A Novel Technique of Feature Extraction Based on Local and Global Similarity Measure for Protein Classification

Neha Bharill and Aruna Tiwari
Department of Computer Science and Engineering, Indian Institute of Technology, PACL Campus, Indore, India

Keywords: Bioinformatics, Probability-based Features, Position-specific Information, Binary Feed Forward Neural Network, Protein Classification.

Abstract: The paper aims to propose a novel approach for extracting features from protein sequences. This approach extracts only 6 features for each protein sequence which are computed by globally considering the probabilities of occurrences of the amino acids in different position of the sequences within the superfamily which locally belongs to the six exchange groups. Then, these features are used as an input for Neural Network learning algorithm named as Boolean-Like Training Algorithm (BLTA). The BLTA classifier is used to classify the protein sequences obtained from the Protein Information Resource (PIR). To investigate the efficacy of proposed feature extraction approach, the experimentation is performed on two superfamilies, namely Ras and Globin. Across tenfold cross validation, the highest Classification Accuracy achieved by proposed approach is 94.32±3.52 with Computational Time 6.54±0.10 (s) is remarkably better in comparison to the Classification Accuracies achieved by other approaches. The experimental results demonstrate that the proposed approach extracts the minimum number of features for each protein sequence. Therefore, it results in considerably potential improvement in Classification Accuracy and takes less Computational Time for protein sequence classification in comparison with other well-known feature extraction approaches.

1 INTRODUCTION

In the recent years, Bioinformatics (Iqbal et al., 2014) is emerged as a forefront research area. It is referred as conceptualization of biology in terms of macromolecules. Due to dramatic evolution of technology and continuous effort of Genome Project, a large amount of protein, DNA and RNA sequences are generated on regular basis. In this regard, many techniques have been proposed by the researcher to analyze and interpret the DNA, RNA and protein sequences. Among these, protein sequence classification (Vipsita and Rath, 2013) is an important problem, which determines the superfamily of an unknown protein sequence. The major advantage of category grouping is that molecular analysis is performed globally within a superfamily instead of local analysis. A protein sequence contains the characters from 20 different amino acid alphabets that can occur in any order. The problem of protein classification are formally stated in (Wang et al., 2001). Given a unlabeled protein sequence $\mathcal{S}$ and a set of known superfamilies F=$\{F_1, F_2, ..., F_f\}$, the problem is to determine with certain degree of accuracy whether the protein sequence $\mathcal{S}$ belongs to one of superfamilies from set F_i, $i = 1, ..., f$. Therefore, classification of unknown protein sequences into one of known superfamilies is an important task. This will help in identifying the structure and function of unknown protein sequences. It also results in saving the large expenses incurred in performing the experiments in laboratory. One of the most important practical applications is in drug discovery. For example, suppose a sequence $\mathcal{S}$ is obtained from disease $\mathcal{D}$ and it is inferred by classification method that sequence $\mathcal{S}$ belongs to the superfamily F_i. So, to treat the disease $\mathcal{D}$ one can use the combination of existing drugs of superfamily F_i.

In past, many feature extraction approaches (Vergara and Estévez, 2014) have been proposed by the researchers to deal with the protein sequence classification problem. The n-gram encoding schemes (Wang et al., 2001), (Solovyov and Lipkin, 2013) for extracting features from protein sequence used the local and global similarities by counting the occurrences of two amino acids within a protein sequence. Further, the extracted features are used as an input to Bayesian Neural Network classifier. Although, the n-gram encoding scheme works reasonably well. But its major

Bharill N. and Tiwari A..
A Novel Technique of Feature Extraction Based on Local and Global Similarity Measure for Protein Classification.
DOI: 10.5220/0005283702190224
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2015), pages 219-224
ISBN: 978-989-758-070-3

drawback is that it fails to consider the positional significance of the residue pairs which is an important consideration in superfamily classification. In addition to it, the number of features extracted by this approach is extremely large (≥ 62). This imposes a major limitation on many classification approaches. Also, these algorithms works on large number of features therefore they have high-computational complexity. In 2005, Bandyopadhyay proposed another feature extraction approach that overcomes some of the limitations of (Wang et al., 2001). This approach limit the number of features and correspondingly it extracts 20 features for each protein sequence. Once the features are extracted, they are used as an input to fuzzy genetic clustering strategy to evolve a set of prototypes for each superfamily. Finally, it uses the nearest neighbour (NN) rule to classify a set of unknown sequences into a particular superfamily. But, this approach only considers the global positional information corresponding to each amino acid. Thus, it fails to consider the local positioning of each amino acid in the respective sequence. Another approach proposed by (Mansouri et al., 2008) which extract only relevant features from the protein sequences by counting the occurrence probability of six exchange groups in each sequence. Then, it uses these extracted features as an input for generating some of the interpretable fuzzy rules which is used to assign protein sequences into appropriate superfamily. This approach suffers from a major drawback that the features extracted by this approach only considers the local positioning of each sequence within an exchange group. It fails to consider the global probability of occurrence of each amino acid in entire superfamily. Hence, the above discussed classifiers do not capture both the global and local similarity. Thus, the relevant features are not extracted due to which it results in degradation of classification accuracy and have high-computational time.

In this paper, the proposed new feature extraction approach overcomes the limitations of existing feature extraction approaches. It capture both the global and local similarity of each protein sequence for extracting features and only 6 relevant features corresponding to each protein sequence are extracted. Firstly, it computes the global probability of each amino acid present within a sequence by counting the positional information of amino acid in all the sequences. Then, the local similarities are determined based on the concept of weighting scheme (Karchin and Hughey, 1998). Further, the computed weights of each amino acid within a sequence is encoded to their respective six exchange groups where the exchange groups are effective equivalence classes of amino acids derived from PAM (Dayhoff and Schwartz, 1978). Once, the features are extracted then these features are fed as an input to the Boolean-Like Training Algorithm (BLTA) (Gray and Michel, 1992) to perform the classification of unknown sequences into the superfamilies. To validate the efficacy of proposed feature extraction approach, the comparison is done by implementing other feature extraction approaches and evaluating their performance on BLTA classifier. The observation can be drawn from the experimental results that the proposed approach limit the number of features extracted corresponding to each protein sequence by capturing both local and global similarity measure thus, leads to the lesser Computational Time and higher Classification Accuracy.

The rest of the paper is organized as follows. The description of proposed model is illustrated in Section 2. In Section 3, the experimental results are reported. Finally, Section 4 concludes this paper.

2 PROPOSED FEATURE EXTRACTION APPROACH

Protein sequence contains characters from the amino acid that can be viewed as a text strings which is formally represented by a set $\mathcal{A}$=$\{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$. The protein sequence can be of any length and contains the combination of these amino acids in any order. Therefore, the most important issue in applying any algorithm for the protein sequence classification is encoding of these protein sequences in terms of feature vectors and then applying these feature vectors as an input to any learning algorithm for classification. For proper classification of sequences into superfamilies a relevant input representation is needed. Thus, the success of learning algorithm depends on the kind of input data available. The proposed new feature extraction approach extracts only 6 relevant features corresponding to each protein sequence by capturing both the global and local similarity of protein sequences. Next, section is presented with the proposed method for computing the global similarity corresponding to all the protein sequences belongs to the superfamily. Section 2.2, describe the proposed local similarity measure which

Sequence	Position				
	1	2	3	4	5
1	M	K	G	D	H
2	M	K	A	V	Y
3	M	K	G	V	H
4	M	A	K	A	S
5	M	K	G	V	H

Figure 1: Primary Structure of Five Related Proteins.

incorporate the global similarity measure computed for all the protein sequences. Finally, in Section 2.3, the proposed encoding method is presented which evaluate the feature vector once the global and local similarity measures for all the protein sequence is determined.

2.1 Global Similarity Measure for Feature Extraction

Given, a set $\mathcal{S}$ consist of all the sequences of a protein superfamily F_i, $i = 1,...,f$ which is formally represented as $\mathcal{S}$=$\{s_1, s_2, ..., s_n\}$ where n represent the number of sequences. The protein sequences which belong to the same superfamily share the structural similarities with each other as shown in Fig. 1. All the sequences present in Fig. 1 are unaligned and taken from same superfamily and consist of total 9 amino acids represented by a set $\mathcal{A}$=$\{A, D, G, H, K, M, S, V, Y\}$. These unaligned sequences are aligned using BioEdit tool. The global similarity measure is determined by calculating the probability of occurrence of each amino acid in a particular position with respect to the total number of sequences present in the superfamily. It is mathematically represented by

$$(Probability)_{ij} = (Occurence)_{ij}/n \qquad (1)$$

where $(Probability)_{ij}$ represent the probability of occurrence of i^{th} amino acid at j^{th} position, $(Occurence)_{ij}$ denote the frequency of i^{th} amino acid at j^{th} position and n represent the total number of sequences in a particular superfamily. For example, as shown in Fig. 1, the amino acid G occurs in the third position three times out of five sequences, therefore probability of occurrence of G is $(Probability)_{ij}$=$\frac{3}{5}$. Thus in the same way, the global similarity measure for all the sequences shown in Fig. 1 is computed and presented in Table 1. Once the global similarity measure is evaluated, the position specific weight of each amino acid is calculated. This is discussed in the subsequent section.

Table 1: Global Similarity Measure of Protein Sequences present in Fig. 1.

Amino acids	Position1	Position2	Position3	Position4	Position5
A	0	0.2	0.2	0.2	0
D	0	0	0	0.2	0
G	0	0	0.6	0	0
H	0	0	0	0	0.6
K	0	0.8	0.2	0	0
M	1	0	0	0	0
S	0	0	0	0	0.2
V	0	0	0	0.6	0
Y	0	0	0	0	0.2

Table 2: Feature Vector of Each Sequence Present in Fig. 1.

Sequence	e_1	e_2	e_3	e_4	e_5	e_6
1	1.4	0.2	0	0.6	1	0
2	0.8	0	0	0.2	1.6	0.2
3	1.4	0	0	0.6	1.6	0
4	0.2	0	0	0.6	1	0
5	1.4	0	0	0.6	1.6	0

2.2 Local Similarity Measure for Feature Extraction

Given a protein sequence, the weight of each amino is evaluated by adding all the position specific occurrences of amino acid at that place and the respective probability of occurrence of amino acid in that place from the entire super family. It is mathematically represented as:

$$Weight(i) = (PSO)_i \times (Probability)_{ij} \qquad (2)$$

where $Weight(i)$ denote the weight of i^{th} amino acid, $(PSO)_i$ represent the position specific occurrence of i^{th} amino acid and $(Probability)_{ij}$ represent the probability of occurrence of i^{th} amino acid at j^{th} position. For example, in Fig. 1 corresponding to the sequence1 i.e. MKGDH, the weight of each amino acid is calculated as follows:
Weight(M)=1×1.0=1.0, Weight(K)=1×0.8=0.8
Weight(G)=1×0.6=0.6, Weight(D)=1×0.2=0.2
Weight(H)=1×0.6=0.6
The weights of all other amino acids present in Fig. 1 with respect to the sequence1 is zero. This is because these amino acids are not present in sequence1. Hence, for all other remaining sequences the weight of each amino acid present within the sequence are calculated in the similar manner.

Therefore, for sequence2 i.e. MKAVY, the weight of each amino acid is evaluated as follows.
Weight(M)=1×1.0=1.0, Weight(K)=1×0.8=0.8
Weight(A)=1×0.2=0.2, Weight(V)=1×0.6=0.6
Weight(Y)=1×0.2=0.2

Similarly, for sequence3 i.e. MKGVH, the weight of each amino acid is evaluated as follows.
Weight(M)=1×1.0=1.0, Weight(K)=1×0.8=0.8
Weight(G)=1×0.6=0.6, Weight(V)=1×0.6=0.6
Weight(H)=1×0.6=0.6

For sequence4 i.e. MAKAS, the weight calculation of each amino acid is presented as follows.
Weight(M)=1×1.0=1.0, Weight(K)=1×0.2=0.2
Weight(S)=1×0.2=0.2,Weight(A)=1×0.2+1×0.2=0.4

For sequence5 i.e. MKGVH, the weight calculation of each amino acid is presented as follows.
Weight(M)=1×1.0=1.0, Weight(K)=1×0.8=0.8
Weight(G)=1×0.6=0.6, Weight(V)=1×0.6=0.6
Weight(H)=1×0.6=0.6

Furthermore, these amino acids within the sequence share some structural similarity with each other. Thus, encoding of these amino acids present within the sequence is another important issue in order to represent these amino acids as a feature vector. Therefore, encoding method is presented in the subsequent section.

2.3 Encoding of Protein Sequences

According to PAM (Dayhoff and Schwartz, 1978), the amino acids belong to the six exchange groups. This is because these amino acids within the group exhibits high evolutionary similarity. The Six-letter exchange groups are formally represented as: e_1=$\{H,R,K\}$, e_2=$\{D,E,N,Q\}$, e_3=$\{C\}$, e_4=$\{S,T,P,A,G\}$, e_5=$\{M,I,L,V\}$ and e_6=$\{F,Y,W\}$. For a given protein sequence1 MKGDH present in Fig. 1, the amino acids M$\in e_5$, K$\in e_1$, G$\in e_4$, D$\in e_2$, H$\in e_1$. The encoding of these amino acids is done by finding the belongingness of each amino acid to the specific group and assign the addition of weight values of amino acids to the specific group. The weight value of amino acid M i.e. 1 is assign to the exchange group e_5, the amino acid K and H both belongs to the exchange group e_1, so the addition of their weight values i.e. $Weight(K)+Weight(H)=0.8+0.6=1.4$ is assign as an overall weight to the exchange group e_1. Similarly, the amino acids G and D belongs to the exchange group e_4 and e_2, so the weight values of G i.e. 0.6 and D i.e. 0.2 is assign to the exchange groups e_4 and e_2. Thus, one can observe that none of the amino acid from sequence1 belongs to the exchange e_3 and e_6 so the weight values assign to the exchange groups e_3 and e_6 is 0. Hence for sequence1 MKGDH, the feature vectors is obtained as $\{(e_1,1.4),(e_2,0.2),(e_3,0),(e_4,0.6),(e_5,1),(e_6,0)\}$. The feature vectors for remaining sequences shown in Fig. 1 are determined in the similar manner and presented in Table 2. The feature vectors generated using the proposed method consider both the local and global similarity and thus extract only 6 relevant features corresponding to each protein sequence. Therefore, it works effectively with any classification algorithm when applied with protein sequence data.

3 EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the experimentation is carried out to investigate the performance of the proposed approach on BLTA classifier (Gray and Michel, 1992). All codes are written in the MATLAB computing environment and tested on Intel(R) Xeon(R) $E5-1607$ Workstation PC. The data used for the experimental purpose are obtained from the International Protein Sequence Database (Barker et al., 2004), release 2012, in the Protein Information Resource (PIR). Table 3, illustrate the information of the two superfamilies used in the experimentation. In all the experiments, the 10-fold cross validation test is performed and corresponding results are reported subsequently.

Table 3: Data used in the experiments.

Name of superfamilies	Number of sequences	Minimum length of sequence	Maximum length of sequence
RAS	500	171	296
Globin	500	128	339

3.1 Parameter Specification

The proposed feature extraction approach is compared with (Mansouri et al., 2008), (Bandyopadhyay, 2005), (Wang et al., 2001) using four parameters i.e. Mean (M), Standard Deviation (SD), Classification Accuracy (CA) and Total Computational Time (TCT). These four parameters are computed by each approach only after the features are extracted corresponding to all protein sequences and classification is performed with BLTA classifier.
The Mean (M) is defined as follows

$$M=\frac{CCS}{n} \tag{3}$$

where CCS is the number of correctly classified sequences, n is the total number of sequences. The mean determines the number of protein sequences are correctly classified by each approach from the total number of sequences.
The Standard Deviation (SD) is defined as

$$SD=\Big[\frac{1}{n-1}\sum_{i=1}^{n}(P_i-M)^2\Big]^{\frac{1}{2}} \tag{4}$$

where P_i denote the i^{th} protein sequence. The SD is computed corresponding to each approach, it evaluates the overall variation occur in the mean across ten fold validation.
The total Classification Accuracy (CA) is defined as

$$CA=(M\pm SD)\times 100 \tag{5}$$

The CA is computed for each approach which determines the total classification accuracy by considering the overall variation in mean and standard deviation across ten fold cross validation.
The Total Computational Time (TCT) is defined as

$$TCT=FET+CT \tag{6}$$

Table 4: The Number of Features Extracted and the Neurons Required by all the Approaches in Each Layer.

Approaches	Consideration of Methods	Number of features extracted	Number of Inputs in Input Layer	Number of neurons		
				Hidden Layer	Inhibition Layer	Output Layer
Proposed approach	Global and Local Similarity	6	18	500	500	2
Mansoori	Local Similarity	6	18	500	500	2
Bandyopadhyay	Global Similarity	20	60	500	500	2
Wang	Global and Local Similarity	400	1200	500	500	2

where FET is the total feature extraction time, CT denotes the classification time. The TCT is a sum of time required in extracting the feature by a particular approach including the time required in classifying protein sequences to the superfamilies.

3.2 Performance Comparison with Other Approaches

The number of features extracted and the neurons required by all the approaches while evaluating the performance on BLTA classifier are summarized in Table 4. It can be observed from the table that, the number of features extracted by the proposed approach is similar to the number of features extracted by (Mansouri et al., 2008). But, the proposed approach consider both the local and global similarity measure whereas the (Mansouri et al., 2008) only considers the local similarity measure to compute the features corresponding to each sequence belongs to their respective superfamily. On the contrary, the other approaches developed by (Bandyopadhyay, 2005), (Wang et al., 2001) extract 20 and 400 features corresponding to each sequence and thus, it results in extraction of many irrelevant features for the classification of unknown protein sequence.To judge the effectiveness of proposed approach, exhaustive results across ten fold cross validation along with the performance comparison with three different existing feature extraction approaches (Mansouri et al., 2008), (Bandyopadhyay, 2005), (Wang et al., 2001) on BLTA classifier by varying *m*-circle values is reported in Table 5. Furthermore, of proposed approach . The four parameters i.e. Mean, Standard Deviation, Classification Accuracy and the Computational Time (seconds) corresponding to all the approaches is calculated by varying *m*-circle values of BLTA classifier. It is found that on protein data

Table 5: Comparison of Results in terms of Mean, Standard Deviations, Classification Accuracy and Computational Time with other Feature Extraction Approaches by Varying *m*-circle values of BLTA Classifier is Reported.

Number of *m*-circle	Proposed feature extraction approach				(Mansouri et al., 2008) proposed by Mansoori			
	Mean	Standard Deviation	Classification Accuracy	Computational Time (seconds)	Mean	Standard Deviation	Classification Accuracy	Computational Time (seconds)
2	0.9225	0.0450	92.26±4.50	6.87±0.11	0.8469	0.0100	84.70±1.00	7.63±0.17
4	0.9290	0.0424	92.90±4.25	6.55±0.24	0.8477	0.0108	84.78±1.08	7.21±0.06
8	0.9431	0.0352	94.32±3.52	6.54±0.10	0.8493	0.0106	84.94±1.06	7.23±0.14
16	0.9379	0.0385	93.79±3.86	6.48±0.09	0.8510	0.0084	85.10±0.85	7.15±0.07
32	0.9379	0.0385	93.79± 3.86	6.44±0.11	0.8510	0.0084	85.10±0.85	7.15±0.06
64	0.9379	0.0385	93.79±3.86	6.57±0.23	0.8542	0.0054	85.42±0.55	7.11±0.07
128	0.9379	0.0385	93.79±3.86	6.50±0.30	0.8542	0.0054	85.42±0.55	7.11±0.06
256	0.9379	0.0385	93.79±3.86	6.58±0.21	0.8542	0.0054	85.42±0.55	7.14±0.09
512	0.9379	0.0385	93.79±3.86	6.50±0.14	0.8542	0.0054	85.42±0.55	7.15±0.1
1024	0.9379	0.0385	93.79±3.86	6.41±0.08	0.8542	0.0054	85.42±0.55	7.13±0.14

Number of *m*-circle	(Bandyopadhyay, 2005) proposed by Bandyopadhyay				(Wang et al., 2001) proposed by Wang			
	Mean	Standard Deviation	Classification Accuracy	Computational Time (seconds)	Mean	Standard Deviation	Classification Accuracy	Computational Time (seconds)
2	0.6734	0.0829	67.34±8.30	10.60±0.18	0.5141	0.0026	51.41±0.27	63.98±1.26
4	0.6744	0.0835	67.45±8.36	10.22±0.11	0.5141	0.0026	51.41±0.27	63.44±1.24
8	0.6748	0.0838	67.49±8.38	10.18±0.14	0.5141	0.0026	51.41±0.27	63.41±1.31
16	0.6750	0.0838	67.51±8.38	10.20±0.13	0.5141	0.0026	51.41±0.27	63.40±1.26
32	0.6750	0.0838	67.51±8.38	10.13±0.14	0.5141	0.0026	51.41±0.27	63.40±1.28
64	0.6750	0.0838	67.51±8.38	10.13±0.09	0.5141	0.0026	51.41±0.27	63.42±1.20
128	0.6750	0.0838	67.51±8.38	10.14±0.14	0.5141	0.0026	51.41±0.27	63.54±1.26
256	0.6750	0.0838	67.51±8.38	10.13±0.09	0.5141	0.0026	51.41±0.27	63.49±1.31
512	0.6750	0.0838	67.51±8.38	10.14±0.17	0.5141	0.0026	51.41±0.27	63.39±1.20
1024	0.6750	0.0838	67.51±8.38	10.16±0.09	0.5141	0.0026	51.41±0.27	63.40±1.22

set, highest Classification Accuracy achieved by proposed technique is 94.32 ± 3.52 on *m*-circle value 8 with Computational Time 6.54 ± 0.10 (s). Instead, it attains 92.26 ± 4.50 as the minimum Classification Accuracy with Computational Time 6.87 ± 0.11 (s) on *m*-circle value 2. On the other hand, method given by (Mansouri et al., 2008) attains 85.42 ± 0.55 as maximum Classification Accuracy for *m*-circle values $\{64, ..., 1024\}$ with Computational Time varies from $\{7.11 \pm 0.06, .., 7.15 \pm 0.1\}$ (s) whereas it gives the minimum Classification Accuracy 84.70 ± 1.00 on *m*-circle value 2 with Computational Time 7.63 ± 0.17 (s). On the contrary, the method proposed by (Bandyopadhyay, 2005) achieves the best Classification Accuracy rate 67.51 ± 8.38 on *m*-circle values $\{16, ..., 1024\}$ with Computational Time from $\{10.13 \pm 0.09, .., 10.20 \pm 0.13\}$ (s). Although, it gives the worst Classification Accuracy rate 67.34 ± 8.30 for *m*-circle value 2 with Computational Time 10.60 ± 0.18 (s). The other method developed by (Wang et al., 2001), exhibits 51.41 ± 0.27 as minimum and maximum Classification Accuracy rate for all the values of *m*-circle with Computational Time varies from $\{63.39 \pm 1.20, .., 63.98 \pm 1.26\}$(s). Moreover, exhaustive results reported in Table 5, justify the significance of proposed approach due to the improvements in Classification Accuracy rate as well as in Computational Time when compared with the methods proposed by (Mansouri et al., 2008), (Bandyopadhyay, 2005), (Wang et al., 2001).

4 CONCLUSIONS

In this paper, a novel feature extraction approach is proposed for classifying the protein sequences into the superfamilies. The proposed approach compute both the local and global similarity measures for extracting relevant features corresponding to each protein sequence. The global similarity measure is calculated by considering probability of occurrence of the positional variance of each amino acid among all the sequences within the superfamily. However, the local similarity measure is produced by evaluating a weighting scheme (Karchin and Hughey, 1998) of the global probability and then assigns the weighted probability of each amino acid to the six exchange groups (Dayhoff and Schwartz, 1978). Finally, the 6 features are extracted corresponding to each protein sequence which is classified using Boolean-Like Training Algorithm (BLTA) (Gray and Michel, 1992).

The experimental work is carried out on two superfamilies Ras and Globin to probe the efficacy of the proposed approach on BLTA classifier in comparison with other approaches (Mansouri et al., 2008), (Bandyopadhyay, 2005), (Wang et al., 2001). Moreover, the results are analyzed and reported in terms of four parameters-Mean, Standard Deviation, Classification Accuracy and Computational Time with variation in *m*-circle values of BLTA classifier. The observation can be drawn from the experimental results, that the proposed approach extract very limited number of features in comparison with other methods. Therefore, it outperforms on the BLTA classifier and thus, achieves best Classification Accuracy 94.32 ± 3.52 with Computational Time 6.54 ± 0.10 (s) on *m*-circle value 8. Hence, its performance is much higher in comparison to other methods (Mansouri et al., 2008), (Bandyopadhyay, 2005), (Wang et al., 2001) in terms of Classification Accuracy and Computational Time.

REFERENCES

Bandyopadhyay, S. (2005). An efficient technique for superfamily classification of amino acid sequences: feature extraction, fuzzy clustering and prototype selection. *Fuzzy Sets and Systems*, 152(1):5–16.

Barker, W., Garavelli, J., Huang, H., McGarvey, P., Orcutt, B., G.Y.Srinivasarao, Xiao, C., Yeh, L., Ledley, R., Janda, J., F.Pfeiffer, H.W.Mewes, A. T., and Wu, C. (2004). The protein information resource (pir). *Nucleic Acids Research*, 28(1):41–44.

Dayhoff, M. and Schwartz, R. (1978). A model of evolutionary change in proteins. In *In Atlas of protein sequence and structure*. Citeseer.

Gray, D. and Michel, A. (1992). A training algorithm for binary feedforward neural networks. *Neural Networks, IEEE Transactions on*, 3(2):176–194.

Iqbal, M. J., Faye, I., Samir, B. B., and Said, A. M. (2014). Efficient feature selection and classification of protein sequence data in bioinformatics. *The Scientific World Journal*, 2014.

Karchin, R. and Hughey, R. (1998). Weighting hidden markov models for maximum discrimination. *Bioinformatics*, 14(9):772–782.

Mansouri, E., A.M. Zou, S. Katebi, H. M. R. B., and Sadr, A. (2008). Generating fuzzy rules for protein classification. *Iranian Journal of Fuzzy Systems*.

Solovyov, A. and Lipkin, W. I. (2013). Centroid based clustering of high throughput sequencing reads based on n-mer counts. *BMC bioinformatics*, 14(1):268.

Vergara, J. R. and Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186.

Vipsita, S. and Rath, S. K. (2013). Two-stage approach for protein superfamily classification. *Computational Biology Journal*, 2013.

Wang, J., Ma, Q., Shasha, D., and Wu, C. (2001). New techniques for extracting features from protein sequences. *IBM Systems Journal*, 40(2):426–441.