity and Accuracy on Training-Testing Dataset (70-30%) in the Prediction of RMSD using Random Forest

K-fold cross validation is used to measure accuracy of thepredictive model. The original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model and the remaining k-1 subsamples are used as training data. The crossvalidation process is then repeated k times (the folds) with each of the k subsamples used exactly once as the validation data. Further, the k results from the folds are can be averaged to produce a single estimation. The advantage of this model over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. Here, 10-fold (k=10) cross validation is used to measure the robustness of the best selected model.

5. Experimental Results

In this section, we analyze the prediction results of all the four

machine learning classification models on the testing dataset. All the four methods are run on their default parameters as shown in Table VI. The accuracy is calculated using eq. (8) and is shown in Table VII for all the models on 50-50, 60-40, 70-30 and 80- 20 training-testing partitions. It is evident that the random forest have the highest sensitivity and accuracy pair of (0.71,69.39%), (0.72,70.33%), (0.72,70.33%), and (0.73,71.35%) on the trainingtesting partitions respectively.

Further, 10-fold cross validation is used to measure robustness of the random forest. Figure 5(a) and Figure 5(b) shows the sensitivity and accuracy respectively for the 10 folds. Cross validation results show a uniform performance in accuracy using random forest. The results validates that random forest outperforms the machine learning models in the classification.

6. Conclusion

In this work we explore the machine learning classification models with six physical and chemical properties to predict the structure of protein in the absence of its true native state. The results indicate that random forest outperforms the other existing classification models. The work can be extended for more physical and chemical properties and other computational methods to enhance the performance of machine learning methods. The data set and source code used in the study are available at http://bit.ly/ RMSDClassification-DS.

References

[1] B. Rost and C. Sander, “Improved prediction of proteinsecondary structure by use of sequence profiles and neuralnetworks”, Proceedings of the National Academy of Sciences, vol. 90, no. 16, **(1993)**, pp. 7558–7562.

[2] B. Rost, C. Sander,, “Prediction of protein secondarystructure at better than 70% accuracy”, JMB, vol. 232, no. 2, **(1993)**, pp. 584–599.

[3] J. Cheng, H. Saigo and P. Baldi, “Large-scale prediction of disulphide bridges using kernel methods, two-dimensionalrecursive neural networks, and weighted graph matching”, Proteins: Structure, Function, and Bioinformatics, vol. 62, no. 3, **(2005)**, pp. 617–629.

[4] D. T. Jones, *et al.*, “GenTHREADER: an efficient and reliableprotein fold recognition method for genomic sequences”, JMB, vol. 287, no. 4, **(1999)**, pp. 797–815.

[5] D. Kim, D. Xu, J. Guo, K. Ellrott and Y. Xu, “PROSPECTII: protein structure prediction program for genome-scale applications”, Protein engineering, vol. 16, no. 9, **(2003)**, pp. 641–650.

[6] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac and A. K. Dunker, “exploiting heterogeneous sequence properties improves prediction of protein disorder”, Proteins: Structure, Function, and Bioinformatics, vol. 61, no. S7, **(2005)**, pp. 176–182.

[7] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton and D. T. Jones, “The DISOPRED server for the prediction of proteindisorder”, Bioinformatics, vol. 20, no. 13, **(2004)**, pp. 2138–2139.

[8] J. Cheng, M. J. Sweredoski and P. Baldi, “Accurate prediction of protein disordered regions by mining protein structure data”, Data Mining and Knowledge Discovery, vol. 11, no. 3, **(2005)**, pp. 213–222.

[9] A. A. Travers, “DNA conformation and protein binding. Annualreview of biochemistry, vol. 58, no. 1, **(1989)**, pp. 427–452.

[10] A. Krogh, B. E. Larsson, G. Von Heijne, E. L. L. Sonnhammer, *et al.*, “Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes”, Journal of molecular biology, vol. 305, no. 3, **(2001)**, pp. 567–580.

[11] K. Bryson, D. Cozzetto and D. T. Jones, "Computer-assisted protein domain boundary prediction using the Dom-Pred server", Current Protein and Peptide Science, vol. 8, no. 2, **(2007)**, pp. 181–188.

[12] P. Fariselli, O. Olmea, A. Valencia and R. Casadio, "Prediction of contact maps with neural networks and correlated mutations", Protein engineering, vol. 14, no. 11, **(2001)**, pp. 835–843.

[13] O. Olmea and A. Valencia, "Improving contact predictions by the combination of correlated mutations and other sources of sequence information", Folding and Design, vol. 2, **(1997)**, pp. S25–S32.

[14] P. Baldi and G. Pollastri, "A machine learning strategy forprotein analysis", Intelligent Systems, vol. 17, no. 2, **(2002)**, pp. 28–35.

[15] K. T. Simons, C. Kooperberg, E. Huang, D. Baker, *et al.*, "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions", JMB, vol. 268, no. 1, **(1997)**, pp. 209–225.

[16] B. Wallner and A. Elofsson, "Prediction of global and localmodel quality in CASP7 using Pcons and ProQ", Proteins: Structure, Function, and Bioinformatics, vol. 69, no. S8, **(2007)**, pp. 184–193.

[17] J. Qiu, W. Sheffler, D. Baker and W. S. Noble, "Ranking predicted protein structures with support vector regression", Proteins: Structure, Function, and Bioinformatics, vol. 71, no. 3, **(2007)**, pp. 1175–1182.

[18] **(2012)**, www.scfbio-iitd.res.in/software/pcsm/dataset/PublicDecoys.

[19] M. R. Betancourt and J. Skolnick, "Universal similarity measure for comparing protein structures", Biopolymers, vol. 59, no. 5, **(2001)**, pp. 305–309.

[20] W. Kabsch, "A discussion of the solution for the bestrotation to relate two sets of vectors", ActaCrystallographicaSection A: Crystal Physics, Diffraction, Theoretical and General Crystallography, vol. 34, no. 5, **(1978)**, pp. 827–828.

[21] **(2012)**, http://zhanglab.ccmb.med.umich.edu/TMscore/RMSD.f.

[22] E. Durham, B. Dorr, N. Woetzel, R. Staritzbichler and J. Meiler, "Solvent accessible surface area approximations for rapid and accurate protein structure prediction", JMM, vol. 15, no. 9, **(2009)**, pp. 1093–1108.

[23] J. Janin, Surface and inside volumes in globular proteins, **(1979)**.

[24] A. Mittal and B. Jayaram, "Backbones of folded proteinsreveal novel invariant amino acid neighborhoods", Journal of Bio-molecular Structure and Dynamics, vol. 28, no. 4, **(2011)**, pp. 443–454.

[25] N. Arora and B. Jayaram, "Strength of hydrogen bonds ina helices", JCC, vol. 18, **(1997)**, pp. 1245–1252.

[26] P. Narang, K. Bhushan, S. Bose and B. Jayaram, "Protein structure evaluation using an all-atom energybased empirical scoring function", Journal of Bio-molecular Structure and Dynamics, vol. 23, no. 4, **(2006)**, pp. 385–406.

[27] T. Z. Sen, R. L. Jernigan, J. Garnier and A. Kloczkowski, "GOR V server for protein secondary structure prediction", Bioinformatics, vol. 21, no. 11, **(2005)**, pp. 2787–2788.

[28] A. K. Qin, V. L. Huang and P. N. Suganthan, "Differential evolution algorithm with strategy adaptation for global numerical optimization", Evolutionary Computation, IEEE Transactions on, vol. 13, no. 2, **(2009)**, pp. 398–417.

[29] D. Karaboga and B. Akay, "A comparative study of artificial bee colony algorithm", Applied Mathematics and Computation, vol. 214, no. 1, **(2009)**, pp. 108–132.

[30] K. Diwold, A. Aderhold, A. Scheidler and M. Middendorf, "Performance evaluation of artificial bee colony optimizationand new selection schemes", Memetic Computing, vol. 1, no. 1, **(2011)**, pp. 1–14.

[31] M. El-Abd, "Performance assessment of foraging algorithms vs. evolutionary algorithms", Information Sciences, vol. 182, no. 1, **(2011)**, pp. 243–263.

[32] D. Karaboga and B. Akay, "A modified artificial bee colony (ABC) algorithm for constrained optimization problems", AppliedSoft Computing, vol. 11, no. 3, **(2011)**, pp. 3021–3031.

[33] B. Akay and D. Karaboga, "A modified artificial bee colonyalgorithm for real-parameter optimization", Information Sciences, vol. 192, no. 3,, **(2012)**, pp. 120–142.

[34] J. R. Quinlan, "Induction of decision trees", Machine learning, vol. 1, no. 1, **(1986)**, pp. 81–106.

[35] A. Liaw and M. Wiener, "Classification and Regression by random Forest", R News, vol. 2, no. 3, **(2002)**, pp. 18–22.

[36] S. S. Keerthi and E. G. Gilbert, "Convergence of a generalized SMO algorithm for SVM classifier design", Machine Learning, vol. 46, no. 1, **(2002)**, pp. 351–360.

[37] J. C. Fernandez Caballero, F. J. Martinez, C. Hervas and P. A. Gutierrez, "Sensitivity Versus Accuracy in Multiclass Problems Using Memetic Pareto Evolutionary Neural Networks", IEEETransactions on Neural Networks, vol. 21, **(2010)**, pp. 750–770.

[38] M. Saraswat and K. Arya, "Supervised Leukocyte Segmentation in Tissue Images Using Multi-objective Optimization Technique", Engineering Applications of Artificial Intelligence, **(2013)**.