Feature Encoding Technique For Efficient Classification Of Protein Sequences

G.Bhanu Priya 1 | Dr.G.Lavanya Devi 2

1(M.Tech,Dept of CS&SE,College of engineering(A),Andhra University,Visakhapatnam,India,gpriyabhanu@gmail.com)
2(Asst Professor,Dept of CS&SE,College of engineering(A),Andhra university,Vizag,India,lavanyaadevi@yahoo.co.in)

Abstract——Bioinformatics is emerging as a new area of research recently by combining computer science and biology for automatic analysis and modeling of biological data. The volume of data generated from the next generation sequencing projects is growing enormously. The data consists of DNA, RNA and protein sequences which contain extremely important information about genes, their structure and function. Computational techniques which involve machine learning and pattern recognition are becoming useful in biological data mining. The process of classifying protein sequences into family/superfamily based on the primary sequence is a very complex and open problem. Although, there are many problems in protein superfamily classification, however the three major issues are the selection of suitable feature encoding method, extraction of an optimized subset of features having higher discriminatory information for the representation of protein sequence and adaptation of an appropriate classification technique that classify sequences with highest classification accuracy. In this paper, we propose a distance based feature encoding technique for extraction of features; the performance of the proposed technique is validated with different classifiers, which show better results than the previously available techniques. The average classification accuracy achieved is 91.2% on the benchmark dataset downloaded from the renowned UniProtKB database.

Keywords- Feature encoding; Data Mining; Feature selection; Superfamily; Protein Classification Algorithm;

I. INTRODUCTION

Bioinformatics is the application of computer science in the field of biology to investigate the biological data to get useful information from it. The size of the sequencing data about Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA) and proteins is increasing every day. The essential building blocks of life are proteins and it comprises of 20 amino acids. Three different hidden characteristics of new proteins. The sequence analysis using classification of sequences into families/super families will be very useful to recognize the structure and function of unknown protein sequences. This research can be practically applied in drug discovery, drug design and identification of genetic diseases. Since, the disorder protein sequences can be successfully classified into different superfamilies.

II. LITERATURE REVIEW

In the past, several approaches were developed to classify protein sequences into different superfamilies. Currently BLAST, FASTA and Hidden Markov Models (HMM) are the major methods used to find sequence similarity and are in use of the biological analysis of the available protein sequences. The BLAST or PSI-BLAST shows which part of the protein sequence best matches with the already present sequences in the database. BLAST employs nearest neighbor method to get a resemblance among sequences. Many computational intelligence approaches using machine learning and pattern recognition also have been developed. These use...
popular classifications and clustering algorithms like, Neural Network (NN), Decision Tree (DT), Fuzzy ARTMAP Model, Genetic Algorithms (GA), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), K Nearest Neighbor (KNN), and K-Means etc. A brief review of some most recent techniques is presented below. Jong et al. [1] proposed a new feature extraction method based on the Position Specific Scoring Matrix (PSSM) to classify yeast protein sequences data into different families. The authors used four different encoding methods to represent each protein sequence. Different criteria like accuracy, specificity, recall (sensitivity), precision, F-measure and correlation coefficient are used for the evaluation of encoding and classification algorithms. The maximum accuracy obtained is 72.5%. Bandyopadhyay et al. [2] extracted features from protein sequences using a 1-gram method. A variable length fuzzy genetic clustering algorithm is used to obtain a number of prototypes for each superfamily. Three superfamilies: Globin, Ras, Cytosine were used in the experiment. The overall accuracy result obtained is 81.3 %. The authors in [3] used physiochemical properties as features to the feed-forward and probabilistic neural network for protein superfamily classification. Mansoori et al. [4] extracted features from the protein sequence by using 2-gram technique and 2-gram exchange group. Some best features were selected through a feature ranking method. Fuzzy rules sequences that can effectively classify protein sequences into different functional classes. Swati et al., [8] adopted adoptive multi objective genetic algorithm (AMOGA) to optimize the structure of radial basis function network (RBFN). A 2-gram encoding method is used for the representation of protein sequences and principal component analysis (PCA) for feature selection. The proposed technique was further used for classification of protein sequences into superfamilies. Saha et al. [9] presented a review of three different classification models such as neural network, fuzzy ARTMAP and Rough set classifier. The authors developed its own GUI based model using JAVA, to reduce the computational overheads and increase the classification accuracy. The proposed technique reduces the feature size and shows good results of subcellular localization datasets. From the literature, it is found that there are sequences that even have sequence similarity among them, but the structure and function of these protein sequences are not the same.

III. METHODOLOGY

Following are different modules or phases of the proposed methodology.

For Example: Let \( i \) be one of the amino acid.

\[
\]

The proposed distance based feature encoding technique finds all the occurrence positions of each amino acid in a sequence. For example: \( p^1_i \) shows the first occurrence position of amino acid \( i \) in a sequence, \( p^2_i \) is the second occurrence position and
\( p^n_i \) is the last occurrence position of the amino acid \( i \) in a sequence, where \( n_i \) is the maximum number of times the specific symbol \( i \) occurs in the sequence. The vector of positions for the amino acid \( i \) in the given sequence is obtained as it follows:
\[
 p^i = (p^i_1, p^i_2, ..., p^i_{n_i})
\]

From the above vector of positions, two features are calculated:

**the mean:**
\[
 \mu^p_i = \frac{1}{n_i} \sum_{j=1}^{n_i} p^i_j
\]

**the variance:**
\[
 \text{var}_p^i = \frac{1}{n_i} \sum_{j=1}^{n_i} (p^i_j - \mu^p_i)^2
\]

Next, we find the differences between consecutives positions of the amino acid \( i \) in the given sequence.
\[
d^i_j = p^i_{j+1} - p^i_j \text{ for all } 1 \leq j \leq n_i - 1
\]

From the vector of distances of consecutives positions, two more features are obtained by calculating the mean and the variance.
\[
 \mu^d_i = \frac{1}{n_i-1} \sum_{j=1}^{n_i-1} d^i_j
\]
\[
 \text{var}_d^i = \frac{1}{n_i-1} \sum_{j=1}^{n_i-1} (d^i_j - \mu^d_i)^2
\]

Similarly, we find positions and distance vectors for other amino acid symbols in the given example sequence. To extract the feature vector from the above calculated positions and distance vector for each amino acid symbol:

\[
d^A_{n_i} = p^A_{n_i} - p^A_{n_i-1} = 16 - 6 = 10
\]

The calculation of distance at decomposition level 3 obtained from the successive distances of the previous level for symbol A is:
\[
dd^A_1 = d^A_2 - d^A_1 = 1 - 2 = -1
\]
\[
dd^A_2 = d^A_3 - d^A_2 = 2 - 1 = 1
\]
\[
dd^A_{n_i-2} = d^A_{n_i-3} - d^A_{n_i-2} = 10 - 2 = 8
\]

Similarly, we find positions and distance vectors for other amino acid symbols in the given example sequence. To extract the feature vector from the above calculated positions and distance vector for each
symbol and for each sequence, the mean and the variance is taken of the above decompositions respectively.

\[ \text{mean}(p_1^A, p_2^A, p_3^A, p_4^A, p_5^A); \]
\[ \text{var}(p_1^A, p_2^A, p_3^A, p_4^A, p_5^A); \]
\[ \text{mean}(d_1^A, d_2^A, d_3^A, d_4^A); \]
\[ \text{var}(d_1^A, d_2^A, d_3^A, d_4^A); \]
\[ \text{mean}(dd_1^A, dd_2^A, dd_3^A); \]
\[ \text{var}(dd_1^A, dd_2^A, dd_3^A); \]

Extracted feature vector for symbol A from the given sample sequence is:

<table>
<thead>
<tr>
<th>( \mu_p^A )</th>
<th>( \text{var}_p^A )</th>
<th>( \mu_d^A )</th>
<th>( \text{var}_d^A )</th>
<th>( \mu_{dd}^A )</th>
<th>( \text{var}_{dd}^A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>34.50</td>
<td>3.75</td>
<td>17.58</td>
<td>2.66</td>
<td>22.3333</td>
</tr>
</tbody>
</table>

Similarly, the feature vector for other symbols can be constructed for each sequence. Using the decompositions of position and distance up-to level 3 as shown above; we obtained 6 features for each amino acid alphabet in each sequence. Any unknown sequence can be encoded using the proposed encoding technique in a straightforward way with a minimum number of features. Because, there are 20 amino acid symbols, as a result the total number of features becomes 120. While using above decompositions up-to level 4, the total number of features will be 160. After encoding of the protein sequences, the extracted feature vector space has a high dimensional data. to reduce the feature size, a subset of more informative features is selected. The selection of features is performed using the statistical metric, only 20 features are used. The total number of features is reduced to 20 using the statistical metric with decompositions. The total number of features obtained by applying the proposed feature encoding technique with decompositions is 80.

The performance of the proposed feature encoding and selection technique. The average experimental results are obtained over tenfold cross validations using feature encoding technique with decompositions. The total number of features obtained by applying the proposed feature encoding technique with decompositions up-to level-2 is 80. Using the feature selection statistical metric, only 20 most prominent features are selected to classify the sequences into the respective superfamilies. Features extracted in this experiment are 120; the size of these features is reduced to 20 using the statistical metric based feature selection technique described in the previous section.

\[ \text{Accuracy/precision} = \frac{TP + TN}{TP + TN + FP + FN} \]
\[ \text{Sensitivity/recall} = \frac{TP}{TP + FN} \]
\[ \text{Specificity} = \frac{TN}{TN + FP} \]
\[ \text{Precision} = \frac{TP}{TP + FP} \]
\[ \text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]
\[ \text{MCC} = \frac{(TP + TN) - (FP + FN)}{\sqrt{(TP + TN)(TP + FN)(TN + FP)(TN + FN)}} \]

where,

\( TP = \) Number of true positive, \( TN = \) Number of true negative, \( FP = \) Number of false positive, \( FN = \) Number of false negative.

V. EXPERIMENTAL RESULTS

For the experiments, we considered the Yeast protein sequences downloaded from UniProtKB database. Three functional categories used in the experiments are: metabolism, transcription and cellular transport, transport facilities. These families play an important role to perform various critical functions inside cell in the human body. The selected sequences are chosen randomly. For the simulation of the results, MATLAB version 7.12 on a 3.0 GHz dual core CPU with 4 GB RAM has been used. The detail of the dataset used in the experiments is shown below.

<table>
<thead>
<tr>
<th>Family Name</th>
<th>Number of Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolism</td>
<td>752</td>
</tr>
<tr>
<td>Transcription</td>
<td>520</td>
</tr>
<tr>
<td>Cellular Transport</td>
<td>563</td>
</tr>
<tr>
<td>Total</td>
<td>1837</td>
</tr>
</tbody>
</table>

For validation of the experimental results, a tenfold cross validation technique is used. Further, in each fold the sequences that are not used during the system training. The following set of statistical performance measure metrics are used for validation and evaluation of the proposed results [1], [5].

VI. DISCUSSION AND EVALUATION OF RESULTS
Finding the protein structure, function and evolutionary information from the primary protein sequence is a critical problem in Bioinformatics. This study is based on the classification of three functional families of the Yeast proteins. The sequence length of the selected dataset varies from few amino acids to thousands amino acids. The sequences in the Yeast proteins have lots of sequence variations, therefore shows difficulties, when classifying sequences into functional families. In this study, we proposed a new distance based feature encoding technique to find the distance between the amino acids in a sequence. A statistical metric for feature subset selection is employed to select the minimum number of features to represent the protein sequence. The well-known classifiers: naive Bayes, decision tree, neural network, random forest and multilayer perceptron classification algorithms are adopted to classify the data extracted from the selected families. The performance of the proposed technique is validated using classifiers performance measure metrics like true positive rate (TPR), false positive rate (FPR), specificity, sensitivity, recall, F-measure and Mathews Correlation Coefficient (MCC). MCC is also an effective criterion for assessing the performance of feature encoding technique. The value of MCC ranges from +1 to -1, where +1 shows best performance whereas -1 represents poor performance. The comparison of results in the figure 5, 6 and 7 reveals efficient and accurate analysis and modeling of the unknown protein sequences.

Figure 5 Comparison of Performance Measure Metrics with Decompositions up-to Level-1

Figure 6 Comparison of Performance Measure Metrics with Decompositions up-to Level-2

Figure 7 Comparison of Performance Measure Metrics with Decompositions up-to Level-3

Figure 8 Comparisons of the Best Classification Accuracy Results obtained from Different Classifiers on Yeast Protein Sequences

average accuracy obtained by applying the decision tree is 91.2%. The previous accuracy obtained on same dataset is 72.5% [1]. Figure 8 depicts the overall best classification accuracy results obtained using the proposed and previous technique.
VII. CONCLUSION

In this paper, a computational technique for protein sequence classification using distance based feature encoding technique is proposed to represent the variable length protein sequence. Using this technique, the features are extracted by finding the distance of each amino acid from the first amino acid in a sequence at different level of decompositions. The feature size is then reduced by using a statistical metric by representing a protein sequence with the minimum number of informative features. The encoding technique is validated with five different classification algorithms which exhibit significant improvements in the classification accuracy, specificity, precision, recall and F-measure. The result indicates that the features derived from the distance based feature encoding method are suitable for highly accurate protein sequence classification. With this technique remotely homologous sequence data can be represented in a way that enables classification algorithms to classify them in an optimal manner. In future, the proposed technique can be extended to higher level of decompositions to achieve more improved result.

REFERENCES


Authors

G. Bhanu Priya received her B.Tech degree in Computer Science and Engineering from JNTU-K and currently pursuing M.Tech degree in Computer Science and Technology with specialization in Bioinformatics from Andhra University.