International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:4, No:8, 2010

A Novel Approach for Protein Classification Using Fourier Transform

A. F. Ali, and D. M. Shawky

***Abstract*—**Discovering new biological knowledge from the high-throughput biological data is a major challenge to bioinformatics today. To address this challenge, we developed a new approach for protein classification. Proteins that are evolutionarily- and thereby functionally- related are said to belong to the same classification. Identifying protein classification is of fundamental importance to document the diversity of the known protein universe. It also provides a means to determine the functional roles of newly discovered protein sequences. Our goal is to predict the functional classification of novel protein sequences based on a set of features extracted from each protein sequence. The proposed technique used datasets extracted from the Structural Classification of Proteins (SCOP) database. A set of spectral domain features based on Fast Fourier Transform (FFT) is used. The proposed classifier uses multilayer back propagation (MLBP) neural network for protein classification. The maximum classification accuracy is about 91% when applying the classifier to the full four levels of the SCOP database. However, it reaches a maximum of 96% when limiting the classification to the family level. The classification results reveal that spectral domain contains information that can be used for classification with high accuracy. In addition, the results emphasize that sequence similarity measures are of great importance especially at the family level.

***Keywords*—**Bioinformatics, Artificial Neural Networks, Protein Sequence Analysis, Feature Extraction.

I. Introduction

PROTEINS are macromolecules that serve as building blocks and functional components of a cell, and account for the second largest fraction of the cellular weight after water. in addition, proteins are responsible for some of the most important functions in an organism, such as constitution of the organs (structural proteins), the catalysis of biochemical reactions necessary for metabolism (enzymes), and the maintenance of the cellular environment [1]. Identifying protein classification is of fundamental importance to document the diversity of the known protein universe. a number of protein classification databases exist, including the structural classification of proteins (SCOP) [2], class, architecture, topology, and homologous superfamily (cath) [3], and protein family (pfam) [4] databases. Methods used to generate these classifications include sequence-only automated methods such as profile hidden Markov models (profile HMM) and position-specific scoring matrices (PSSM) as well automated structural alignment and hand curation [3], [4]. This paper provides a new set of spectral domain features for the prediction of protein's SCOP classification. For each protein sequence in the SCOP database, we calculate the molecular weight for every amino acid in the sequence. Fast Fourier Transform (FFT) is then calculated for every sequence of molecular weights. Compared to other proteins in different classification levels, the spectrum is different which suggests that FFT coefficients of molecular weights may be used as a good discriminating feature. We then use a neural network classifier to classify protein sequences based on these features. We conducted several experiments in order to determine the classification accuracy as a function of the number of FFT coefficients, and the number of neurons used in the classifier. In the experiments conducted, it was found out that the highest accuracy is achieved when we use only the lowest 40-60 coefficients of FFT of the generated spectrum, and 30-60 neurons in the hidden layer of a 3-layers back propagation neural network. Thus, given a novel protein sequence, we are able to predict its SCOP classification with high accuracy, using the decision made by this classifier.

The remainder of this paper is organized as follows. Section II provides a brief discussion of the related techniques for protein classification. Section III describes the details of our approach. In Section IV, the results of our approach are summarized. Finally, conclusions and directions for future work are provided in Section V.

II. Related Work and Objectives

To understand how proteins function, we need to build a global picture of the protein universe [5]. Newly-discovered protein structures are growing exponentially, hence, the protein universe is constantly changing. In order to understand the functions of proteins and their relationships to each other, classifications of proteins should be updated frequently [6].

Considerable research has addressed the problem of protein classification. Traditionally, protein classification has relied on sequence alignment methods such as BLAST [7], where a protein's function is inferred from proteins of similar sequence whose function is known. However, this approach is only reliable for high sequence similarity values, and even then, the transfer of function is not complete between proteins [8]. Furthermore, the inference of function from proteins whose function was already inferred can lead to the propagation of errors [9]. Comparing protein structure in addition to sequence has been suggested as a way to predict protein function [10]-[12], since a protein's structure is directly related to its function and is conserved at low levels of sequence similarity [13]. In addition, some authors advocate the use of automated methods to predict structural features and classes from protein sequences as a step to predict function [13], [14]. As an

A. F. Ali is with the Biomedical Engineering Department, Faculty of Engineering, Helwan University, Cairo, Egypt, e-mail: ahmed.farag@mcit.gov.eg.

D. M. Shawky is with the Department of Engineering Mathematics, Faculty of Engineering, Cairo University, e-mail: doaashawky@yahoo.com.

alternative to alignment methods, several machine learning approaches have been applied to protein function prediction/classification from sequence and structure data [15]. Support vector machines (SVMs) have been the most popular (e.g. [14]), although other methods such as decision trees [16], [17], Markov chains [18], and neural networks [19], [20] have also been used with some success. SVMs in particular, have proven more reliable than sequence alignment methods in situations of low sequence similarity, such as remote homology detection [21], [22]. However, much of the focus of SVM-based studies has been on finding useful representations of the sequence data. Some authors used explicit feature vectors using global sequence properties such as amino acid composition [15], [23], and [24], in which the order and pattern of the amino acids in the protein sequence are ignored. Others used more complex Kernel functions, involving sequence similarity or pattern matching [25]-[29], most of which are computationally intensive, and do not differ much from alignment methods. Theoretically, the sequence of amino acids of a protein contains all the necessary information to predict its function. Thus an approach that can deal directly with sequences could be advantageous [25]. Only the techniques based on sequence alignments considered the sequence of amino acids, however, they measure sequence similarity locally and suffer from high complexity especially for long protein sequences.

Despite the power of the mathematical tools of signal processing, their application to protein sequences has been minimal. One of the very first applications was in the computation of the hydrophobic moment of protein domains [30] and in detecting periodicities in secondary structure (α-helix, β-sheet and 310-helix) [31]. Wavelet analysis of the hydrophobicity signal has been used to locate the secondary structure content relating the periodicity observed in the signal to the known values of secondary structure period [32]. The Fourier spectrum is computed from the hydrophobicity and secondary structure signal of a protein, and the power spectrum served as the feature input into a neural network [33]. To the best of our knowledge, none of the existing techniques apply Fourier transform for detecting the sequence similarity and classifying the unknown protein sequence according to some features detected from the spectral domain.

III. Proposed Technique

Fig. 1 depicts the analysis steps of our approach. For each protein in SCOP database, a set of features is calculated. The most discriminating features are selected, and passed to the classifier. Finally, the classification decision is taken. The following sections will discuss each process separately.

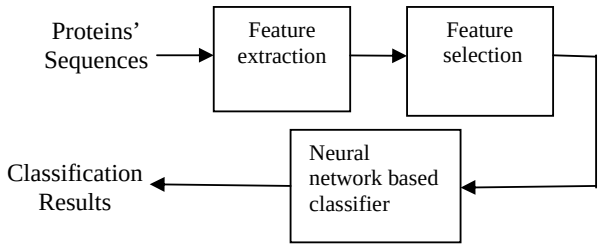

Fig. 1. Block diagram of the proposed approach

A. Used Database

Protein classification schemes employ different heuristics, similarity metrics, and different degrees of automation. SCOP is one of the classification schemes which is created mainly by manual inspection [5]. This is perhaps the reason that it is accepted by many researchers as the most accurate classification scheme (or the ground truth) [34]. SCOP is a database of known structural and evolutionary relationships amongst proteins of known structures [2]. It has been created as a hierarchy of several obligatory levels. The fundamental unit of classification is a domain in the experimentally determined protein structure. Protein domains are grouped at different levels according to their sequence, structural and functional relationships [2]. Proceeding from bottom to top, the SCOP hierarchy comprises the following levels: protein Species, representing a distinct protein sequence; Protein, grouping together similar sequences of essentially the same functions ; Family containing proteins with related sequences but typically distinct functions; and Superfamily bridging together protein families with common functional and structural features inferred to be from a common evolutionary ancestor. Near the root, the basis of classification is purely structural: structurally similar superfamilies with different characteristic features are grouped into Folds, which are further arranged into Classes based mainly on their secondary structure content and organization. The seven main classes in the latest release contain 92927 domains organized into 3464 families, 1777 superfamilies and 1086 folds. The SCOP domains correspond to 34 495 entries in the Protein Data Bank (PDB) [35]. Statistics of the current and previous releases, summaries and full histories of changes and other information are available from the SCOP website (http://scop.mrc-lmb.cam.ac.uk/scop/) together with parseable files encoding all SCOP data [36]. The sequences and structures of SCOP domains are available from the ASTRAL compendium [37], and hidden Markov models of SCOP domains are available from the SUPERFAMILY database [38]. Since the creation of SCOP in 1994, the number of known protein structures has grown more than 20-fold, whereas the numbers of SCOP folds, superfamilies and families have increased 4-fold, 5-fold and 7-fold, respectively [36].

SCOP is updated manually every six months [36]. However, automated classification schemes have the advantage that the view of the protein universe can be updated frequently to include newly-discovered protein structures in a timely manner.

B. Feature Extraction

In this paper, spectral domain features based on FFT of molecular weight of each protein sequence are used. The total number of protein sequences used is 14762. This number was obtained after applying a data cleaning and preprocessing stage over the 15273 available sequences in the used version of SCOP database.

For each protein sequence in the database, a tuple of four attributes is produced to represent its SCOP classification. The tuple consists of class.fold.superfamily.family (e.g. d.136.1.1) according to which all sequences are sorted in a 2-dimensional array. The first column represents a sequence's SCOP classification, and the second column includes the sequence

itself. After sorting protein sequences, they are divided into 3399 categories based on their unique full four digit classification.

After the preprocessing step, we created fixed length sequences for each protein by padding each sequence with zeroes until it reaches 752 characters which is the length of the longest protein sequence in the database used. Molecular weight of each amino acid in each sequence is calculated, and then it is transformed to spectral domain using Fast Fourier Transform. The spectrum is averaged for all sequences belonging to the same SCOP classification. Two levels for classification were used. Firstly, all four levels up to the class level were used. Secondly, only the family level was used. In both experiments, the averaged spectra of the different classes are compared. Figures 2-4 show the averaged spectrum of a.1.1.1, b.34.11.4, and c.72.34.11 proteins respectively. The x-axis represents the frequency components in Hz, while the y-axis represents the average magnitude of Fourier coefficients of molecular weights.

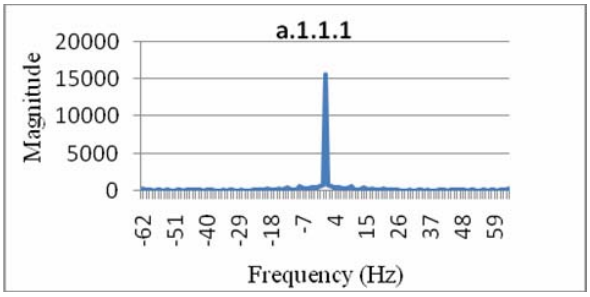

Fig. 2. The average spectrum of molecular weight of each amino acid in a.1.1.1 proteins

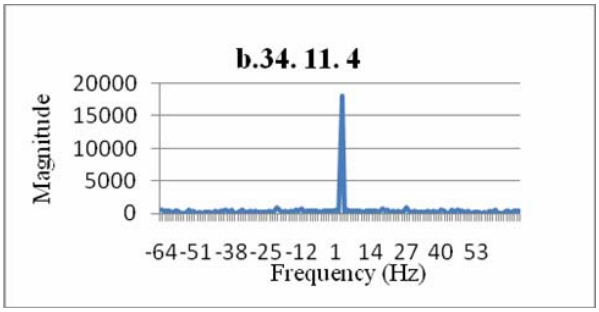

Fig. 3. The average spectrum of molecular weight of each amino acid in b.34.11.4 proteins

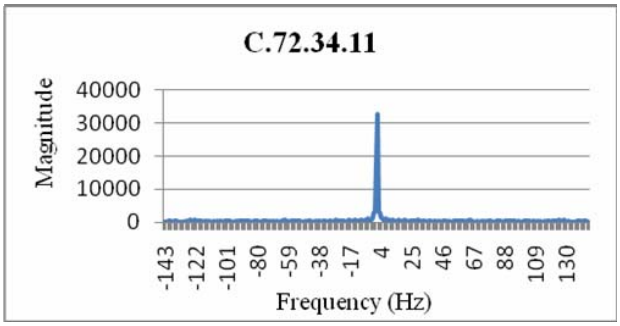

Fig. 4. The average spectrum of molecular weight of each amino acid in c.72.34.11 proteins

From the example spectra shown in the above figures, we notice that the spectrum of different protein families is different, which suggests that there is some information in the spectral domain that can be used for classification. The differences are obvious regarding the peak of each spectrum which occupies a frequency range around 1, 4, and 7 Hz for a.1.1.1, b.34.11.4, and c.72.34.11 proteins respectively. In addition, the maximum magnitude of Fourier coefficients is more than 15000 for a.1.1.1 and b.34.11.4 proteins. Meanwhile it exceeds 30000 for c.72.34.11 proteins.

C. Best Features Selection

The 14762 spectra generated are divided into 60% for training, and 40% for validation. The training data is used as an input to a neural network classifier who is learned to classify the inputs according to their spectra. The network consists of 3 layers. The number of neurons in the input layer is set equal to the number of FFT coefficients used (features). The number of neurons in the hidden layer is varied in every experiment we conducted. In addition, the number of neurons used in the output layer is set equal to log 2m, where m is the number of different classes to be identified.

The performance of the proposed classifier depends on two factors, the number of neurons in the hidden layer, in addition to the number of features to be used. Thus, to find the optimum operating point, we increased the used number of features gradually from 30 to 60. For each number of features belonging to this range, the number of neurons in the hidden layer is increased gradually. The reason for choosing this range of values is that we noticed from the experiments conducted that the accuracy reached an acceptable value (above 70%) when the number of features used reached 30. Above 60 features, the accuracy was saturated. Similarly, the performance reached an acceptable value when the number of neurons varied from 30 to 40. Figure 5 shows the accuracy calculated at 38 neurons in the hidden layer with variable FFT coefficients ranging from the lowest 30 to 60 coefficients. Moreover, Fig. 6 shows the effect of varying both the number of neurons and the number of FFT coefficients on the accuracy. Experimental results indicate, as shown in Fig. 6, that the maximum obtained accuracy in these settings is about 91% which occurs when we use 38 neurons and 43 features. These empirical values indicate that the frequency domain

contains information that can be used for proteins classification with an acceptable performance.

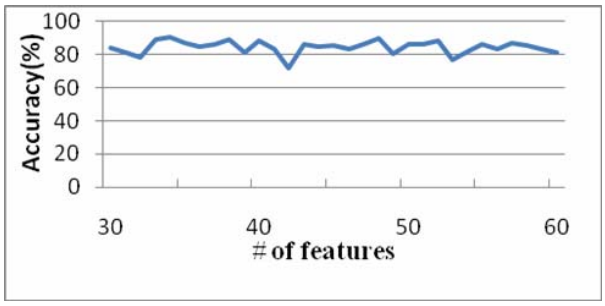

Fig. 5. The accuracy vs. the number of FFT coefficients

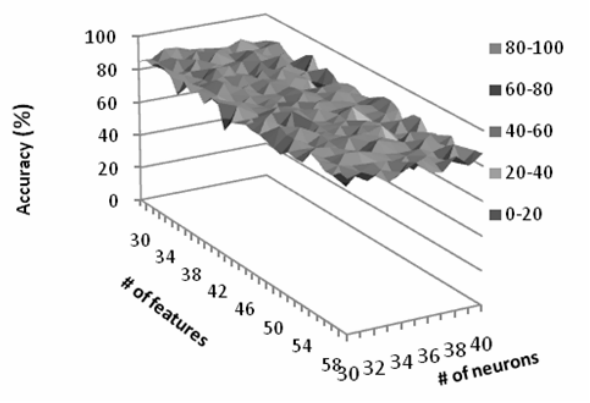

Fig. 6. The accuracy vs. the number of FFT coefficients and the number of neurons

IV. Results

To the best of our knowledge, no study was found in the literature for the classification of proteins into full four-digit classes. Considering the difficulty of the task, and after conducting several experiments, it was found out that the best accuracy with the minimum number of features and neurons is about 98% in the training step and 91% in the validation step when only the lowest 43 FFT coefficients are used as features, with 38 neurons in the hidden layer. Table 1 shows classification accuracy for some Enzyme Commission (EC) families. The lowest obtained accuracy is 77% for 3.4.22.45 proteins. When we limited the classification level to the family level, the classification accuracy was raised to a maximum of 100% in the training step. In addition, it reaches a maximum of 96% in the validation step. Table 2 gives the classification accuracies for the same proteins listed in Table 1, when the classification level is limited to the family level only. The improvements show the relevance of molecular weights as a discriminating feature for all proteins at the family level. These improvements were expected since sequence similarities are the main similarity measures at the family level. Both experiments show that spectral domain contains information about sequences that can be used for classification which is the main contribution of the conducted work.

TABLE 1
CLASSIFICATION ACCURACIES FOR TRAINING AND VALIDATION FOR SOME EC PROTEINS UP TO THE CLASS LEVEL

EC family	Accuracy		EC class	Accuracy	
	Training	Validation		Training	Validation
1.1.1.1	0.98	0.91	2.1.1.33	0.97	0.86
1.1.1.23	0.96	0.88	2.1.1.45	0.97	0.87
1.1.1.27	0.98	0.85	2.1.1.56	0.89	0.86
1.1.1.37	0.97	0.87	3.4.22.28	0.98	0.88
1.2.1.12	0.93	0.87	3.4.22.29	0.99	0.90
1.2.1.38	0.96	0.81	3.4.22.44	0.98	0.84
1.2.1.41	0.89	0.90	3.4.22.45	0.94	0.77
1.2.1.70	0.88	0.90	6.3.4.2	0.9	0.88
2.1.1.14	0.89	0.91	6.3.4.4	0.95	0.90
2.1.1.31	0.89	0.90	6.3.4.5	0.99	0.88

TABLE II
CLASSIFICATION ACCURACIES FOR TRAINING AND VALIDATION FOR SOME EC PROTEINS LIMITED TO THE FAMILY LEVEL ONLY

EC super-family	Accuracy	
	Training	Validation
1.1.1	1	0.96
1.2.1	1	0.94
2.1.1	0.97	0.96
3.4.22	0.98	0.95
6.3.4	0.95	0.92

In Fig. 7, the Receiver Operating Characteristic curve (ROC) of the used features is shown for the four level classification experiments. It is to be noticed that the area under the curve is near one which emphasize that the suggested features are good representatives for the classes to be identified.

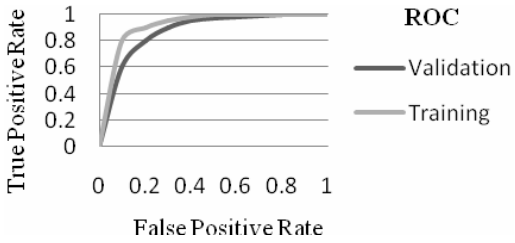

Fig. 7. The ROC for the used features

V. Conclusions And Future Work

In this paper, a new set of features for protein's SCOP classification is proposed. The features used are FFT coefficients of the spectrum of the molecular weight of each SCOP protein. Experimental results show that the best

accuracy occurs when using only the lowest 43 FFT coefficients as features.

Fourier analysis provides a useful DSP means for the description of protein sequences where it is used to extract characteristic bands from these sequences. In our approach, the sequence-scale analysis with Fourier analysis gave a multi-resolution similarity comparison between protein sequences. Using Fourier transform, we took into account not only the local pair-wise amino acid but also the information contained in coarser spatial resolution. Also, this Fourier based method did not require the complex sequence alignment processing for sequences. Therefore, proteins with different sequence lengths could be compared easily.

Currently, we are applying more intelligent feature selection techniques. Moreover, we intend to use more physical features for the sequences, in addition to other structural measures to further improve the classification accuracy up to the class level.

REFERENCES

[1] J. Zhao, "Multivariate Statistical Analysis of Protein Variation", A Ph. D. dissertation, available at http://www.lib.ncsu.edu/theses/available/etd-12092005-003538/unrestricted/etd.pdf

[2] A. Murzin, S. Brenner, T. Hubbard, and C. Chothia, "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536-540, 1995.

[3] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton, "CATH- A Hierarchic Classification of Protein Domain Structures," *Structure*, vol. 5, no. 4, pp. 1093-1108, 1997.

[4] A. Bateman, L. Coin, R. Durbin, R. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer, D. Holme, C. Yeats, and S. Eddy, "The Pfam protein Families Database," *Nucleic Acids Res*., vol. 32, no. 36, pp. D138-D141, 2004.

[5] O. Camoglu, T. Can, A. Singh, and Y. Wang, "Decision Tree Based Information Integration for Automated Protein Classification," Journal of Bioinformatics and Computational Biology (JBCB), Vol. 3, No. 3, pp. 717-742, 2005.

[6] O. André, F. Daniel, F. António, "Peptide programs: applying fragment programs to protein classification", *Proceeding of the 2nd International Workshop on Data and Text Mining in Bioinformatics*, pp. 37-44, 2008.

[7] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res*., vol. 25, no. 17, pp. 3389-3402, 1997.

[8] W. Tian, and J. Skolnick, "How well is enzyme function conserved as a function of pairwise sequence identity?", *Molecular Biological*, vol. 3, no.4, pp. 863-882, 2003.

[9] D. Devos, and A. Valencia, "Intrinsic errors in genome annotation", *Trends Genetics*, vol. 17, no.8, pp. 429-431, 2001.

[10] E. N. Baker, V. L. Arcus, and J. S. Lott, "Protein structure prediction and analysis as a tool for functional genomics", *Appl. Bioinformatics*, vol. 2, no. 3, pp. 3-10, 2003.

[11] M. Grotthuss, D. Plewczynski, K. Ginalski, L. Rychlewski, and E. I. Shakhnovich, "PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics", *BMC Bioinform*atics, vol. 7, no. 1, pp. 53-56, 2006.

[12] J. C. Whisstock, and A. M. Lesk, "Prediction of protein function from protein sequence and structure", *Q Rev Biophys*., vol. 36, no. 3, pp. 307-340, 2003.

[13] I. Friedberg, "Automated protein function prediction the genomic challenge", *Brief Bioinfo*rmatics, vol. 7, no. 3, pp. 225-242, 2006.

[14] I., Melvin, E. Ie, J. Wetson, W. S. Noble, and C. Leslie, "Multi-class protein classification using adaptive codes", *J Mach. Learn. Res*., vol. 8, pp. 1557-1581, 2007.

[15] L. Y. Han , C. Z. Cai, Z. L. Ji, Z. W Cao., J. Cui, and Y. Z. Chen, " Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach", *Nucleic Acids Res*., vol. 32, no. 21, pp. 6437-6444, 2004.

[16] R. E. Langlois, M. B. Carson, N. Bhardwaj, and H. Lu "Learning to translate sequence and structure to function: Identifying DNA binding and membrane binding proteins" , *Annals of Biomedical Engineering*, vol. 35, no. 6, pp. 1043-1052, 2007.

[17] Z. R. Yang, and R. Hamer, "Bio-basis function neural networks in protein data mining", *Current Pharmaceutical Design*, vol. 13, no. 14, pp. 1403-1413, 2007.

[18] J. Busch, P. Ferrari, A. Flesia, S. P. Grynberg, and F. Leonardi," Testing statistical hypothesis on random trees and applications to the protein classification problem", Annals of Applied Statistics, Vol.3, No.2, pp.542-563, 2009.

[19] M. Q. Yang, J. Y. Yang, and O. K. Ersoy, "Classification of proteins multiple-labelled and single-labelled with protein functional classes", *Int. J Gen. Syst*., vol. 36, no.1, pp. 91-109, 2007.

[20] C. Pasquier, V. Promponas, and S. J. Hamodrakas, "PRED-CLASS: Cascading Neural networks for generalized protein classification and genome wide applications", *Proteins, PROTEINS: Structure, Function, and Genetics*, vol. 44, no.1, pp. 361-369, 2001.

[21] B. J. Webb-Robertson, C. Oehmen, and M. Matzke, "SVM-BALSA: Remote homology detection based on Bayesian sequence alignment", *Computational Biological Chemistry*, vol. 29, no. 6, pp. 440-443, 2005.

[22] Z. D. Zhang, S. Kochhar, and M. G. Grigorov, " Descriptor-based protein remote homology identification", *Protein Science*, vol. 14, no.2, pp. 431-444, 2005.

[23] N. Bhardwaj, R. E. Langlois, G. J Zhao, and H. Lu " Kernel-based machine learning protocol for predicting DNA binding proteins", *Nucleic Acids Res*, vol. 33, no. 20, pp. 6486-6493, 2005.

[24] P. D. Dobson, and A. J. Doig, "Predicting enzyme class from protein structure without alignments", *Journal of Molecular Biology*, vol. 345, no. 1, pp. 187-199, 2005.

[25] Y. D. Cai, and A. J. Doig, "Prediction of Saccharomyces cerevisiae protein functional class from functional domain composition", *Bioinformatics*, vol. 20, no.8, pp. 1292-1300, 2004.

[26] Q. W. Dong, X. L. Wang, and L. Lin, "Application of latent semantic analysis to protein remote homology detection", *Bioinformatics*, vol. 22, no. 3, pp. 285-290, 2005.

[27] R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie, "Profile-based string kernels for remote homology detection and motif extraction", *Journal of Bioinformatics and Computational Biology*, vol. 3, no.3, pp. 527-550, 2005.

[28] H. Rangwala, and G. Karypis, "Profile-based direct kernels for remote homology detection and fold recognition", *Bioinformatics*, vol. 2, no.23, pp. 4239-4247, 2005.

[29] L. Nanni, S. Mazzara, L. Pattini, and A. Lumini, "Protein classification combining surface analysis and primary structure", *Protein Engineering: Design and Selection*, vol. 22, no. 4, pp. 267-272, 2009.

[30] D. Eisenberg, R. Weiss, and T. Terwilliger, "The Helical Hydrophobic Moment: A Measure of the Amphiphilicity of a Helix", *Nature*, vol.4, pp. 299-371, 1982.

[31] D. Eisenberg, E. Schwarz, M., Komaromy and R. Wall, "Analysis of Membrane and Surface Protein Sequences with the Hydrophobic Moment Plot", *Journal of Molecular Biology*, vol.42, no.1, pp. 125-179, 1984.

[32] L. Pattini, L. Riva, and S. Cerutti, "A wavelet based method to predict the alpha helix content in the secondary structure of globular proteins", *Proceedings of the IEEE-EMBS*, pp.132-133 , 2002.

[33] A. Shepherd, G. Gorse, and J. Thornton, "A novel approach to the recognition of protein architecture from sequence using Fourier analysis and neural networks", *Proteins*, vol. 50, no.2, pp. 290-302, 2003.

[34] A. Antonina, H. Dave, C. John-Marc, and E. Steven, "Data growth and its impact on the SCOP database: new developments", *Nucleic Acids Res*., vol. 36, no. 1, pp. 1-7, 2008.

[35] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, "The Protein Data Bank", *Nucleic Acids Res*., vol. 28, no. 1, pp.235–242, 2000.

[36] L. Lo Conte, S.E. Brenner, T.J.P. Hubbard, C. Chothia, and A.G. Murzin, "SCOP database in 2002: refinements accommodate structural genomics", *Nucleic Acids Res*., vol. 30, no.1, pp. 264–267, 2002.

[37] J. M. Chandonia, G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S.E. Brenner, "The ASTRAL compendium in 2004", *Nucleic Acids Res*., vol. 32, no.1, pp. 189–192, 2004.

[38] D. Wilson, M. Madera, C. Vogel, C. Chothia, and J. Gough, "The SUPERFAMILY database in 2007: families and functions", *Nucleic Acids Res*., vol. 35, Database Issue, pp. 308–313, 2007.