A Survey on Protein Sequence Classification with Data Mining Techniques

C. Nandini[1], I. Laurence Aroquiaraj[2]

ABSTRACT: *Today biological data like protein sequence and DNA sequence is increasing at vast rate due to improvement of technologies. From this vast data we have to derive the hidden knowledge so that it is used in wide range of areas to design drug, to indentify disease, and in classification of protein sequence etc. Today most of the researchers show their keen interest in finding unknown protein sequence and classify them in already defined protein sequence family. Data mining consist of number of techniques used to extract the knowledge for this vast biological data. One of the important areas of research is to classify protein sequence into different families, classes and sub classes. Feature selection also plays an important role in sequence classification as it increase the accuracy of classifier SVM. A number of different classification techniques or algorithms have been proposed by different researchers to classify protein sequences. This paper explains various techniques used by different researches in classifying the proteins and also provides an overview of different protein sequence classification methods.*

Keywords: Protein sequence, SVM classifier, DNA sequence, Feature selection.

1 INTRODUCTION

BIOINFORMATICS is the field which consist of more than one branch of learning. As It is the field of science which combines together computer science, statistics, mathematics, and engineering to analyse and interpret biological data, it is called as interdisciplinary field. It develops a methods and software tools for understanding biological data. It is the application of computer technology to the management of biological information. The main objectives of bioinformatics were storing, extracting, organizing, analysing, interpreting and utilizing the information from the biological sequences and molecules. Molecular biology is one of the areas of bioinformatics that use techniques of bioinformatics such as image and signal processing that allow extraction of useful results from large amount of raw data. It also plays an important role in the analysis of gene expression, protein expression, protein sequence and regulation. The genomics and proteomics generated a vast amount of biological data like DNA sequence, protein sequence and other data which are available for prediction. These data not only grow due to data sample that are available in our daily life but also due to increasing number of candidate features of various information. There are some classification techniques are there that specially used for protein sequence to extract specific features from sequences and these features are depend on functional and structural properties of 20 amino acids and these features has to be reduced so that it can increase the efficiency of the classifier which is used to classify the protein sequence. This paper explains various feature extraction techniques and sequence classification techniques used by different researchers and also provide an overview of different protein sequence classification methods. The remainder of the paper is organized as follows: Section 2 provide with basic of sequence classification. In section 3, we discuss various methods developed by researches related to protein sequence. Section 4 with challenges in research and section 5 presented with conclusion.

2. BASIC OF SEQUENCE CLASSIFICATION

2.1 Sequence Classification Methods:

The sequence classification methods can be divided into three large categories. (i) The first category is feature based classification, which transforms a sequence into a feature vector and then apply conventional classification methods. Feature selection plays an important role in this kind of methods. (ii) The second category is sequence distance based classification. The distance function which measures the similarity between sequences determines the quality of the classification significantly. (iii) The third category is model based classification, such as using hidden markov model (HMM), neural network, SVM and other statistical models to classify sequences.

2.2 Protein Sequence Database:

A number of protein sequence databases have been available. To conduct their research, the researchers can

download the sequences of various kinds of proteins datasets from the databases were it is stored. Some of the well known databases that available are The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), UCI Machine Learning Repository, Protein Data Bank (PDB), Universal Protein Resource (UniProt), Swiss-Port, SCOP etc.

An example of protein sequence of Gallus dataset is given below:

MLGKNDPMCLVLVLLGLTALLGICQGGTGCYGSVSRIDTT
GASCRTAKPEGLSYCGVRASRTIAERDLGSMNKYKVLIKRV
GEALCIEPAVIAGIISTESHAGKILKNGWGDRGNGFGLMQV
DKRYHKIEGTWNGEAHIRQGTRILIDMVKKIQRKFPRWTR
DQQLKGGISAYNAGVGNVRSYERMDIGTLHDDYSNDVVA
RAQYFKQHGY

3. RELATED WORKS:

An overview of some of the classification techniques that have been developed to classify the protein are given below:

3.1. Neural Network Model:

Cathy wu, Michael Berry, Sailaja Shivakumar and Jerry Mclarty proposed neural network method for protein sequence classification based on already known structure/function of protein. The methods contains three-layers input, hidden and output layer, input layer is used to represent sequence data, the hidden layer is to get information in non-linear parameters, and the output layer to represent sequence classes. The sequences are encoded in neural input vector by hashing method that counts the occurrence of n-gram. The SVD (Singular Value Decomposition) is used to reduce the size of n-gram vectors and to extract semantics from the n-gram patterns. The SVD method was evaluated using many different n-gram vectors. The SVD computation can reduce the size of the network (i.e., input vector and weight matrix) by tens and hundreds of fold and improved the classification accuracy of the network. The SVD method also applies to nucleic acid sequences with very good results. Then a full-scale protein classification system has been implemented on a Cray supercomputer to classify unknown sequences into 3311 PIR (Protein Identification Resource) superfamilies/families. The accuracy level is close to 90%. The system could be used to reduce the database search time and is being used to help organize the PIR protein sequence database.

3.2. Word Segmentation Techniques:

Yang Yang1, Bao-Liang Lu1; 2_ And Wen-Yun Yang proposed a word segmentation technique that is a segmentation-based feature extraction method for protein sequence classification. The extracted features include selected words, i.e., substrings of the sequences, and also motifs specified in public database. They are segmented out and their occurrence frequencies are recorded as the feature vector values. Here to find the most discriminate features, text processing techniques were adopted and they selected high-frequency k-mers and conducted a segmentation to calculate the feature vectors for predicting protein subcellular localization. The experiments were conducted on two protein data sets, one is a set of SCOP families, and the other is G-Protein-Coupled Receptors (GPCRs family). The method proposed not only results in an extremely condensed feature set, but also achieves higher accuracy.

The method consists of three major steps. First step is building a dictionary by collecting all the 20 amino acids and certain number of meaningful k-mers according to some criterion. In the second step, segmentation algorithm is applied and corresponding matching process is conducted on the dictionary. In third step, sequences are converted into feature vectors based on the segmentation results. To built a dictionary, statistical methods were used on training data set. First a maximum word length *MaxLen* should be set, here number four is the best bound of k which specifies the set of k-mers from which words are selected and the most frequently presented strings are usually words, thus k-mers' appearance times are calculated and k-mers which widely presented are put into dictionary. Then to find the discriminating words tf - idf and entropy value 9 could be used.

Next segmentation is done by matching sequences with words in the dictionary; Maximum Matching (MM) algorithm is used. In the algorithm, given a dictionary *D* and a sequence *S* whose length is *N*, *segNo[1 _ _ _N]* and *wordLen[1 _ _ _N]* are two arrays recording the number of segments which have been identified from each amino acid to the end of *S*, and the length of word segmented from each position, respectively. They are initialized as zero arrays. *maxLen* stands for the maximum length of words. We selected N top ranked words

according to certain measure, and converted each protein to N-dimensional feature space. The LibSVM version 2.612 is chosen as our classifier. The proposed method achieves higher accuracy with 92.6 and 88.8% accuracy, respectively.

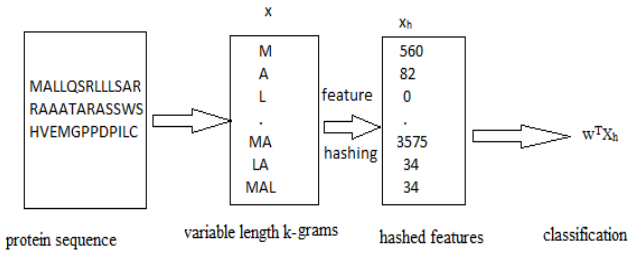
3.3. Feature Hashing Technique:

Cornelia Caragea, Adrian Silvescu, Prasenjit Mitra has used the feature hashing technique for protein sequence classification, where the original high dimensional space is "reduced" to lower-dimensional space by using a hash function to the features ie, mapping features to hash key, where multiple features can be mapped to the same hash key and aggregating their counts. Here the feature hashing with the "bag of k-grams" and found that this technique is an effective approach to reduce the dimensionality on protein sequence classification tasks. This new approach is also used for text classification, which is very inexpensive and very effective approach to reduce the number of features. Three issues were handled in this method:

1. What is the effect of hash size on performance of protein sequence classifiers that use hash features and what is the hash size which at which the performance stars degrading, due to hash collision.

2. How effective is the hashing on k-gram representation?

3. What is the performance of hash technique compare to feature selection?



To answer the above question, first pre-process the data by generating the k-grams for the collected sequence ie, generating all contiguous sub-sequences of length k, for various values of k, which was done by sliding a window of length k over sequences in each data set. If k-gram does not appear in the data set, it is not taken as feature. The feature hashing is applied in two setting, first generate all k-gram with fixed length ie, k=3, second generate all k-grams with various length k ie, k=1,2,3& 4. Thus this setting uses the union of k-gram ie, unigram, 2-gram, 3-gram or 4-gram is hashed into hash key with variable length. Then the Support Vector Machine (SVM) classifier on hash features for above both setting and investigates the influence of hash size on the performance of the classifier. It is noted the performance of SVMs trained on fixed length k-gram is worse than that of SVM classifiers trained on variable k-gram representation, it shows the highest performance with accuracy of 82.33%. The result of the experiment showed that feature hashing is an effective approach for reducing the dimensionality on protein sequence classification tasks.

3.4. Feature Selection Technique:

K.Radha, C.Akila, The feature selection is also known as attributing selection or variable subset selection. It does not alter the original representation of features. It is the process of selecting a subset of relevant features to use in model construction. Feature selection techniques remove the repeated and irrelevant features to improve the classification of sequences. It preserves the original semantics of the features. The goal of this technique is to find the best features subsets that improve the scoring function above. The accuracy obtained by using this method is 80.13%. The main reason for using feature selection technique is, it is a simplification models to make them easier to interpret by researchers, shorter training times and enhanced generalization by reducing overfitting. Feature selection techniques are used where there are many features and comparatively few samples. Latent Dirichlet Allocation (LDA), Probabilistic Latent Semantic Analysis (PLSA) and Latent Semantic Indexing (LSI) are some of the feature selection techniques designed to find the hidden topics.

3.5. Rough Set Classifier:

Ramadevi Yellasiri,C.R.Rao proposed Rought Set Classifier. It is the machine learning method which has the concept of set theory to make decision. The indiscernibility relation that produces minimal decision rules from training examples is the important notation in this method. To identify the set of feature, if-else rule is used on the decision table. This method can handle large volume of protein sequence for classification based on structural and functional properties of protein. It is a hybridized tool comprising sequence Arithmetic, Rough Set Theory and Concept Lattice which reduce the domain search space to 9% without losing the potentiality of classification of protein. The accuracy level of this classifier is 97.7%. Instead of classifying protein sequence into classes or sub classes, this model provide small know sequence from a long unknown protein sequence. Thus this model required extra time and space for further classification of the output sequence into classes or subclasses.

3.6. KNN and Support Vector Machine (SVM):

Amit kumar Banerjee & Vadlamani Ravi &U.S.N Murty & Sengupta proposed KNN and Support vector machine method to compute protein physicochemical and structural properties of protein sequences for classification of HK protein family and their variable influence for different bacterial genera. KNN is the simple but effective classification algorithm used here. KNN applies all the training inputs during testing phase of data classification. All input vectors in this algorithm are assumed to be in an n-dimensional Euclidean feature space, and classes are arranged based on the k-closest training inputs computed through Euclidean distances in the metric space. The Euclidean distance for $d(x_i, x_j)$, where x_i and x_j are two different input vectors in the feature space are computed by applying the equation:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{k=n} \left(ak(x_i) - ak(x_j)\right)^2}$$

SVM is a universal constructive learning procedure based on the statistical learning theory [1]. It is a machine learning technique based on the structural risk minimization principle that minimizing the true error and performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories and correctly classifies data points as much as possible. SVM first formulates a constrained optimization problem and then solves it using Constrained Quadratic Programming. Optimal hyperplane for classification is done using the dot product functions in feature space called kernels. The solution of the optimal hyperplane can be written as a combination of a few input points that are called support vectors. Given a set of points b $\in R$ with I=1…N. Each point x_i belongs to either of two classes with the label $y_i \in \{-1, +1\}$, The set S is linearly separable if there exist $w \in R^n$ and b $\in R$ such that,

$$\begin{cases} w^T\varphi(X_i) + b \geq +1, & y_{i=+1,} \\ w^T\varphi(X_i) + b \leq -1, & y_{i=-1,} \end{cases}$$

which is equivalent to

$$[\, w^T\varphi(X_i) + b] \geq 1, i = 1 \ldots\ldots, N.$$

The nonlinear function $\varphi(.)$ maps the input space to a high -dimensional feature space. In this feature space, the above inequalities basically construct a hyperplane $w^T\varphi(x) + b =$ 0 discriminating between both classes for a typical two-dimensional case. By minimizing $w^T w$, the margin between two classes is maximized. SVM and KNN algorithms were applied and the accuracy for both the algorithms were successful in attaining the mark of 91 % but tremendously varied in number of selected attributes which was 32 for SVM and only 18 in the case of KNN out of a total of 57 attributes.

3.7. N-gram patterns:

John K. Vries,[1*] Xiong Liu,[1,2] and Ivet Bahar[1] proposed N-gram patterns that provide useful means for classifying and characterizing proteins. An n-gram pattern (NP{n,m}) in a protein sequence is a set of n residues and m wildcards in a window of size n+m. Each window of n+m residues in a sequence is associated with a collection of NP{n,m} patterns based on the combinatorics of n+m objects taken m at a time. Any protein sequence can be parsed into a series of overlapping n-gram patterns by advancing a window of size n+m along the sequence. NP{n,m} patterns that are shared between sequences shows relationships. Theoretically, NP{4,2} patterns also shows the secondary structure propensity since it contain all possible n-grams for 1 <n< 4 and a window of 6 residues is taken to predict periodicities in 2 < n< 5 range. The accuracy of this method is 75.2%. In [10], the N-gram is used for feature extraction as the protein is made up of combination of 20 amino acids. The n-gram features are a pair of values (vi, ci), where vi is the feature i and ci is the counts of this feature in a protein sequence for i = 1, ⋯ , 20n. In general, a feature is the number of occurrences of an animal in a protein sequence. These features are all the possible combinations of n letters from the set. For example, the 2-gram (400 in total) features are (AA, AC, ⋯ , AY, CA, CC, ⋯ , CY, ⋯ , Y A, ⋯ ,YY). Consider a protein sequence example V AAGT V AGT, its corresponding 2-gram feature vector are {(V A, 2),(AA, 1),(AG, 2),(GT, 2),(TV, 1)}.Each sets of n-grams features, i.e., e_n and a_n, from a protein sequence will be scaled separately to avoid skew in the counts value using the formula: $\bar{x} = \frac{x}{L-n+1}$ where x represents the count of generic gram feature, x is the normalized x, which will be the inputs of the classifiers; L is the length of the protein sequence and n is the size of n-gram features.

3.8. Extreme Learning Machine (ELM):

Dianhui Wang, Guang-Bin Huang proposed Extreme Learning Machine (ELM) with sigmoidal activation function and Gaussian RBF kernel function for protein sequence

analysis which is recently developed algorithm . Protein sequences are classified into the same class if they have high homology in terms of feature patterns extracted through sequence alignment algorithms. A comparative study also conducted between ELM and the neural network classifier - Backpropagation Neural Networks. Results show that ELM needs up to four orders of magnitude less training time compared to BP Network. The classification accuracy of ELM is also higher than that of BP network. ELM is a new learning algorithm for Single-hidden Layer Feedforward neural Networks (SLFNs) which are either additive neurons or kernel based schemes. For additive neurons based SLFNs one may randomly choose and fix the input weights ie, linking the input layer to the hidden layer and linking the hidden layer to the output layer of SLFNs ie, analytically determine the output weights. ELM tends to have good generalization performance and can be implemented easily. The ELM algorithm is a adjustment methods that can avoid the difficulties like non-differential activation functions nor prevent the troubling issues such as stopping criteria, learning rate, learning epoches, and local minima.The three main steps involved in ELM algorithm can be summarized as:Given a training set $\aleph$ = $\{(xi, ti) | xi \in \mathbb{R}^n, ti \in \mathbb{R}^m, i = 1, \ldots, N\}$, activation function g or kernel function ϕ, and hidden neuron or kernel number $\tilde{N}$, *step* 1 Assign randomly input weight w*i* and bias *bi* for additive neurons or impact width μi and center σi, $i = 1,\ldots \tilde{N}$,. *step* 2 Calculate the hidden layer output matrix **H**. *step* 3 Calculate the output weight β: $\beta = \mathbf{H}\dagger\mathbf{T}$. The average best performance obtained by ELM with sigmoid and RBF kernels are 88.03% and 87.612% respectively while the average best generalization performance obtained by BP is 86.929%, thus ELM can obtain better generalization than BP.

3.9. Discriminative Descriptors Substitution Matrix (DDSM):

Rabie Saidi[1,2,3,4,] Mondher Maddouri [4,5,] and Engelbert Mephu Nguifo proposed a novel encoding method that uses amino-acid substitution matrix to define protein sequence classification based on similarities between motif. This paper deal with preprocessing of supervised classification. This method neglects the fact that some amino acids have similar properties and therefore they can substitute each other while changing neither the structure nor the function. The similarity between motifs is based on the similarity between 20 amino-acid .Since the mutation between 20 amino-acid are scored by 20*20 matrixes called substitution matrix. $S_m(X,Y)$ is the substitution score of the motif Y by the motif X and it is computed using formula:

$$S_{m(X,Y)=\sum_{i=1}^{k} S(X[i],Y[i])}$$

For better classification DDSM also consider combination with (DDSM &C4.5),(DDSM & SVM) and (DDSM & NB). The higher accuracy can be obtained by using combination DDSM & SVM) and (DDSM & NB) with accuracy level 82.3% and 85.9%.

TABLE 1

Summary of Above Techniques

S. No	Summary of Above Techniques		
	Techniques	**Accuracy Level**	**Classification Based On**
1.	Neural Network Model	90%	structure/function
2.	Word Segmentation Techniques	92.6 and 88.8%	segmentation-based feature extraction
3.	Feature Hashing Technique	82.33%	variable length k-gram
4.	Feature Selection Technique	80.13%	Based on attribute or variable subset
5.	Rough Set Classifier	97.7%	specific feature
6.	KNN and Support Vector Machine (SVM)	For both 91%	physicochemical and structural
7.	N-gram pattern	75.2%	NP{n,m} Pattern based

8.	Extreme Learning Machine (ELM)	88.03% and 87.612%	Feature pattern Based
9.	Discriminative Descriptors Substitution Matrix (DDSM)	82.3% and 85.9%	Motif

4. CHALLENGES IN RESEARCH:

In a Bioinformatics the protein sequence classification is one of the important area were the protein sequence has to be classified into different families, classes or sub classes. There are number of Data mining technique available to do this classification. Even though the techniques are available their also some challenging issues and difficulties faced by the researchers while undergone this task. Some important challenges in research are:

- In protein sequence classification, the dataset that used for research containing large volume of features which may be insignificant or even affect the process of learning. Such noisy data may cause a worse situation such as confusing the mined information or hiding the impact caused by the true value.
- In case self organized map (SOM) network there is no chance of back propagation. But to reach a particular goal and increase the accuracy level of the classification back propagation is most important technique. In back propagation based model, there is a chance to move to the previous steps.
- Computational cost is a major problem which is being faced when input dataset is very large. This has to taken in account by the researchers while finding new techniques.
- One of the possible ways to predict the function of sequence is by predicting the folded 3D structure of a protein from its sequence and then uses features of that structure to infer catalysis (biological reaction), binding partners, or other functional properties. Function prediction based on structure has been one of the "Grand Challenge".
- As data by data analysis are conducted in FUZZY ARTMAP model, storage and time consuming is very high in this model and the computational complexity also high in FUZZY model. This model also failed to process the physical relationships which are most important in this purpose. Hence the performance of this model has to be improved.
- Even though the Rough set classifier used to extract a specific features from the sequence it took extra time and need extra space for computation and also it only gives knowledge based information.
- Decision trees and neural networks, can only take input data as a vector of features and the sequences also do not have explicit features and the dimensionality of potential features may still be very high and the sequential nature of features is difficult to capture and the computation cost is very high for this.
- Besides accuracy, other challenges in sequence classification are to speed up classification in order to handling a large amount of data and to train an interpretable classifier to gain knowledge about characteristics of protein sequences.
- In some cases like protein sequence pattern recognition, some techniques fails to classify patterns with continuous features as the number of attributes is very large.
- Feature selection method required large memory storage to store the vocabularies of k-grams which is very difficult for today's large collection of sequences.
- Drawback in Support and confidence based feature selection measure is, in this method is, it is not easy to select a good value for the minsup and minconf, this may affect the classification quality, it also does not consider the multiple occurrence of sequence feature in an input sequence.
- In case of feature hashing, there is a significant loss of information while hash collision occurred between highly frequent features, with different class distribution.

The above mentioned points are some of the challenging problems that have to be mainly focused by the researchers while finding new techniques for protein sequence classification.

5. CONCLUSION:

In this paper, we provide a brief survey on protein sequence classification. We group sequence classification methods in feature hash based methods, segmentation-based feature extraction method, sequence distance based methods and structural based method. We also review the comparison study of sequence classification methods applied in different application domains. In this paper, we also highlight some of challenging issues faced by the researchers while doing

classification task for protein sequence. From this survey we clearly found that there is a problem in classifying complex sequence data which is still open at large. In future, different analysis has to be done and techniques has to found to increase the accuracy rate, better performance with less computational time and cost present challenges for future studies.

REFERENCES:

[1] Amit Kumar Banerjee & Vadlamani Ravi & U. S. N. Murty & Neelava Sengupta & Batepatti Karuna The Application of Intelligent Techniques for Classification of Bacteria Using Protein Sequence-Derived Features

[2] Padro Gabriel Ferreira, Paulo J.Azevedo, "protein sequence classification through Relevant sequence Mining and Bayes Classifers"

[3] R. Karchin, K. Karplus, and D. Haussler. Classifying G-protein coupled receptors with support vector machines. Bioinformatics, 18(1):147–159, 2002

[4] C. C. Chang and C. J. Lin. LIBSVM: a library for support vector machines. Software available at http://www. csie. ntu. edu. tw/cjlin/libsvm, 80:604–611, 2001.

[5] B.Y.M. Cheng, J.G. Carbonell, and J. Klein-Seetharaman. Protein Classification Based on Text Document Classification Techniques. Proteins: Structure, Function and Bioinformatics, 58:955– 970, 2004.

[6] C. Leslie, E. Eskin, and W.S. Noble. The spectrum kernel: A string kernel for SVM protein classification. Proceedings of the Pacific Symposium on Biocomputing, 7:566–575, 2002.

[7] C. Leslie, E. Eskin, J. Weston, and W.S. Noble. Mismatch string kernels for SVM protein classification.Advances in Neural Information Processing Systems, 15:1441–1448, 2003.

[8] V. John K. Vries,[1*] Xiong Liu,[1,2] and Ivet Bahar[1].The relationship between N-gram patterns and protein secondary structure.(2007).

[9] Zhengzheng Xing, Jian Pei, Eamonn Keogh."A Brief Survey on Sequence Classification"(2006).

[10] Dianhui Wang, Guang-Bin Huang. "Protein Sequence Classification Using Extreme Learnin Machine". Proceedings of International Joint Conference on Neural Networks (IJCNN2005), July 31-August 4, 2005, Montreal, Canada.

[11] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in Proceedings of International Joint Conference on Neural Networks (IJCNN2004), (Budapest, Hungary), 25-29 July, 2004.

[12] G.-B. Huang and C.-K. Siew, "Extreme learning machine: RBF network case," in Proceedings of the Eighth International Conference on Control, Automation, Robotics and Vision (ICARCV 2004),(Kunming, China), 6-9 Dec, 2004.

[13] G.-B. Huang and C.-K. Siew, "Extreme learning machine with randomly assigned RBF kernels," International Journal of Information Technology, vol. 11, no. 1, 2005.

[14] Dr.S.Vijayarani[1] and Ms.S.Deepa[2],"Protein sequence classification in Data mining-A study,"International Journal of Information Technology, modelling and computing (IJITMC) vol.2, no2, May 2014.

[15] Cathy Wu, Michael Berry, Sailaja Sivakumar and Jerry Mclarty Neural Networks for Full-Scale Protein Sequence Classification: Sequence Encoding with Singular Value Decomposition (1995).

[16] Yang Yang[1], Bao-Liang lu1;[2]_ and Wen-Yun Yang1," classification of protein sequences based on word segmentation methods, October 3, 2007 13:34 Proceedings Trim Size: 9.75in x 6.5in.

[17] Pv Nageswara Rao[1](nagesh@gitam.edu), T.Uma Devi1, Dsvgk Kaladhar1,Gr Sridhar[2],Allam Appa Rao[3] "A probabilistic neural network approach for protein superfamily classification "Journal of Theoretical and Applied Information Technology. © 2005 - 2009 JATIT.

[18] [1]Md. Arif Rahman, [2] Md. Alam Hossain, [3] Nazmun Nahar, [4] Must. Reshma Sultana,"An Efficient Technique for Protein Sequence Classification Using Data Mining, International Journal of Information Technology.

[19] ZhihuaWEI [1,2,3,] Duoqian MIAO[1,2],Jean-Hugues CHAUCHAT[3], Rui ZHAO[1], Wen LI [1,2]N-grams based feature selection and text representation for Chinese Text Classification International

[20] Journal of Computational Intelligence Systems, Vol.2, No. 4 (December, 2009), 365-374

[21] Suprativ Saha[1] and Rituparna Chaki[2],APPLICATION OF DATA MINING IN PROTEIN SEQUENCE CLASSIFICATION International Journal of Database Management Systems (IJDMS) Vol.4, No.5, October 2012 DOI: 10.5121/ijdms.2012.4508 103

[22] Yang Yang[1,2] and Bao-Liang Lu[1,2], Extracting Features from Protein Sequences Using Chinese Segmentation Techniques for Subcellular Localization.

[23] John Thomas, Naren Ramakrishnan, and Chris Bailey-Kellogg, Graphical Models of Residue Coupling in Protein Families, IEEE Transactions on Computational Biology and Bioinformatics.

[24] Rabie Saidi[1,2,3,4], Mondher Maddouri[4,5] and Engelbert Mephu Nguifo[*1,2], Protein sequences classification By means of feature extraction with substitution matrices.

[25] *Matthew N. Davies[1],Andrew Secker[2]'Alex A. Freitas[2],Jon Timmis[3],Edward Clark[3],Darren R. Flower[1],Alignment-independent techniques for protein classification.

[26] Murty, U. S. N., Banerjee, A. K., & Arora, N. (2009). Journal of Proteomics & Bioinformatics, 2, 97–107.

[27] Xiao, Y., & Segal, M. R. (2008). Bioinformatics, 24(9), 1198–1205.

[28] Zhang, L., Shao, C., Zheng, D., & Gao, Y. (2006). Molecular & Cellular Proteomics, 5(7), 1224–1232.

[29] Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines. Cambridge: Cambridge University Press.

[30] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proceedings of International Joint Conference on Neural Networks (IJCNN2004)*,(Budapest, Hungary), 25-29 July, 2004.

[31] G.-B. Huang and C.-K. Siew, "Extreme learning machine with randomly assigned RBF kernels," *International Journal of Information Technology*, vol. 11, no. 1, 2005.

[32] K. Blekas, D. I. Fotiadis, and A. Likas. Motif-based protein sequence classification using neural networks. Journal of Computational Biology, 12(1):64–82, 2005.

[33] X. Ji, J. Bailey, and G. Dong. Mining minimal distinguishing subsequence patterns with gap constraints. Knowl. Inf. Syst., 11(3):259–286, 2007.

[34] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden markov support vector machines. In ICML '03: The Twentieth

International Conference on Machine Learning, pages 3–10, 2003.

[35] K. Blekas, D. I. Fotiadis, and A. Likas. Motif-based protein sequence classification using neural networks. Journal of Computational Biology, 12(1):64–82, 2005.

[36] B. Cheng, J. Carbonell, and J.Klein-Seetharaman. Protein classification based on text document classification techniques. Proteins, 1(58):855–970, 2005.

[37] N. A. Chuzhanova, A. J. Jones, and S. Margetts. Feature selection for genetic sequence classification. Bioinformatics, 14(2):139–143, 1998.

[38] M. Deshpande and G. Karypis. Evaluation of techniques for classifying biological sequences. In PAKDD '02: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pages 417–431, 2002.

[39] J. J. R. Diez, C. A. Gonz´alez, and H. Bostr¨om. Boosting interval based literals. Intell. Data Anal., 5(3):245–262, 2001.

[40] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. J. Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. PVLDB, 1(2):1542–1552, 2008.

[41] G. Dong and P. Jian. Sequence Data Mining, pages 47– 65. Springer US, 2007.

[42] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. Chapter 3. Markov Chain and Hidden Markov Model. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, pages 47–65. Cambridge University Press, 1998.

[43] Z. Xing, J. Pei, G. Dong, and P. S. Yu. Mining sequence classifiers for early prediction. In SDM'08: Proceedings of the 2008 SIAM international conference on data mining, pages 644–655, 2008.

[44] Z. Xing, J. Pei, and P. S. Yu. Early classification on time series: A nearest neighbor approach. In IJCAI'09: Proceedings of the 21st International Joint Conference on Artificial Intelligence, pages 1297–1302, 2009.

[45] Milan Randic et. Al., "Novel 2-D graphical representation of proteins", Chemical Physics Letters, Elsevier, doi:10.1016/j.cplett.2005.11.091.

[46] Manoj Kumar Gupta, Rajdeep Niyogi, Manoj Misra, "A 2D Graphical Representation of Protein Sequence and Their Similarity Analysis with Probabilistic Method", MATCH Commun. Math. Comput. Chem. 72 (2014) 519- 532, ISSN 0340 – 6253. (4 groups).

[47] Milan Randic, Jure Zupan,, Alexandru, T. Balban, „ Unique graphical representation of protein sequences based on nucleotide triplet codons, Chemical Physics Letters 397 (2004) 247-252.

[48] Zu-Guo, Vo Anh, Ka-Sing Lau, "Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analysis", Journal of Theoretical Biology, Elsevier, dio: 10.1016/j.jtbi.2003.09.009. (5 groups).

[49] Chun Li, Lili Xing, Xin Wang, "2-D graphical representation of protein sequences and its application to coronavirus phylogeny", BMB reports, July 2007, Page 217-222.

[50] Yusen Zhang, Xiangtian Yu, "Analysis of protein sequence similarity"978-1-4244-6439-5/10/2010 IEEE, pp. 1255-1258.

[51] Yu-hua Yao, Fen Kong, Qi Dai, Ping-an He, " A Sequence segmented method applied to the Simiarity analysis of Long Protein Sequnce", MATCH Commun. Math.Comput. Chem. 70 (2013) 431-450

[52] Soumen Ghosh et. al., "Classification of Amino Acids of a Protein on the basis of Fuzzy set theory", International Journal of Modern Sciences and Engineering Technology, ISSN 2349-3755, Volume 1, Issue 6, 2014, pp.30-35.

[53] Wistrand M, Kall L, Sonnhammer EL. A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein Sci*. **2006**; 15:509-21.

[54] Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **2002**; 18:147-159.

[55] Papasaikas PK, Bagos PG, Litou ZI, Promponas VJ, Hamodrakas SJ. PRED-GPCR: GPCR recognition and family classification server. *Nucleic Acids Res*., **2004**; 32:W380-2.

[56] Guo YZ, Li ML, Wang KL, Wen ZN, Lu MC, Liu LX, Lin J. Fast fourier transform-based support vector machine for prediction of G-protein coupled receptor subfamilies *Acta Biochim Biophys Sin (Shanghai)* **2005**; 37:759-66.

[57] Bhasin M, Raghava GP. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res*. **2004**; 32:W383-9.