European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 10, 2020

Hybrid of PSO-GSA based Clustering Approach for Predicting Structural Class Prediction using Random Forest Method

Sarneet Kaur[1], Ashok Sharma[2], Parveen Singh[3]

[#]Computer Application[1], Computer Science[2,3], LPU University[1,2] ,University of Jammu[3]

E-mail:[1] sarneetkaur@live.com, *[2]*ashok.23877@lpu.co.in, *[3]* imparven@gmail.com

Abstract**: Protein can be classified in different classes. A lot of research is being performed for analyzing the structure and classes of protein. There are many problems associated with protein structure. Some of them are folding problem and protein structure prediction (PSP) etc. PSP is the most considerable open problem in field of biology. In the present work different algorithms like particle swarm optimization (PSO), gravitational search algorithm (GSA) and K-Mean clustering algorithms are used to classify different structures of protein. A random forest (RF) classifier is proposed for analyzing and comparing different protein classes in terms of other conventionally available algorithms in terms of various performance parameters like accuracy, recall, precision and specificity. The proposed classifier proved better than other classifiers in terms of accuracy and can be helpful in predicting the protein structures. A hybrid PSO-GSA algorithm is also proposed which provided improved performance as compared to single algorithms and can be utilized for analysis of protein structure.

Keywords: Gravitational Search Algorithm, K-Mean Clustering Algorithm, Protein Structure Prediction, Particle Swarm Optimization, Random Forest Classifier.

1. INTRODUCTION:

Proteins are consisting of a chain of amino acids (AA), which can be organized as secondary structures of three main types: helices (termed as α structure), the strands (termed as β structure), and the coils. Levitt and Chothia firstly defined and structured the protein classes [1]. On the basis of the pioneer work, the authors distinguished four different structure classes of globular proteins as: (1) an all-α class, where only small quantity of strands is included in proteins; (2) an all-β class in which only a small quantity of helices is included in proteins; (3) α/β class in which both helices as well as the strands are included and the strands are mainly parallel with each other; (4) an α + β class, where both helices as well as the strands are included and the strands are mainly anti-parallel with each other [2]. This structural class knowledge of proteins is helpful for understanding a wider problem called as protein structure prediction (PSP). The knowledge of these structural classes are useful for predicting the accuracy of the secondary structure, and reducing the possible conformations of search space for the tertiary arrangement [3, 4].

A lot of important biological functions are determined by protein's spatial structure [5]. Presently, two processes are utilized for determining the protein structure which is Nuclear Magnetic Resonance (NMR) as well as X-Ray Crystallography (XRC). However, a lot of time and money are consumed in both processes. Accordingly, there is an enormous gap between the volume of decoded and cataloged sequences of protein structures [6]. So presently, studies and research is undergoing to involve silicon methods for predicting the native structure of protein with an aim of reducing this gap and promoting the development

of precise, cost effective and time saving models. Hence, this method can predict the structure of protein using different computational techniques is called as Protein Structure Prediction (represented as PSP). PSP is a major concern/problem in analyzing the spatial structure of protein [7].

For solving this PSP problem in numeral computational techniques are suggested by literature using several problem concepts classified as: threading modeling, homology modeling, and the ab-initio modeling etc. The ab-initio modeling aims in predicting the protein's native conformation using its main sequence and physicochemical properties of amino acids (AA) like the hydrophilic or hydrophobic interaction [8]. The ab-initio modeling PSP problem is approaching with the use of off-lattice and on-lattice modeling. On-lattice model limits the structure of protein in a lattice. The Hydrophobic-Hydrophilic (HH) is an approach using on-lattice type of modeling which is proposed by Dill (in 1985). This is possibly a lesser complex model ab-initio (on-lattice type) [9]. In spite of its simplicity, this HH model is also verified by Berger & Leighton (in 1998). Hence, HH model is circulated to other abstractions of PSP in which superior degree of freedom is presented. A distinguished disadvantage in on-lattice model is that there is not enough detail in protein representations, so it is difficult to reproduce a genuine protein structure [10].

The reason of native structure predicting of small proteins using ab-initio model is that these are inexpensive conformation evaluation, and it presents enormous and multimodal space for search. So the need of the hour is to design and develop simplified models for protein like HH model for reducing the time complexity and degree of freedom [11]. The main objectives of such models are testing, development, and contrasting of several methods. A simplified three-dimensional model like AB off-lattice can be used for demonstrating two-phase optimization efficiency by utilizing Differential Evolution (DE) algorithms [12].

Protein–protein interactions can also be performed by using a Bayesian framework in a superior approach based on an unsupervised technique of learning, in which the models of network studies are presented in given form of protein [13]. Direct mapping match is undergone utilizing hyper parameters in PPI modeling form. For molecular search, parameterized BLOSUM metrics are used for sending back the alignment models of existing proteins. A simulation model is performed by considering the data value interconnections for identifying the model efficiency [14].

1.1 Protein Structure Prediction:

Protein structures are present in every living being and are composed of a series of amino acids (AA) linked together with peptide bonds [15]. Each component of amino acid is characterized with a central carbon (Cα) atom which is attached by an atom of hydrogen, a carboxyl (COOH) group, an amine (NH_2) group, and side-chain (radical R) [16]. Numerous classes of amino acids are existed in nature, out of these only 20 classes are proteinogenic. Amino acids are classified as per their affinity with water (termed as hydrophobicity): hydrophilic and hydrophobic amino acids. Proteins can be synthesized in cells ribosome as per template provided by a messenger RNA (mRNA). During this synthesis process, the protein is folded into an exclusive 3D configuration. This process is also termed as protein folding process. Most stable 3D structures (termed as native structure) are available under the physiological conditions, which allow a protein atom to carry out its function. However, any failure in folding into the proposed 3D structure typically lead to protein atoms with dissimilar properties that becomes inactive. Such incorrectly folded (misfolded) proteins are harmful for the body organism. The most challenging and important problem existed in applications related to Molecular Biology like drug design, needs a superior understanding of this protein folding problem. In modern Computational Biology, there exist two problems related to protein folding. First problem is correct prediction of protein structure from its

primary structure, whereas second problem is prediction of pathways of protein folding, which composed of establishing the event sequences that lead to folding of protein's primary structure (a linear chain of amino acids) to the native structure. This lead to a problem called as protein structure prediction (PSP) that has a major practical importance in present. Different models for studying the PSP problem in protein structure are proposed by Computer science and Physics fields [17]. However, system multidimensionality creates a problem in computing PSP as simulation using computational models generally take into account all protein atoms which is normally unfeasible ($> 104^0$ of freedom) [18], even if the most powerful resources for computation are used. Accordingly, numerous simplified models for abstracting this protein structure are proposed. The simplest model for computation of PSP problem is Hydrophobic-Polar (HP), this model exist in two dimensions (2D-HP) and three dimensions (3D-HP) [19]. This computational method for finding a solution of PSP problem using these HP models is termed as *NP*-complete [20]. Some other models like AB *off-lattice* and 3D Side-Chain model for HP (3DHP-SC) are also used for solving PSP problem [21-22].

2. PROPOSED METHOD

Protein secondary structure (PSS) described as primary folded structures, are produced inside polypeptide because of interactions among backbone atoms. In general, four classes of protein structure exists in nature: (1) an all-α class, where only small quantity of strands is included in proteins; (2) an all-β class in which only a small quantity of helices is included in proteins; (3) α/β class in which both helices as well as the strands are included and the strands are mainly parallel with each other; (4) an α + β class, where both helices as well as the strands are included and the strands are mainly anti-parallel with each other. PSS classification is vital for diverse biological functions which include: recognition of protein fold, prediction of tertiary structure, DNA-binding prediction, and conformation search area reduction. In current article, a model based on machine learning for PSS classification is proposed. Here both sequence-based as well as structure-based features are considered. Firstly, preprocessing on protein data is applied, then a clustering technique i.e hybrid model of PSO and GSA optimization with collaboration of *K*-Mean clustering is proposed. Selected clusters are used for training of classifier random forest, and evaluation of performance parameter.

2.1 METHODOLOGY

1. Read data from excel and apply pre-processing on data for refining dataset. we used FC699 dataset for Protein secondary structure (PSS)

2. after that apply k-mean clustering on data for make initial cluster and find out centroid point that centroid points take input for optimization algorithm or take initial population for generate by kmean clustering.

3. Initialize PSO and GSA parameters like C1, C2 and G0 and number of population, maximum iterations;

4. Generate best solution of clustering with the help of hybrid of PSO and GSA optimization.

5. Initialize Random forest for classification and evaluation of performance parameter.

3.PROPOSED ALGORITHM

In this section, we have explained the proposed algorithm which used for clustering and classification of various proteins. The flowchart of the proposed algorithm is shown in Figure 1. Next, the detail description of each component of the flowchart is explained below.

3. 1 K-Means algorithm

Macqueen in 1967 developed a simple clustering algorithm termed as *k*-means algorithm. This algorithm is based on an uncomplicated and unsupervised partitioned cluster algorithm in which the data is clustered on the basis of given *k*-value of data. An iteration technique is utilized for producing independent data produced into a variety of clusters with their data properties similar to each other. There are two separate segments in this algorithm. The first segment provides a methodology for random selection of *k* center by any user, whereas the second segment recalculates the average value of the different clusters formed previously.

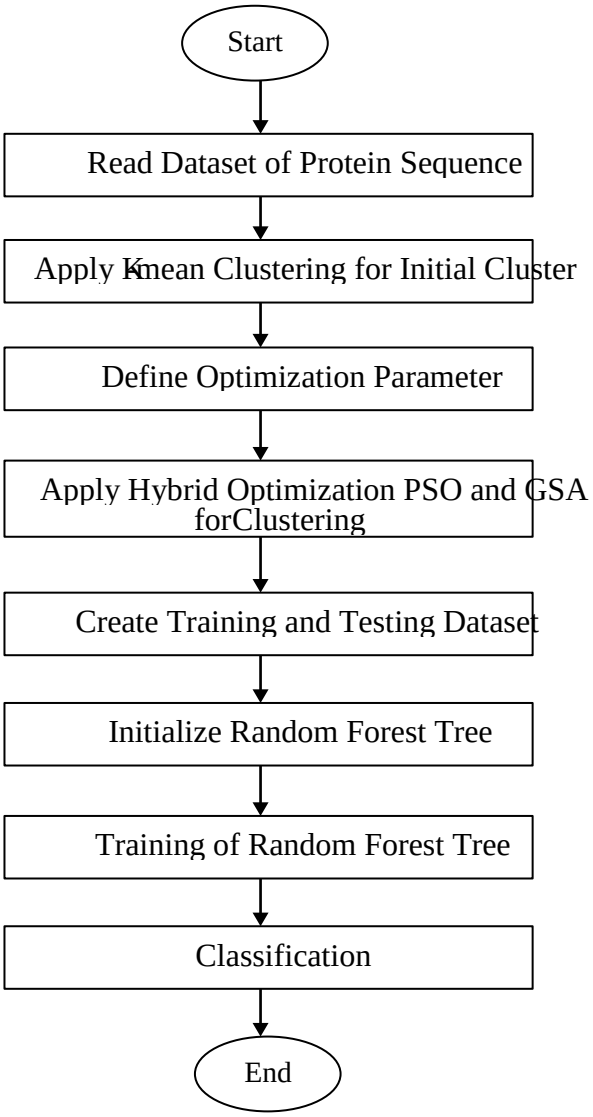


Fig. 1 Flowchart of the Proposed Technique

In the first segment, numerous metrics for distance calculation (like Euclidean distance) are considered for taking the individual object into the closest center. Thus each identified object in all the clusters is considered and early grouping of these objects is done in the same way to

finish the first segment. In the next phase, the average value of previously shaped clusters is recalculated. This process of iteration is continued until the criterion function is allotted a highest value. Iteration is stopped when this value reaches to minimum.

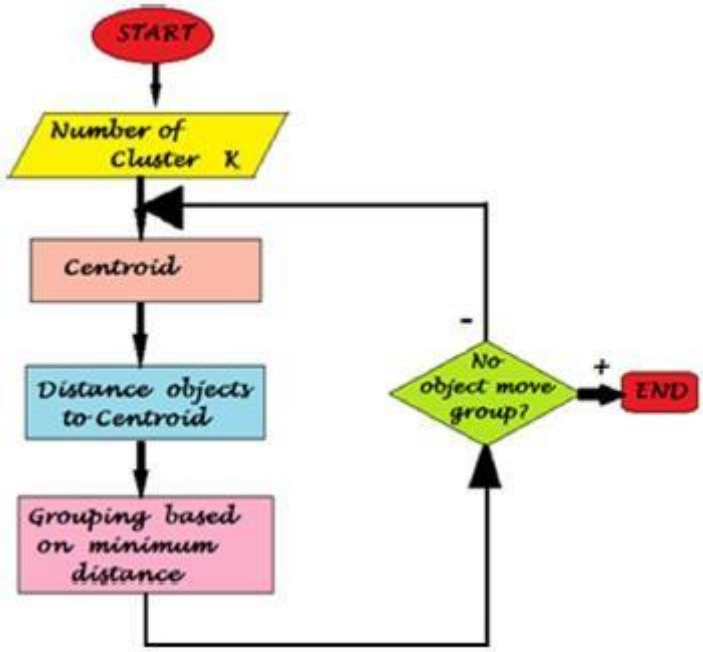


Fig. 2: k –Means Clustering Algorithm

Various calculations needed for clustering by *k*-means algorithms are specified as:

$$d(n, z) = \min_{1<i<k} d(n, z_i) \quad (1)$$

$$d(N, Z) = \frac{\sum d(n_i, Z)^2}{L} \quad (2)$$

Here N = {$n1$,…. nL} is a set for k centers, also Z ={Z_1,…Z_k} is a mean square distance which is computed between the cluster center and given data point. To complete the operation, a complexity analysis is performed.

Generally for the process of grouping, different steps performed by *k*-means algorithm are indicated in Fig. 2. In the starting phase similar average data are grouped assuming an initial value of neighboring average data. Afterwards the initial data is calculated by the average value of a cluster of individual data. Then, an initial data for the individual is again assumed for the identified group of neighboring average data. Lastly, the classification process is checked for the next data and then to next until data is not changing and same data value is resembled with the previous one. After this process, the clustering is stopped and the result is produced. If there is any failure in the checking process, the process is again repeated until the same data value is achieved.

3.2 Particle Swarm Optimization (PSO)

PSO algorithm is provoked by the organized movement of bird flocks and fishes [23]. The PSO is composed of swarm of elements which interact with one another in a constant search area. A prospective solution of any problem can be represented with position of each element and representation is done like an *n*-dimensional space vector. The particles in PSO "fly" throughout the n-dimensional search area, and socio-cognitive affinity decides the possible

change in their positions to imitate the success accomplished by further particles. The life experience of every particle in the swarm is different from other particles and the quality of each particle is evaluated by its own experiences. As an individual in social gathering, each particle has knowledge about the behavior of its neighbors. The information of the cognitive factor is also termed as individual learning whereas as information of social factor is termed as cultural transmission. Hence every individual's decision is made by accounting both cognitive as well as social factors, which lead the swarm population to an evolving behavior [24]. The algorithm for PSO is as indicated by Algorithm 1.

Algorithm 1: PSO Algorithm

```
Set parameters: n, φ_p, φ_g
for i = 1 to n do
    Initialize the positions x⃗_i and velocities v⃗_i randomly
    Evaluate fitness f(x⃗_i)
    Initialize the particle's best known position to its initial position: p⃗_i = x⃗_i
    if f(p⃗_i) is better than f(g⃗) then
        Update the swarm's best known position: g⃗ = p⃗_i
    end if
end for
while stop condition not met do
    for i = 1 to n do
        Update particles' velocity: v⃗_i = v⃗_i + φ_p * r_p * (p⃗_i − x⃗_i) + φ_g * r_g * (g⃗ − x⃗_i)
        Update particles' position: x⃗_i = x⃗_i + v⃗_i
        if f(x⃗_i) is better than f(p⃗_i) then
            p⃗_i = x⃗_i
            if f(p⃗_i) is better than f(g⃗) then
                g⃗ = p⃗_i
            end if
        end if
    end for
end while
Postprocess results and visualization
```

3.3 Gravitational Search Algorithm (GSA)

GSA is formed on the basis of the gravitational law and the concept of interaction of masses [25]. The Newton theory of physics is utilized by GSA algorithm and its search instruments are the mass collectors. GSA contains an isolated organism of different masses. On the basis of gravitational force, each mass in the organism can notice the condition of the other mass. So by using the gravitational force, the information can be transferred between diverse masses. In GSA, agent is an object whose performance is calculated with its mass. All such objects interact with one another by the gravitational force which causes the combined movement of these objects towards the heavier mass object. The heavy masse objects form a superior solution of this problem. A solution to the given problem is provided by agent's position, is used for determining its mass [26].

Algorithm 2: Gravitational Search Algorithm

1: Set parameters: n, α, G_0
2: **for** $i = 1$ **to** n **do**
3: Initialize the positions $\vec{x}_i$ randomly
4: Initialize velocities $\vec{v}_i$ and acceleration $\vec{a}_i$ to zero
5: **end for**
6: **while** stop condition not met **do**
7: Evaluate the fitness of each agent
8: Update G, $best$ and $worst$ of the population
9: Calculate mass (M) and acceleration ($\vec{a}_i$)
10: Update velocity $\vec{v}_i$ and position $\vec{x}_i$
11: **end while**
12: Postprocess results and visualization

4. PERFORMANCE EVALUATION

K-mean clustering algorithm can be used for predicting the structures of four classes of protein α (A), β (B), α/β (C), and α + β (D). The performance can be analyzed feature wise and class wise in terms of different parameters like: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP illustrates the correctly marked positive samples of protein. TN illustrates the correctly marked negative samples of protein. FP illustrates the incorrectly marked positive samples of protein. FN illustrates the incorrectly marked negative samples of protein. FP is also termed as a type-I error. FN is also termed as a type-II error. For understanding the PSP problem, both this errors are taken into consideration.

Various calculations of these parameters (TP, TN, FP, and FN) can be considered for evaluating the performance of various parameters such as Accuracy, Precision, Specificity, and Recall (Sensitivity) as provided in Equations (3-6). Accuracy is one of the most frequently used parameter for indicating performance of the appropriately classified samples out of total samples. Precision determines the preciseness in a model for correctly classifying the correct positive samples out of total positive samples. Recall determines the correct positive samples out of all available correct positive samples (TP+FN). Specificity determines actual negative samples out of total negative samples.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

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

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$pecificity = \frac{TN}{TN+FP} \quad (6)$$

5. RESULTS & DISCUSSIONS

The main objective of this clustering process is grouping of similar objects in a same group (cluster). A set of measurement or attributes are used to define each object. For determining any similar objects, the similarity is measured between them. Numerous similarity measures

are provided in the literatures. In current article, Euclidean distance is used for calculating the similarity between different objects. Euclidean distance is provided by following equation:
1 2

$$distanc(o_i, o_j) = (\sum^{m}{}_{p=1}|o_{ip} - o_{jp}|^2) \quad (7)$$

Here, m represents total no. of attributes, o_{ip} represents the attribute number's value, p for the object 'i' (o_i). For solving the problem of data clustering, the standard algorithms (PSO and GSO) are adapted for reaching the centroid of clusters.

5.1 Performance Evaluation For Protein Structural Class A

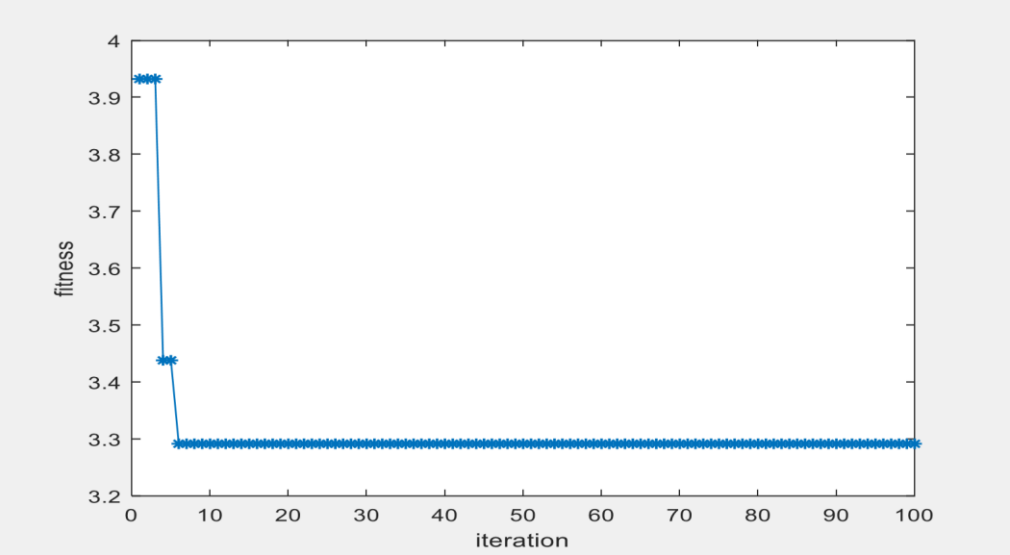


Fig. 3: Variation of fitness function with no. of iterations for Class A

Figure 3 provides the variation of the fitness function as per the number of iterations for Protein Structural Class A. Figure 4 highlights the performance of different parameters (accuracy, recall, precision, and specificity) values (in
%) accomplished by proposed Random Forest (RF) classifier with FC699 represented test data. Figure 5 (Protein Structural Class A) highlights the comparison of accuracy values accomplished by the proposed RF classifier with FC699 test data with other classifier like SVM, Ada boost, RF etc. As provided by figure 5, accuracy of proposed RF classifier is much higher than other classifiers.

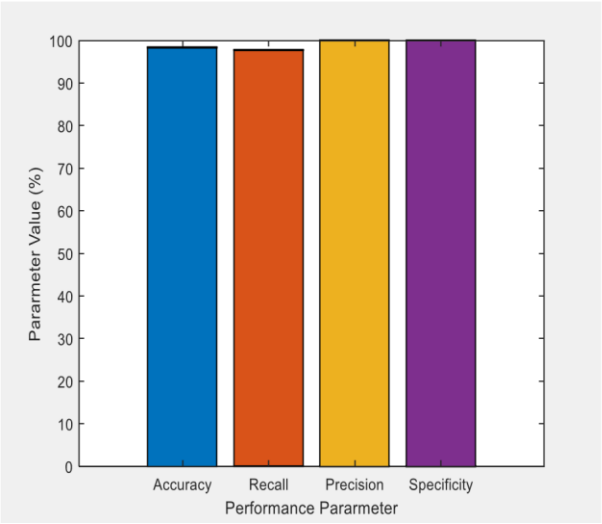


Fig. 4: Performance of different parameters

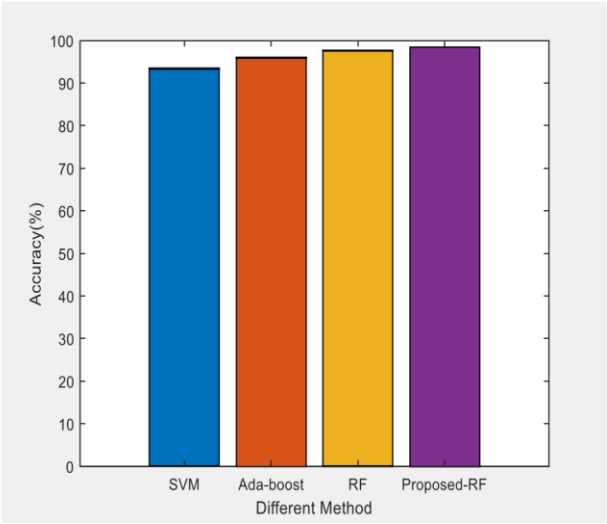


Fig. 5: Accuracy of different methods

Table 1: Accuracy comparison of proposed RF method with other methods

Sr.No.	Technique	Accuracy (%)
1	SVM	93.36
2	Ada-boost	96
3	RF	97.56
4	Proposed-RF	98.36

Table 1 compares the accuracy values of proposed RF method with different methods like SVM, Ada-boost, and RF for prediction of protein structure for class A using FC699 Data sets. Table 2 provides the performance parameters values (accuracy, recall, precision, and specificity) accomplished by the proposed RF classifier.

Table 2: Performance value of different parameters for proposed RF method

Sr.No.	Performance (%)	Proposed RF
1	Accuracy	98.36
2	Recall	97.72
3	Precision	100
4	Specificity	100

5.2 PERFORMANCE EVALUATION FOR PROTEIN STRUCTURAL CLASS B

Figure 6 provides the variation of the fitness function as per the number of iterations for Protein Structural Class B.

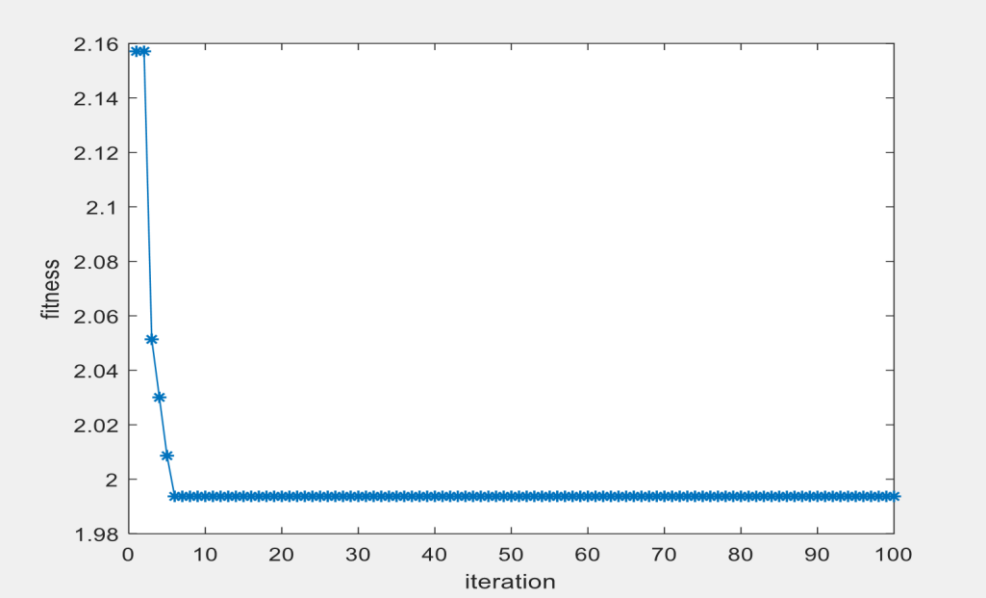
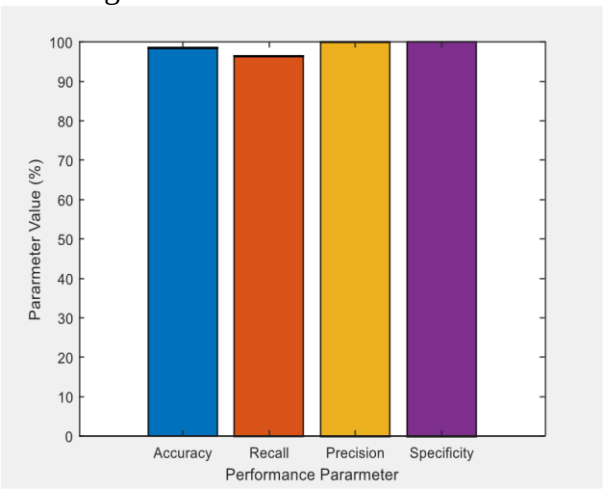


Fig. 6: Variation of fitness function with no. of iterations for Class-B

Figure 7 highlights the performance of different parameters (accuracy, recall, precision, and specificity) values (in

%) accomplished by proposed Random Forest (RF) classifier with FC699 represented test data. Figure 8 (Protein Structural Class B) highlights the comparison of accuracy values accomplished by the proposed RF classifier with FC699 test data with other classifier like SVM, Ada boost, RF etc. As provided by figure 8, accuracy of proposed RF classifier is much higher than other classifiers.



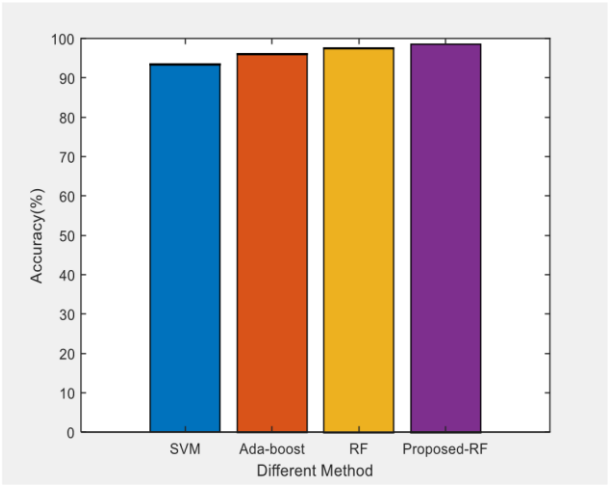


Fig. 7: Performance of different parameters

Fig. 8: Accuracy of different methods

Table 3: Accuracy comparison of proposed RF method with other methods

Sr.No.	Technique	Accuracy(%)
1	SVM	93.36
2	Ada-boost	96
3	RF	97.56
4	Proposed-RF	98.47

Table 3 compares the accuracy values of proposed RF method with different methods like SVM, Ada-boost, and RF for prediction of protein structure for class B using FC699 Data sets. Table 4 provides the performance parameters values (accuracy, recall, precision, and specificity) accomplished by the proposed RF classifier.

Table 4: Performance value of different parameters for proposed RF method

Sr.No.	Performance (%)	Proposed RF
1	Accuracy	98.47
2	Recall	96.36
3	Precision	100
4	Specificity	100

5.3 PERFORMANCE EVALUATION FOR PROTEIN STRUCTURAL CLASS C

Figure 9 provides the variation of the fitness function as per the number of iterations for Protein Structural Class C.

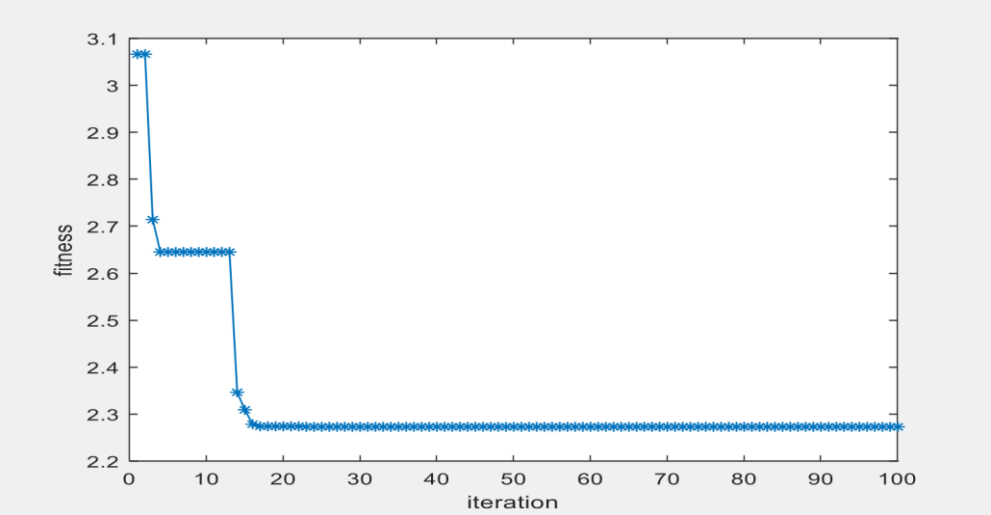
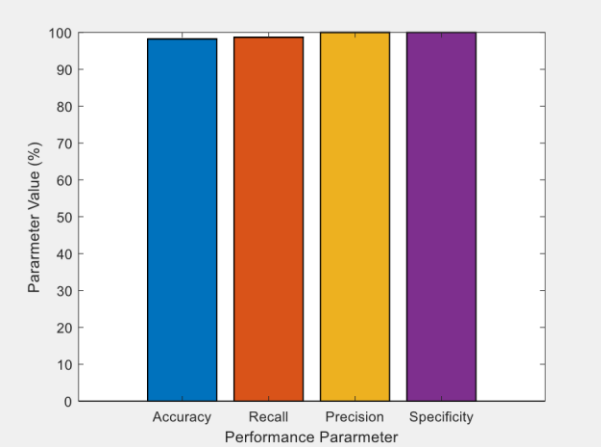


Fig. 9: Variation of fitness function with no. of iterations for Class-C

Figure 10 highlights the performance of different parameters (accuracy, recall, precision, and specificity) values (in

%) accomplished by proposed Random Forest (RF) classifier with FC699 represented test data. Figure 11 (Protein

Structural Class C) highlights the comparison of accuracy values accomplished by the proposed RF classifier with FC699 test data with other classifier like SVM, Ada boost, RF etc.



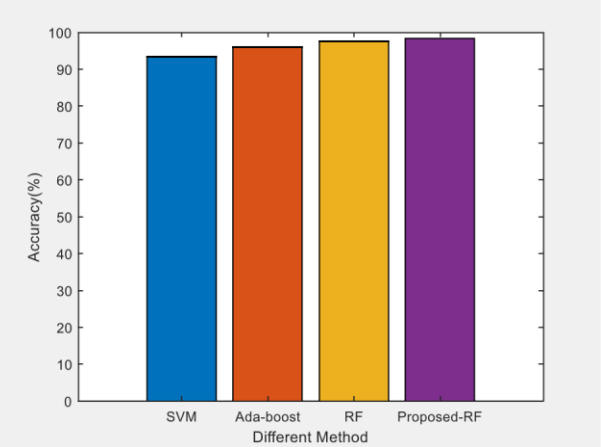


Fig. 10: Performance of different parameters Fig. 11: Accuracy of different methods

As provided by figure 10, accuracy of proposed RF classifier is much higher than other classifiers.

Table 5: Accuracy comparison of proposed RF method with other methods

Sr.No.	Technique	Accuracy(%)
1	SVM	93.36
2	Ada-boost	96
3	RF	97.56
4	Proposed-RF	98.22

Table 5 compares the accuracy values of proposed RF method with different methods like SVM, Ada-boost, and RF for prediction of protein structure for class C using FC699 Data sets. Table 6 provides the performance parameters values (accuracy, recall, precision, and specificity) accomplished by the proposed RF classifier.

Table 6: Performance value of different parameters for proposed RF method

Sr.No.	Performance (%)	Proposed RF
1	Accuracy	98.22
2	Recall	98.70
3	Precision	100
4	Specificity	100

5.4 PERFORMANCE EVALUATION FOR PROTEIN STRUCTURAL CLASS D

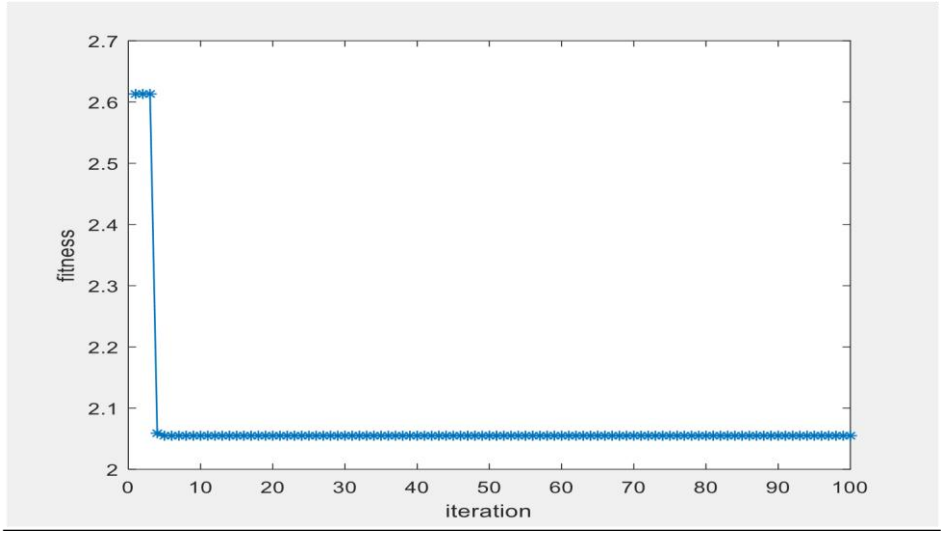
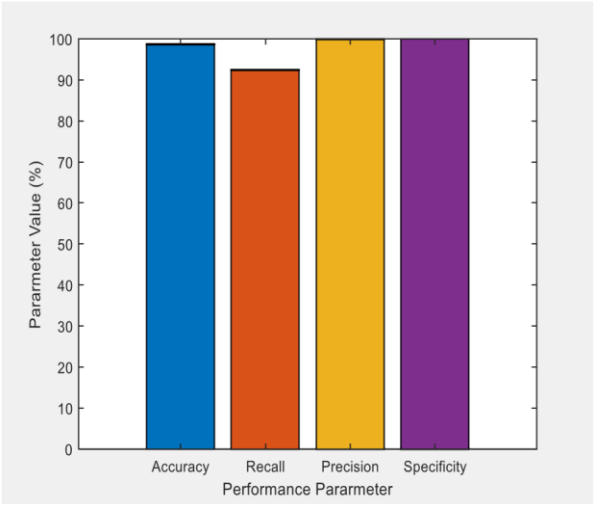


Fig. 12: Variation of fitness function with no. of iterations for Class-D

Figure 12 provides the variation of the fitness function as per the number of iterations for Protein Structural Class D. Figure 13 highlights the performance of different parameters (accuracy, recall, precision, and specificity) values (in
%) accomplished by proposed Random Forest (RF) classifier with FC699 represented test data. Figure 14 (Protein

Structural Class D) highlights the comparison of accuracy values accomplished by the proposed RF classifier with FC699 test data with other classifier like SVM, Ada boost, RF etc. As provided by figure 13, accuracy of proposed RF classifier is much higher than other classifiers.



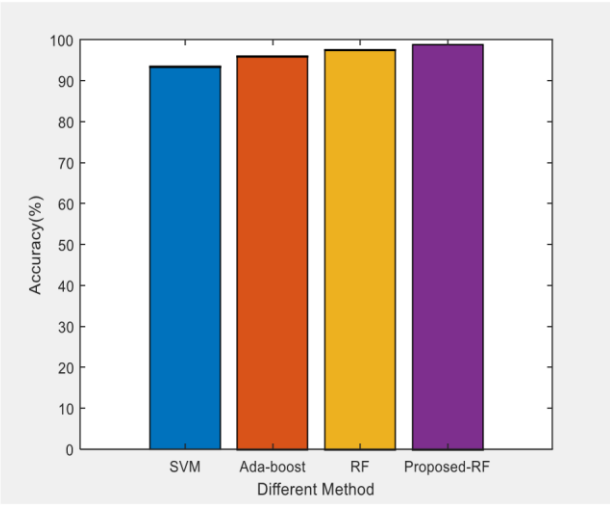


Fig. 13: Performance of Different Parameters Fig. 14: Accuracy of Different Methods

Table 7: Accuracy comparison of proposed RF method with other methods

Sr.No.	Technique	Accuracy(%)
1	SVM	93.36
2	Ada-boost	96

3	RF	97.56
4	Proposed-RF	98.70

Table 8: Performance value of different parameters for proposed RF method

Sr.No.	Performance (%)	Proposed RF
1	Accuracy	98.70
2	Recall	94.73
3	Precision	100
4	Specificity	100

Table 7 compares the accuracy values of proposed RF method with different methods like SVM, Ada-boost, and RF for prediction of protein structure for class D using FC699 Data sets. Table 8 provides the performance parameters values (accuracy, recall, precision, and specificity) accomplished by the proposed RF classifier.

5.5 HYBRID PSO-GSA ALGORITHM

Fig. 15: Combined model to measure different parameters for all classes

Figure 15 provides the performance values of different parameters (accuracy, recall, precision, and specificity) accomplished by proposed RF classifier for FC699 data set for different structural classes of Protein.

Table 9: Parameters of Hybrid PSO-GSA algorithm for clustering of protein structure

Sr.No	Parameter	PSO-GSA
1	Population size	50
2	Iteration	100
3	C1	2
4	C2	2
5	W	0.72
6	G0	1

Table 9 provides the parameters of Hybrid PSO-GSA algorithms for clustering of protein structure classes. Here, C1, C2 are acceleration coefficients, W is inertia weight (PSO), G_0 is used for controlling the search accuracy for GSA algorithm.

6. CONCLUSIONS

In this paper, different structural classes of protein are classified in order to make an understanding of different problems like folding and protein structure prediction etc. Clustering is performed using K-mean algorithm. In current work a random forest (RF) classifier is proposed which is compared with conventional classifiers like SVM, Ada boost, RF etc. in terms of accuracy for all four classes of protein. The accuracy of proposed RF classifier is much higher than other classifiers. Also, the values of performance parameters like accuracy, recall, precision, and specificity are measured for different classes of protein. A Hybrid PSO-GSA algorithm was analyzed and its different parameters are analyzed for classification of protein structure. As suggested in literature, the proposed hybrid PSO-GSA algorithm has proved to achieve better results as compared to single algorithms.

7. REFERENCES

[1] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature, 261 (1996) 552–557.

[2] M. Gromiha, S. Selvaraj, Protein secondary structure prediction in different structural classes. Protein Eng. 11 (1998) 249–251.

[3] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[4] I. Bahar, A.R. Atilgan, R.L. Jernigan, B. Erman, Understanding the recognition of protein structural classes by amino acid composition, Proteins 29 (1997) 172–185.

[5] D. L. Nelson, M. M. Cox, Principles of Biochemistry, seventh Edition, W. H. Freeman and Company, One New York Plaza, New York, NY, 760 USA, 2017.

[6] N. Jana, S. Das, J. Sil, A Metaheuristic Approach to Protein Structure Prediction: Algorithms and Insights from Fitness Landscape Analysis, Vol. 31 of Emergence, Complexity and Computation, Springer International Publishing, Cham, Switzerland, 2018.

[7] F. Campeotto, A. Dal Pal`u, A. Dovier, F. Fioretto, E. Pontelli, A Constraint Solver for Flexible Protein Models, Journal of Artificial Intelligence Research 48 (1) (2013) 953 – 1000.

[8] D. H. Kalegari, H. S. Lopes, An Improved Parallel Differential Evolution Approach for Protein Structure Prediction using both 2D and 3D off-lattice models, in: 2013 IEEE Symposium on Differential Evolution (SDE), IEEE, Singapore, 2013, pp. 143 – 150.

[9] K. A. Dill, Theory for the folding and stability of globular proteins, Biochemistry 24 (6) (1985) 1501 – 1509.

[10] B. Berger, T. Leighton, Protein Folding in the Hydrophobic-hydrophilic (HP) is NP-complete, in: Proceedings of the Second Annual International Conference on Computational Molecular Biology,
RECOMB'98, ACM, New York, NY, US, 1998, pp. 30 – 39.

[11] W. E. Hart, A. Newman, Protein Structure Prediction with Lattice Models, 2005.

[12] F. H. Stillinger, T. Head-Gordon, C. L. Hirshfeld, Toy model for protein folding, Phys. Rev. E 48 (1993) 1469{1477.

[13] Birlutiu A, d'Alche-Buc F, Heskes T (2015) A Bayesian framework for combining protein and network topology information for predicting protein–protein interactions. IEEE Trans Comput Biol Bioinform 12(1):538–550

[14] Song D, Chen J, Chen G, Li N, Li J, Fan J, Bu D, Li SC (2015) Parameterized BLOSUM matrices for protein alignment. IEEE Trans Comput Biol Bioinform 12(3):686–694.

[15] L. Hunter. (1993). *Artificial Intelligence and Molecular Biology*. AAAI Press, Boston, USA, 1 edition.

[16] D.L. Nelson and M.M. Cox. (2008). *Lehninger Principles of Biochemistry*.W.H. Freeman, 5*th* edition.

[17] H.S. Lopes. (2008). Evolutionary algorithms for the protein folding problem: A review and current trends. In T.G. Smolinski, M.M. Milanova, and A-E Hassanien, editors, *Computational Intelligence in Biomedicine and Bioinformatics*, volume I, pages 297–315. Springer-Verlag, Heidelberg, Germany.

[18] A. Liwo, M. Khalili, and H. A. Scheraga. (2005). Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proceedings of the National Academy of Sciences*, 102(7):2362–2367.

[19] K.A. Dill, S. Bromberg, K. Yue, and K.M. Fiebig et al. (1995). Principles of protein folding - a perspective from simple exact models. *Protein Science*, 4(4):561–602.

[20] F.H. Stillinger, T. Head-Gordon, and C. Hirshfeld. (1993). Toy model for protein folding. *Physical Review E*, 48(2):1469–1477.

[21] C.M.V. Ben´ıtez and H.S. Lopes. (2010). Hierarchical parallel genetic algorithm applied to the threedimensional HP side-chain protein folding problem. In *Proc. of IEEE International Conference on Systems, Man and Cybernetics*, pages 2669–2676. IEEE Computer Society.

[22] P. Crescenzi, D. Goldman, C. Papadimitrou, A. Piccolboni, and M. Yannakakis. (1998). On the complexity of protein folding. *Journal of Computational Biolology*, 5:423–446.

[23] J. Kennedy and R.C. Eberhart. (1995). Particle swarm optimization. In *Proc. of the IEEE Int. Conf. on Neural Networks*, pages 1942–1948, Piscataway, USA. IEEE Press.

[24] Tapas Si and Nanda Dulal Jana. (2012). Particle swarm optimisation with differential mutation. *International Journal of Bio-Inspired Computation*, 11(3):212–251.

[25] A. Chatterjee, S.P. Ghoshal, and V. Mukherjee. (2012). A maiden application of gravitational search algorithm with wavelet mutation for the solution of economic load dispatch problems. *International Journal of Bio-Inspired Computation*, 4(1):33–46.

[26] Esmat Rashedi, Hossein Nezamabadi-pour, and Saeid Saryazdi. (2009). GSA: A gravitational search algorithm. *Information Sciences*, 179(13):2232–2248.