Classification of GPCRs Proteins Using a Statistical Encoding Method

Muhammad Javed Iqbal[1], Ibrahima Faye[2], Brahim Belhaouari Samir[3]
[1] Department of Computer Science, University of Engineering and Technology Taxila, Pakistan
[2] Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Malaysia
[3] College of Science, University of Sharjah, United Arab Emirates
javed1797@hormail.com, ibrahima_faye@petronas.com.my, sbelhaouari@sharjah.ac.ae

***Abstract*—Classification of G protein-coupled receptors (GPCRs) according to their functions is an ongoing area of research which is helpful for the pharmaceutical industry in the development of drug targets for major diseases. Currently, more than 40% drugs in the market target GPCRs. The experimental methods of determining their function are very expensive and time consuming. Due to a rapid and constant increase in the GPCRs proteins in the public databases, it is extremely important to develop computational techniques that lessen the gap between the sequenced proteins and proteins with known functions. In this paper, a statistical method was utilized to encode proteins sequences. The encoding technique considers various distances for an amino acid in a sequence at different levels of decompositions. The Neural Network and Support Vector Machines classifiers were compared on 2 well-known GPCRs datasets. The results showed that better performance is achieved using neural network classifier. The classification accuracies were in the range of 94 to 98%.**

Keywords— Bioinformatics; GPCRs; Distance-based Encoding; Superfamily; Performance Measurement;

I. INTRODUCTION

Machine learning techniques have been used in the development of various intelligent systems useful in the prediction of functions and structures of newly discovered biological data such as proteins [1]. The emergence of human genome project generated an enormous amount of multifaceted biological data comprising of Deoxyribonucleic acid (DNA) and proteins [2]. Proteins are the basic substance of life. A proteins family classification problem can be defined as a group of proteins that show significant sequence similarity in them and thus have a similar structure and function [3, 4]. The experimental methods for the annotation of sequences are very time consuming and intensive. Alternately, alignments based techniques also find difficulties in the accurate determination of the functions of unknown proteins. To lessen the issues in the usage of traditional methods, numerous computational intelligence techniques have been developed to effectively solve challenging problems such as the structure and function prediction of a newly discovered protein, protein-protein interactions, mass spectrometry based proteomics, genes regulatory networks, biomarker detection from gene expression data, meta-genomics etc. [5-8].

Among different functionally enriched protein superfamilies, G protein-coupled receptors (GPCRs) is found to be the largest protein superfamily in the human genome. GPCRs control several fundamental physicochemical processes enclosed in a cellular signaling network, such as smell, taste, vision, secretion, neurotransmission, metabolism, cellular differentiation and growth, inflammatory and immune response [9, 10]. These also regulate pathways and mechanisms that are helpful in performing many essential functions in several distinct species, including humans. The GPCRs performs main role in sensing various kinds of signals extending from visual to olfactory [11]. These receptors are also a main target in the design and development of new drugs, since approximately 40% to 60% of the current drugs in the market target GPCRs. A GPCR can interact with one or more G-proteins; a problem encountered is the prediction of the coupling specificity of GPCRs to the G-protein superfamily classes [12].

The GPCRDB database arranges a GPCR superfamily into a hierarchy of 'families, 'sub-families', 'sub-sub-families', and 'types' [13]. Class A is the largest family in the hierarchy of potential GPCR classes, which covers more than 80% of overall human GPCRs.

In this paper, the objective is the classification of distantly related GPCRs of three families, Class A, Class B and Class C. Additionally, it also classifies subfamilies of Class A protein sequences into their respective family based on their primary structure. The statistical distance-based encoding technique captures the statistical significance of amino acids present in the sequences. The neural network (NN) and support vector machines (SVMs) classifiers were compared. These results were also compared with the existing popular techniques of GPCRs classification.

The rest of the paper is organized as follows. Section II summarized the literature review on GPCRs protein sequence classification. The proposed classification framework is illustrated in Section III. Section IV presents the details of the experiments. The results are discussed and evaluated in section V and the conclusion is presented in the section VI.

II. RELATED WORK

Several sequence similarity search based computational methods have been introduced to determine the function of unknown protein sequences. Two main categories of these methods are alignment-based and alignment-free. The alignment-based methods match the example sequence with all the sequences in a database for a particular family and produces a matching score against each comparison. The sequence is assigned to the family, with which the query sequence has obtained the highest similarity score. The popular alignment-based techniques include BLAST, PSI-

BLAST, FASTA, etc. [14]. However, these alignment-based methods are found to be inadequate for comprehensive functional identification of GPCRs, since the sequences in any GPCRs families are highly divergent [9]. The brief description about some most recent and popular techniques of GPCRs classification is presented subsequently.

Zhou et al. described that there exists three types of approaches for GPCRs classification, which are based on proteochemometrics, sequence similarity search and statistical and machine learning based methods [9]. In the first step of feature selection, minimum redundancy maximum relevance (mRMR) was utilized, while in the next step, a genetic algorithm was employed to further select the most important features. The fitness function in genetic algorithm utilizes two objective functions: maximize the classification accuracy and reduce the feature subset size. The classification accuracy was achieved in the range from 80-95%.

Nascimento et al. presented to extracts 120 features based on the patterns of regular expressions [12]. The classification mechanism is completed in two phases. In the first phase of the classification, chi-square statistics was implemented to evaluate the effectiveness of the extracted patterns. In the second phase, a more sophisticated machine learning approach using SVM was employed for the classification of sequences into the respective GPCRs families. Three types of feature matrices were obtained by applying a String Matching algorithm on each sequence and patterns. The data used in the experiments were taken from GPCRDB, which comprised of 769 GPCR sequences and 2565 non- GPCRs. The final feature matrix size was 3334 x 120. Using the Chi-Square, it was proved that the occurrence frequency of patterns in GPCR and non-GPCR are statistically different. An average AUC of approx. 98% was achieved on the GPCR dataset.

Cobanoglu et al. proposed GPCRBind method to classify GPCRs Class A subfamily protein sequences [11]. During Distinguishing Power Evaluation (DPE) step, most distinguished motifs were found from the training data for the purpose of classification. The main task here is to repeatedly construct decision trees from arbitrarily divided training and data to search for the motifs that occurs repeatedly in every decision tree. The ultimate objective of DPE step is the evaluation of the motifs instead of constructing a classifier. The average classification accuracy obtained using the GPCRBind was 90.7%. To obtain better classification results of GPCRs classification, the experiments could be performed using an efficient representation of protein sequences.

Strope et al. investigated the performance, advantages and limitations of both the alignment-based and alignment-free techniques [15]. Eight different classifiers were utilized in the classification: SAM, SVM_Fisher, SVM_Pairwise, SVM_A(rbf), SVM_AA(pol), SVM_AA(sig) and DT. In the experiments average with-in class A classification accuracy on different classifiers was 92%. Whereas, slightly lower accuracy rates (approx. 85%) were observed against non-Class A datasets. The best classification accuracy achieved using SVM_pairwise was approximately 90% which is similar to other classifiers: SVM_AA(rbf) or SVM_AA(pol). The amino acid composition-based classifiers (SVM_AAs and DT) produced better results than SAM and SVM_Fisher. The accuracy obtained with SVM_AAs was approximately 90% or higher. The further intensive experimentation is desired by using different methods of sequence encoding.

The analysis of previous studies has shown that a variety of computational techniques were introduced to classify GPCR protein sequences into different families/subfamilies by finding similarities between them [8]. However, the fast increase in the data has complicated the implementation of such techniques. These techniques undergo performance degradation due to the lack of an appropriate sequence encoding mechanism in order to represent the complex characteristics of many protein sequences. The traditional alignment-based approaches such as BLAST, FASTA and Hidden Markov Models also found difficulties in the annotation of unknown sequences, which have very low or weak similarity between them [6, 8]. Thus, there is a need of a highly accurate and efficient system that can classify the newly discovered protein sequences into existing families in a very short period of time. The implemented framework of sequence encoding and classification would eliminate the major limitations found in the existing techniques. The protein classification framework would be helpful in the exploration, interpretation, analysis of GPRCs proteins.

III. Proposed Framework for Protein Sequence Classification

The framework presented in Figure 1 has been utilized for the classification of GPCRs protein sequences in families and subfamilies based on the primary sequence information solely. The framework is comprised of different phases, in every phase some essential steps involved in machine learning-based models are performed.

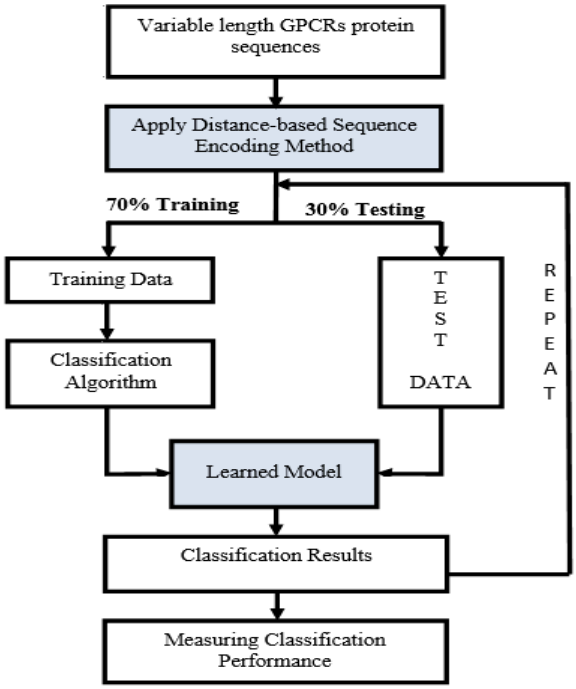

Figure 1 Framework for the classification of GPCRs proteins

The exclusive GPCRs protein sequences belonging to three families: Class A, Class B, and Class C were taken from the GPCRDB or UniProtKB database [16, 17]. For the validation of the proposed work, the data is divided into training and testing patterns. During the training, 70% of sequences were chosen to train the classifier. After successful training of the system, the remaining 30% sequences were incorporated for the testing. In both the training and the testing phases, the first step was the encoding of each amino acid sequence. In the encoding scheme, the sequence was represented using a fixed length feature vector containing a numeric value for each of the amino acids.

Following the selection of specific data in the experiments, a statistical distance-based sequence encoding technique was employed to obtain the statistical information about each sequence. The distance based feature encoding technique finds all the amino acids frequency and their occurring positions in a sequence at different levels of decomposition [18]. Using this technique, each protein is represented with the statistically enriched protein's primary sequence-based features. The objective of the feature extraction is to recognize protein sequence by the measurement of similar values that are present in the homologous protein sequences.

This technique of sequence encoding in the first step extracts the occurrence frequencies of the twenty amino acids in a sequence. In the second step, which is known as the 1st level decomposition, it computes the occurrence positions of each amino acid and exchange groups in a protein sequence. In the 2nd level decomposition, it computes a distance vector from the 1st level decomposition that finds the mutual distance between the successive positions of each amino acid in each sequence. Similarly, in the 3rd level decomposition, it computes the distance between the successive elements of the distance vector, which was obtained at 2nd level decomposition. The number of decomposition levels may increase, which depends upon the complexity of the sequence data. The sequence length of each protein sequence was also considered as one of the features in the proposed sequence encoding algorithm. Consequently, a total of 177 features were extracted from each protein sequence up to 3rd level decomposition. Similarly, for 4th level decompositions, the total number of extracted features would be 229. Additional features can be added by using the distance vectors with the higher levels of decomposition features depending on the complexity of the protein sequences. The major advantage of the utilized sequence encoding algorithm is that it maintains sequence order information of each amino acid in a sequence.

IV. MATERIALS

In order to validate the proposed GPCRs classification framework, protein sequences were taken from GPCRDB database. Two datasets were constructed; Dataset 1 was comprised of three families: Class A, Class B and Class C. While, the Dataset 2 was consisted of three subfamilies within Class A, family: Dopamine, Serotonin, and Chemokine. The total numbers of exclusive sequences considered in both datasets were 1619 and 1306 respectively. The sequence detail of data used in the experiments is shown in Tables 1 and 2.

TABLE 1. DETAILS OF DATASET 1 USED IN THE EXPERIMENTS

Family Name	Number of Sequences
Class A	718
Class B	405
Class C	709
Total Sequences	1832

TABLE 2. DETAILS OF DATASET 2 USED IN THE EXPERIMENTS

Subfamily of Class A	Number of Sequences
Dopamine	228
Serotonin	503
Chemokine	575
Total Sequences	1306

Furthermore, in the experiments, the neural network and support vector machines (SVMs) classifiers were implemented with the following parameter configurations shown in Table 3 and 4.

TABLE 3. PARAMETER CONFIGURATION OF THE NEURAL NETWORK CLASSIFIER

S. No.	Parameter's Name	Value
1	Type of model	Multilayer perceptron Network (MLP)
2	Training Algorithm	Scaled conjugate gradient back propagation algorithm
3	Number of Layers	3
4	No. of Input Neurons	The number of input neurons depends on the size of the features selected in the experiment.
5	No. of hidden layers	1
6	Number of hidden neurons per layer	10-50
7	Hidden layer activation function	Sigmoid $f(x) = \frac{1}{1+e^{-x}}$
8	Output layers	The number of output layers depends on the number of protein families considered in an experiment.
9	Output layers activation function	Linear
10	Learning rate (η)	0.1-0.9
11	Momentum (Υ)	0.1-0.9
12	No. of epochs to train through	500-1000
13	Error function	Mean Square Error (MSE)
14	Initial weights of the neural networks	Within [-0.1,0.1]

15	Validation method	Cross validation
17	Cross validation method	Training 70%, Testing 30%

TABLE 4. PARAMETER CONFIGURATION OF SUPPORT VECTOR MACHINES (SVMS) CLASSIFIER

S. No.	Parameter's Name	Value
1	SVM type	C-SVC (Classification)
2	Type of analysis	Classification
3	Gamma	0.0-0.5
4	Eps	0.001
5	Kernel type	Radial basis kernel function Exp (-gamma*(x-y)(x-y))
6	Degree of kernel	3

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this paper, a statistical distance-based encoding method is used to represent a variable length lower similarity protein sequence with a fixed size vector. The protein sequences of three GPCRs families were investigated in the experiments due to their importance in the pharmaceutical industry. The sequences belong to the families of GPCRs having very weak sequence similarity which exhibits difficulties in classifying them into different classes. Each phase of the proposed framework was employed to perform a specific task of classifying protein sequences into their corresponding families/subfamilies. This mechanism seems to be very helpful in the annotation of unknown protein sequences. In the experiments, neural network and SVMs classifiers were used for classification. After the successful classification of sequences, the classification results also have been verified using GPCR database. The classification results obtained were the average results of various rounds of the classifier. Different performance measurement metrics such as accuracy, true positive rate (TPR), false positive rate (FPR), specificity, sensitivity, recall, F-measure and Mathews Correlation Coefficient (MCC) were investigated for performance evaluation of the protein sequence classification technique. The confusion metrics obtained with the MLP neural network and support vector machines classifiers are shown in Table 5-8 respectively.

TABLE 5 CONFUSION MATRIX USING NEURAL NETWORK CLASSIFIER ON DATESET 1

Data\results	Class A	Class B	Class C
Class A	703	9	6
Class B	7	393	5
Class C	5	9	695

TABLE 6 CONFUSION MATRIX USING NEURAL NETWORK CLASSIFIER ON DATASET 2

Data\results	Dopamine	Serotonin	Chemokine
Dopamine	202	11	15
Serotonin	8	490	5
Chemokine	12	10	553

TABLE 7. CONFUSION MATRIX USING SVMS CLASSIFIER ON DATESET 1

Data\results	Class A	Class B	Class C
Class A	685	20	13
Class B	16	378	11
Class C	11	17	681

TABLE 8. CONFUSION MATRIX USING SVMS CLASSIFIER ON DATASET 2

Data\results	Dopamine	Serotonin	Chemokine
Dopamine	190	22	16
Serotonin	19	471	13
Chemokine	20	14	541

TABLE 9 COMPARISON OF PERFORMANCE MEASURE METRICS ON BOTH DATASETS USING NEURAL NETWORK CLASSIFIER

Performance Measure Metrics	Dataset 1	Dataset 2
Accuracy	97.9%	94.0%
Specificity	98.9%	97.6%
Sensitivity	97.9%	94.1%
Recall	97.9%	94.1%
F-Measure	98.0%	94.3%
MCC	96.9%	91.9%

TABLE 10 COMPARISON OF PERFORMANCE MEASURE METRICS ON BOTH DATASETS USING SVMS CLASSIFIER

Performance Measure Metrics	Dataset 1	Dataset 2
Accuracy	95.3%	90.3%
Specificity	97.6%	96.0%
Sensitivity	94.9%	90.4%
Recall	94.9%	90.4%
F-Measure	94.8%	90.3%
MCC	92.4%	87.3%

Table 5-8 demonstrates the confusion matrices obtained with both datasets using the neural network and SVMs classifiers. From these confusion matrices, we can compute the performance measured metrics, which have been shown in Table 9 and 10. The performance measured metrics includes classification accuracy, specificity, sensitivity, recall, f-measure and MCC. The best accuracy results on Datasets 1 and

2 were in the range of 94% to 97.9%, which shows some improvement from the accuracy results which were in the range of 90.7% to 95% in the previous studies [9, 11]. However, in this paper, comparison with the alignment-based methods is not included, as these do not produce comparable results. The proposed classification framework is able to classify distantly related GPCRs protein sequences into their corresponding families and subfamilies solely based on the information extracted from their amino acid sequence.

VI. Conclusion

The exploitation of computational intelligence techniques have shown promising results in the analysis and modeling of biological data such as protein sequences. In this paper, a statistical distance-based encoding method was employed in the proposed framework to encode protein sequences in the form of a numeric feature vector. The experiments were carried out on three families and subfamilies of GPCRs protein sequences. Using this classification technique, the unknown protein sequence was successfully annotated with an improved accuracy. In the experiments, the MLP neural network classification algorithm has shown substantial improvement in the classification accuracy, specificity, precision, recall and F-measure. A sufficient amount of increase in the classification accuracy was found (i.e. approximately 3-4%). In the future, this framework can be extended to classify groups of protein sequences, which are involved in various kinds of proteomics' diseases.

References

[1] L. Kurgan and Y. Zhou, "Machine learning models in protein bioinformatics," *Current Protein and Peptide Science,* vol. 12, p. 455, 2011.

[2] D. R. Bentley, "The human genome project - An overview," *Medicinal Research Reviews,* vol. 20, pp. 189-196, 2000.

[3] S. Bandyopadhyay, "An efficient technique for superfamily classification of amino acid sequences: Feature extraction, fuzzy clustering and prototype selection," *Fuzzy Sets and Systems,* vol. 152, pp. 5-16, 2005.

[4] S. V. a. S. K. R. D. of, "Two-Stage Approach for Protein Superfamily Classification," *Hindawi Publishing Corporation,Computational Biology Journal,* 2013.

[5] N. Goel, S. Singh, and T. C. Aseri, "A Review of Soft Computing Techniques for Gene Prediction," *ISRN Genomics,* vol. 2013, p. 8, 2013.

[6] M. N. Davies, A. Secker, A. A. Freitas, J. Timmis, E. Clark, and D. R. Flower, "Alignment-independent techniques for protein classification," *Current Proteomics,* vol. 5, pp. 217-223, 2008.

[7] J. S. Bernardes and C. E. Pedreira, "A review of protein function prediction under machine learning perspective," *Recent Patents on Biotechnology,* vol. 7, pp. 122-141, 2013.

[8] S. Saha and R. Chaki, "A brief review of data mining application involving protein sequence classification," *International Journal of Database Management Systems,* vol. 4, pp. 469-477, 2013.

[9] X. Zhou, Z. Dai, and X. Zou, "Classification of G-protein coupled receptors based on support vector machine with maximum relevance minimum redundancy and genetic algorithm," *BMC Bioinformatics,* vol. 11, 2010.

[10] Y. Huang, J. Cai, L. Ji, and Y. Li, "Classifying G-protein coupled receptors with bagging classification tree," *Computational Biology and Chemistry,* vol. 28, pp. 275-280, 2004.

[11] M. C. Cobanoglu, Y. Saygin, and U. Sezerman, "Classification of GPCRs using family specific motifs," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 8, pp. 1495-1508, 2011.

[12] F. Nascimento Jr, I. R. Tsang, and G. D. C. Cavalcanti, "A SVM for GPCR protein prediction using pattern discovery," Barcelona, 2008, pp. 429-434.

[13] F. Horn, E. Bettler, L. Oliveira, F. Campagne, F. E. Cohen, and G. Vriend, "GPCRDB information system for G protein-coupled receptors," *Nucleic Acids Research,* vol. 31, pp. 294-297, 2003.

[14] W. R. Pearson, "Rapid and sensitive sequence comparison with FASTP and FASTA," *Methods in Enzymology,* vol. 183, pp. 63-98, 1990.

[15] P. K. Strope and E. N. Moriyama, "Simple alignment-free methods for protein classification: A case study from G-protein-coupled receptors," *Genomics,* vol. 89, pp. 602-612, 2007.

[16] http://www.gpcr.org/7tm/.

[17] M. N. Davies, A. Secker, A. A. Freitas, M. Mendao, J. Timmis, and D. R. Flower, "On the hierarchical classification of G protein-coupled receptors," *Bioinformatics,* vol. 23, pp. 3113-3118, 2007.

[18] M. J. Iqbal, I. Faye, A. M. D. Said, and B. B. Samir, "Computational Technique for an Efficient Classification of Protein Sequences With Distance-Based Sequence Encoding Algorithm," *Computational Intelligence,* pp. n/a-n/a, 2015.