ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Prediction of Hubness of Human Proteins Using Wavelets and Tumor Proteins in P53 Pathway as a Step towards Cancer Research

Sajeev J[1], Dr. T. Mahalekshmi[2]

Associate Professor, Department of Computer Applications, Sree Narayana Institute of Technology, Kollam, Kerala, India

Professor and Principal, Department of Computer Applications, Sree Narayana Institute of Technology, Kollam, Kerala, India

ABSTRACT: With the rapid development of advanced computing facilities, the amount of biological sequence data has increased exponentially. This has given rise to the need of advanced algorithms to process these data to fish out patterns hidden in these sequences. From the point of view of signals, these biological sequences can be seen as one-dimensional signals. So researchers have been applying signal processing techniques for mining useful information from these sequences. For the last few years wavelet transforms have been used to extract hidden features from the biological sequences. In this paper wavelets are used in the prediction of hubness in human proteins with the help of tumor proteins in P53 pathway. Specifically, we discuss an approach for representing the biological sequence numerically and methods of using wavelet analysis on this numerical sequence for the prediction of hubness of human proteins.

KEYWORDS: Hub proteins, PIN, tumor proteins,P53, wavelet analysis, SVM, NCBI, AAIndex.

I.INTRODUCTION

Bioinformatics is an interdisciplinary field which is used to develop methods and for analyzing different types of biological data[1]. Biological data can be generally classified as Genomics and Proteomics Data.

Hub proteins are highly connected and active as the name suggested[2], [3]. These inevitable proteins are vital for the proper biological functioning of humans [4], [5], [6], [7]. The present work is an attempt to extract hub characteristics of Human Proteins using tumor proteins as the bench mark. The study checks whether the features available in the Tumor Proteins are responsible for the hubness of other human proteins.

The work propose a method which explores the key characteristics of tumor proteins for the purpose of hub feature extraction. It is not mandatory that the characteristics responsible for the hubness in one protein should remain same across other proteins, but the presence or absence of some key characteristics may play the big role for that.

Protein analysis is crucial in drug discovery [8] and cancer research and is an important application area of Bioinformatics. This is the reason why ever going research is taking place in Bioinformatics domain. One such sub area is the studies relating to cellular activities and disease states [1]. Here identification of DNA binding domains, protein sequence features, protein domains and protein structures are very important. So Protein sequences are vital in conducting such analysis and studies.

The sections below discuss background, existing system, data set, proposed method, results and discussion followed by conclusion and future work.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

II.BACKGROUND

In this section important background information are discussed that are pertinent to the proposed method such as Hub proteins, Tumor proteins, Numerical representation of biological sequences, wavelet analysis and application of wavelets in biological sequence analysis.

2.1 Hub Proteins

The highly connected Hub proteins are able to form a network of proteins [9]called PIN (Protein Interaction Network) due to their increased binding characteristics available in the surface. Figure 1 depicts a part of PIN derived from HPRD(Human Protein Reference Database) which gives clear perception that certain proteins such as BTRC, PAEP, UBE2EI carry more connections than other proteins called degree of connectivity. Degree of connectivity of each proteins in figure 1 is given in Table 1.

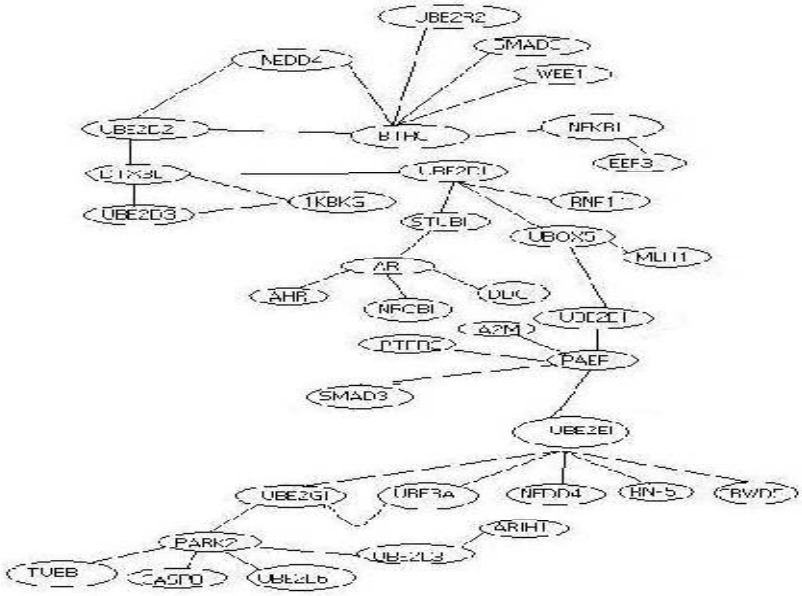

Fig 1. A sub-network of PIN using HPRD Database

TABLE I. Degree of Connectivity of Proteins in the subnetwork given in figure 1.

ID	frequency	ID	frequency	ID	frequency
BTRC	18	RNF5	10	AHR	27
UBE2R2	3	CBWD5	1	NR0B1	11
PAEP	9	AR1H1	1	DDC	1
UBE2L6	10	TUBB	42	A2M	27
UBE2E1	12	UBOX5	7	PTPRC	44
UBE2D1	9	UBE2E1	12	VBE2G1	6
SMAD3	184	IMMT	37	UBE3A	24
WEF1	20	MCH1	6	NEDD4	36
NFKB1	72	AR	138	UBE2D2	6
DTX3L	7	CASP8	111	1KBKG	57
UBE2D3	12	STUB1	29	RNF11	61

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

PIN's are counted as either Random Networks or as Scale Free Networks where Scale Free Networks closely model most of the real world networks [10]. But still a lot of havoc exist about the scale free nature of human biological networks. The large and complex protein interactions direct most biological pathways and processes [3] and surprisingly most of these interactions are directed by hub proteins. This is the reason why they are considered as lethal proteins which are strategically located and if disturbed can lead to biological lethality [10]. Hence study of hub proteins is relevant to understanding the causes of diseases and provides efficient and cost effective solutions.

Hub Proteins participate in significant number of protein interactions and play a vital role in the organization of cellular protein interaction networks [11, 12].This is the reason why Hub proteins are more essential than the non-hub proteins and could be of particular interest as drug targets [13] related to drug discovery.

2.2 Tumor Proteins

Tumor proteins are those proteins which are responsible for cancerous tumors in case of a mutation in human cells.

P53 (also known as tumor protein 53) is a protein in humans encoded by the TP53 gene which is a tumor suppressor gene[13], i.e., its activity stops the formation of tumors. Over the years P53 has been shown to interact with more than hundred proteins, which is evident from the pathway information [14]. This shows the importance of P53 as a Hub protein [15].

2.3 Numerical Representation of Biological Sequences

For further analysis of biological sequences they need to be encoded in a suitable format. Then the input biological sequences can be processed as signals and signal processing techniques such as wavelets can be utilized to extract hidden features out of these sequences. The encoding process is a kind of numerical substitution for each character symbols that forms the biological sequence.

Using these encoding process two types of biological sequences such as DNA nucleotide sequences and protein amino acid sequences can be successfully mapped to required formats for processing. DNA sequence is easier compared with protein sequences for encoding since only 4 character symbols are there with DNA sequences where proteins are represented with 20 amino acids.

In the TABLE 2 given below each amino acid is represented as a complex number which is used for representing the entire amino acid sequence as numeric sequence [16].

The same has been successfully used in cancer research to classify driver genes and passenger genes [16].

TABLE II. The Complex Representation of 20 Amino Acids

Amino Acid Name	Symbol	Complex Number Repr.
Alanine	A	0.61 + 88.3i
Arginine	R	0.60 + 181.2i
Asparagine	N	0.06 + 125.1i
Aspartic	D	0.46 + 110.8i
Cysteine	C	1.07 + 112.4i
Glutamic	E	0.47 + 140.5i
Glutamine	Q	148.7i
Glycine	G	0.07 + 60.0i
Histidine	H	0.61 + 152.6i
Isoleucine	I	2.22 + 168.5i
Leucine	L	1.53 + 168.5i
Lysine	K	1.15 + 175.6i
Methionine	M	1.18 + 162.2i
Phenylalanine	F	2.02 + 189.0i
Proline	P	1.95 + 122.2i
Serine	S	0.05 + 88.7i
Theronine	T	0.05 + 118.2i
Triptophan	W	2.65 + 227.0i
Tyrosine	Y	1.88 + 193.0i
Valine	V	1.32 + 141.4i

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

2.4 Wavelet analysis

Jean Batiste Joseph Fourier, a French mathematician developed the concept of Fourier Trignometric series. Through this concept he represented a periodic function in terms of a weighted sum of cosine and sine functions. This was considered as the origin of wavelets theory. In 1909 Alfred Haar developed Haar Wavelets family which is considered the simplest wavelets. Compared with harmonic functions used in Fourier analysis, wavelets can be used to analyze a given signal in terms of functions that are more finite in time. One of the important property of the Haar wavelets which gave wide acceptance across the globe was the scaling property which give more accurate results in modeling functions. The idea of multiresolution, which is the base theory of versatile wavelets families, was proposed[17]. Using this multiresolution concept, Daubechies [18] created the most frequently used Daubechies wavelets family. This is evident from the above statements that the wavelet theory originated from Fourier Transform.

The characteristics of most of the real-world signals vary in both time and frequency domains. They are also called nonstationary signals.

Fourier Transform is one way to find frequency content and measure the signal composition in frequency. Fourier Transform can be calculated using (1). Here F is the frequency in Hertz and Ωt is the phase in radians:

$$FT\{x(t)\} = X(\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt,\ \Omega = 2\pi F. \quad (1)$$

The FT defines the global representation of the frequency content of a signal over a total period of time. However, it does not give access to the signal's spectral variations during this interval of time. In other words, the time and frequency information cannot be seen at the same time, and thus, a time-frequency representation of the signal is needed. To circumvent this localization problem, Gabor [19] proposed the STFT to analyze only a small section of the signal at a time by using a technique called windowing the signal. This obtains the specific contents of each of the analyzed sections separately. The segment of signals in each section is assumed stationary. Let *g(t)* be the sliding window of a fixed size. STFT is defined in (2), where $g(t-b)e^{-j\Omega t} = \psi^*_{\Omega,b}(t)$ is the complex conjugate of $\psi_{\Omega,b}(t)$.

$$\begin{aligned} STFT_{g(\Omega,b)}\{x(t)\} &= \int_{-\infty}^{\infty} x(t)g(t-b)e^{-j\Omega t}dt \\ &= X_g(\Omega, b). \end{aligned} \quad (2)$$

One of the limitations of STFT is due to the fixed size window used. A narrow window and wide window results in poor frequency resolution and poor time resolution respectively. Also it is really difficult to determine the time intervals where a particular frequency exists. Thus wavelet transform was proposed to get rid of these problems as an alternative to STFT. The definition of continuous wavelet transform is given below [16]:

$$\begin{aligned} CWT_x^{\psi}(a,b) &= X_{\psi}(a,b) \\ &= \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(t)\left[\psi^*\left(\frac{t-b}{a}\right)\right]dt \\ &= \langle x(t), \psi^*_{a,b}(t)\rangle, \end{aligned} \quad (3)$$

Here a and b are the scaling and translation parameters, respectively, and $\psi^*_{a,b}(t) = \frac{1}{\sqrt{a}}\psi^*\left(\frac{t-b}{a}\right)$ is the mother wavelet (base function), a prototype for generating the other window functions. All the windows are the dilated, compressed, and shifted versions of the said mother wavelet. There are different types of wavelet basis functions. In summary, wavelet analysis techniques outrun the traditional FT in the following perspectives [20]:

1. wavelets are suitable for analysis on both stationary and nonstationary signals where FT is less useful in analyzing nonstationary signals;
2. wavelets are well localized in both time and frequency domains, where the standard FT is localized in frequency domain only;
3. the base functions of wavelets can both be scaled and shifted, while the FT can only be scaled; and

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

4. wavelets have solid mathematics foundation and a wider range of applications than FT such as nonlinear regression and compression.

2.4.1 Wavelet families for Biological Sequences

This section illustrates some of the wavelet families commonly used in biological sequence analysis. Wavelet families generally belong to one of the following types. [16].

- Orthogonal wavelets with scaling finite impulse responses (FIR) filters. These wavelets are defined through a lowpass scaling filter. Predefined families of such wavelets include: Haar, Daubechies, Coiflets, and Symlets. .
- Biorthogonal wavelets with scaling finite impulse responses filters. These wavelets are defined through two scaling filters, for reconstruction and decomposition, respectively.
  The BiorSplines wavelet family is an example of a predefined family of this type.
- Wavelets with scaling function. These wavelets are defined using a wavelet function, the mother wavelet, and a scaling function, the father wavelet, in the time domain. The Meyer wavelet family is a predefined family of this type.
- Wavelets without scaling filters and without scaling function. These wavelets are defined through the definition of the wavelet function. The wavelet has a time-domain representation only. Predefined families of such wavelets include Morlet and Mexican_hat.

2.5 Application of Wavelets in biological sequence analysis

The below given figure 2 shows basic operations done as part sequence analysis using wavelet mechanism. Here the biological sequence such as proteins sequence after representing in numerical form undergoes wavelet transform for further protein classification. Here the classification is done for identifying whether the proteins belong to classes hub or non hub.

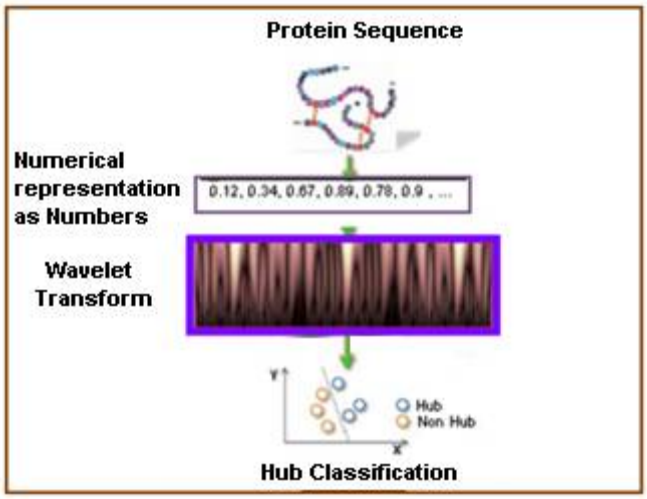

Fig 2. A sample procedure of applying wavelet analysis in Hub protein classification

As an example of the effect of wavelet transforms on biological sequences, Daubechies wavelets function is applied to visualize mutations of a segment of p53 (TP53) gene. Figs. 3a and 3b shows the wavelets coefficients before and after the mutations of a part of p53 gene. The following steps are carried out for the same [16]:

1. Last 1,000 nucleotides from the original DNA sequence (TP53) is extracted.
2. Mapped the sequence segment to complex numbers.
3. Daubechies transform is applied to the complex numbers in scales 2 to 400 at a step length of 2 and obtained a coefficients matrix (visualized in Fig. 3a).
4. Manually mutated 50 base pairs from the extracted 1000 nucleotide sized sequence with nucleotides indices from 401 to 450.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

5. Again Perform Daubechies transform on the artificially mutated sequence segment to obtain a new coefficients matrix (visualized in Fig. 3b).

From the figure 3 the ability of wavelet transform to capture the variations in a DNA sequence due to mutation at different scales is visually noticeable that two rectangular regions in Fig 3a and 3b have a difference in colors. This analysis sheds lights on the possibility of using wavelet techniques for hub protein sequence analysis.

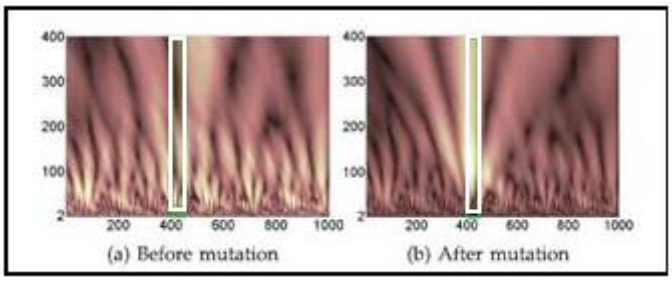

Fig.3 Wavelet transform coefficients visualization on a segment of p53 gene (TP53). The x-axis represent nucleotides indices and y-axis represent the scale numbers [16]. Mutation spots are represented as rectangles. Based on a mutation in the DNA sequence the bar shows a visualized regional difference in wavelet coefficients.

III.EXISTING SYSTEM

The existing methods in the area of PIN maybe classified as experimental and computational. The experimental (large-scale proteomic experiment) techniques though they have vast coverage and sensitivity, do not give much information about the interacting residues [21]. Computational analysis of PIN is based on various attributes like gene proximity, gene fusion events, phylogenic profiling, identification of interacting protein domains and text mining techniques [21]. Each of these approaches has its own strengths and weaknesses especially with regard to sensitivity and specificity. Given below are some of the existing methods in this area.

3.2 Method 1

A method for the prediction of Hubs in Scale-free networks is seen in [11] which is based on List Dominating Set problem (LDS) and is used to model Hubs. This method gives due importance to identifying communities as 'Quasi Cliques' which are sub networks of a given PIN with 'small' number of edges missing in contrast to cliques that are completely connected. If the network is very dense then the system will typically not give a solution. The characterization of community is identified through this method.

3.3 Method 2

Based on one of the findings in [22] that Hub proteins with common interaction partners tend to interact with them through a common interacting motifs, a method has been developed in [23] to predict Hub protein using similar interacting motifs. The input of this method is the binary protein interactions; neither sequence nor structure information is required. By building an interaction network and applying clustering technique this method identifies interacting motifs. These interacting motifs are assigned to Hub proteins and then analyzed [23]. The number of interacting partners, connectivity, has been chosen as 20. The study also revealed that as connectivity decreases sensitivity of the method decreases.

3.4 Hub Classifier

There exists another method called Hub classifier which uses Gene Ontology terms with 84.96% accuracy, 34.41% sensitivity and 90.27% specificity [5]. For using this method to predict whether a target protein is Hub or not gene ontology annotation of the target protein is needed. Michel Hsing et al [1] have stated that the performance of Hub

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

classifier will primarily rely on the number of Gene Ontology annotations available for each species. According to them, the reason for the low sensitivity is "the lack of gene ontology annotations for certain proteins in the training sets" [5].

Three existing methods given have their own problems. In the proposed method we are trying to address these problems using advanced data mining tools and cutting edge sequence based feature sets.

IV.DATA SET

In this paper two sets of data have been used to test the proposed method. The first set is taken from HPRD [24]. From the database whole set of human protein ID"s were obtained. In this database there were 27080 human proteins. Among them 9630 have interactions with others. This information is presented in the form of binary interactions in the database. From this, it was possible to find the count of number of interactions of each protein. This count is taken as the degree of connectivity of the protein and it ranged from 0 to 267. The table 3 below gives the number of proteins having the degree of connectivity k in HPRD database.

TABLE III. Degree of connectivity and number of proteins in hprd database

Degree of Connectivity (k)	Number of proteins
0	17450
1	2237
2	1424
3	1009
4	759
5	618
6	468
7	422
8	287
>> 8	2406
Total	27080

It can be seen from the table III that as the value of k increases the frequency of the protein decreases. Using this information as a frequency table, its mean was calculated and was obtained as 8.0557. It is again evident from the table III that frequency count of the proteins with $k < 9$ is 7224. That is there are 2406 proteins with $k > 8$ which is around 25% of the total interacting proteins. In the proposed method the threshold for connectivity of hub proteins for this database is taken as 8 based on the above analysis.

The second set is taken from the interacting protein set of P53 protein[25]. In this database there were 108 proteins which have shown interaction with P53. The sequence information is obtained from the NCBI and UniProt Database [26].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

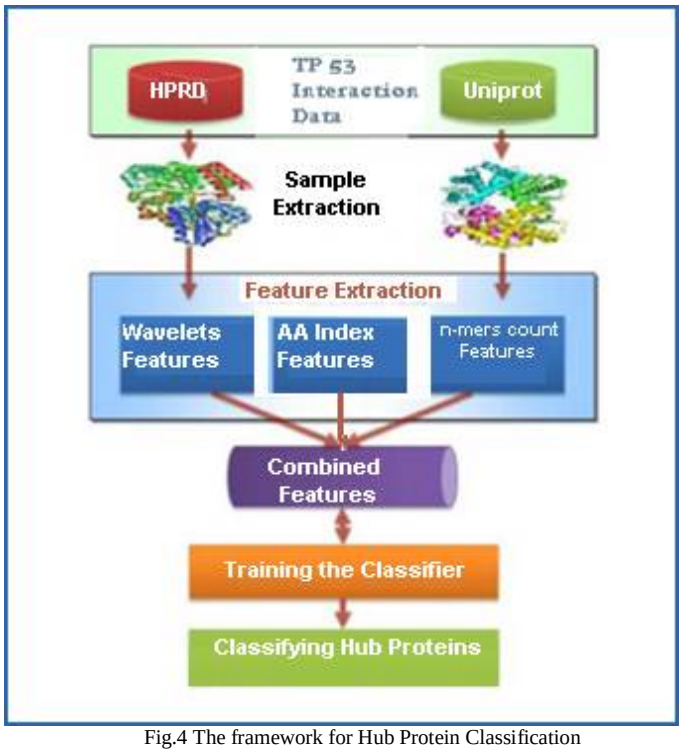

Fig.4 The framework for Hub Protein Classification

V.PROPOSED METHOD

In this section, we introduce an empirical study in which the wavelet analysis is applied to solve one important problem in proteomic research which is the prediction of hubness in human proteins using tumor proteins. We evaluate the effectiveness of the features computed using wavelet analysis and discuss some insights based on the experimental results.

5.2 A Unique Computational Framework

Fig. 4 shows the architecture of the framework. First, the Hub proteins are collected from existing knowledge and downloaded from NCBI, Uniprot and GenBank [27]. Next, the protein samples are extracted according to the window size on the corresponding protein sequences, and those samples are represented by numerical numbers according to a certain mapping scheme. Then, three sets of features are extracted using the samples obtained in the previous step. Finally, a classification technique, SVM-based classifier, is applied to classify the hub and non hub proteins. The steps are discussed as follows:

Step1. Data Collection

a. Hub and Non hub protein data is collected from HPRD data set based on connectivity index
b. P53 interaction data is collected from Uniprot database [28].

Step2. Sample sub sequence extraction using length 100 as size from both data sets.

a. Find the dense area in the sequence of size 10 amino acid residues. A dense area in a protein sequence is the substring s whose sum of hydrophobic value is greater than the average of the hydrophobic value of all amino acids [29], which is calculated as 13 and set as the threshold. Here the length of the substring is set as 10 which is the window size.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

b. 45 amino acids are extracted from the left and right side of the dense area as the size of subsequence is set as 100 for the purpose of feature extraction.

Step 3. Numerical Representation

The original amino acids in the samples extracted are converted to numerical representation based on the mapping scheme in Table 2. In this experiment, only the real component of the complex representation is used [16].

Step 4. Feature extraction – 1044 features were derived from amino acid sequence as part of this step

a. Feature Set I using Wavelet analysis – Using this technique 100 features are derived in the form of wavelet coefficients. The Matlab wavelet toolbox provides a powerful tool for wavelet analysis. In the current experiment, the continuous wavelet transform based on Daubechies wavelets function is used to extract wavelet coefficients from sample sequence data. (The Daubechies wavelets are chosen due to their successful applications in biological sequences analysis [20], [30].) The scales are set to be 2:2:100, where the second 2 represents a sampling step of 2 (similar to the example illustrated in Section 2.5). The obtained COEFS are a 50 by 100 matrix, where each row is a coefficient sequence at a specific scale. The averages of the rows of the coefficients in the matrix are calculated to obtain a 100-dimensional
feature vector.

b. Feature Set II using AAIndex - In addition to the wavelet features, 544 amino acid index (AAindex) features [31] that represent the physicochemical properties of the proteins are also extracted.
c. Feature Set III using 2-mers count - 400 features are generated using Matlab function called n-merscount .This function returns a vector consisting of 400 features as the value of n is set as 2.

Step 5. Classification

Support vector machine is used as classification model. The LIBSVM package [32] is one of the most popular off-the-shelf classifiers. In this study, the LIBSVM classifier is utilized as the classification model

Step 6. Result Evaluation

For the purpose of results evaluation, the "Accuracy," "F1," and "Matthew's correlation coefficient" (MCC) performance metrics are used. Here, "TP" is the total number of true-positive data, "TN" is the total number of true-negative data, "FP" is the total number of false-positive data, and "FN" is the total number of false-negative data. In addition, MCC ranges from -1 to 1. A value of MCC = 1 indicates the best prediction possibility; while MCC = -1 indicates the worst prediction possibility. MCC = 0 is expected for a random prediction. The equations for different criteria are given below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN},$$

$$F1 = \frac{2 \cdot \frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}},$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TN+FN)(TP+FN)(TN+FP)}}$$

VI.EXPERIMENTAL RESULTS

Experiments are conducted to evaluate the contributions and characteristics of five different groups of features. Table 4 shows the group IDs and their corresponding features. The LIBSVM classifier is utilized to evaluate those different groups of features.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

TABLE IV . FEATURE SET AND IT'S SIZE

Group Sl. No	Feature Set	Size of Feature Set
1	Daubechies Wavelets	100
2	AA Index	544
3	2-mers count	400
4	AA Index + Daubechies Wavelet	644
5	AA Index + Daubechies Wavelet+ 2-mers count	1044

TABLE V . RESULTS

Group Sl. No	Accuracy	F1	MCC
1	0.778302	0.618474	0.559625
2	0.858491	0.628378	0.717492
3	0.792453	0.618321	0.585845
4	0.859774	0.628986	0.707831
5	0.839623	0.625	0.679366

From the experimental results shown in the Table V, it could be seen that the AAindex features (Group 2)outperform the Daubechies wavelet features (Group 1) and the n-mers count features (Group 3). The reasons are as follows: First, the dimension of the AAindex features is 544 but the Daubechies wavelet feature and n-mers count features are only of 100 and 400 dimensions respectively. The AAindex features contain more information. In addition, each dimension of the AAindex features represents one kind of physiochemical properties, which determines the protein structure that is related to the hub function based on the classic biological assumption that the structure and property of the protein determine its biological functions.

The wavelet transform captures relatively indirect features of the protein sequence. In terms of Daubechies wavelet features and n-mers count, their performances are comparable and it is not clear which one outperforms the other. However, when the AAindex features are combined with the Daubechies wavelet features, the performance is improved compared to that using n-mers count features or Daubechies wavelet features individually. The same improvement is also seen with group 5 where AA Index, Daubechies Wavelet and n-mers count are used together.

It also suggests that even though the Daubechies wavelet features themselves donot give good performance, they could be utilized to enhance the AAindex-based features. It is easy to draw the conclusion that if a feature set with good performance is combined with the one with worse performance, an average performance is achieved. The reason is that the AAindex feature, which captures the global feature of the protein sequence loses all the information about the sequence position. However, the sequence of the protein also determines the properties of the proteins.

The wavelet-based features capture the sequence or the temporal information of the proteins and complement the AAindex features. The complete results including evaluation criteria, and the values obtained from the experiments are also shown in the table 5 namely Accuracy, F1, and MCC.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

VII.CONCLUSION AND FUTURE WORK

In this paper we have tried to apply wavelets in biological sequences for the purpose of hub feature extraction. First an overview of Protein Interaction Network and a sample PIN is given to understand the basics of hub proteins. We identify three important steps, which are numerical representations of protein sequences, feature extraction and classification based on feature sets extracted, and classification based on 3 sets of features derived. The numerical representation of biological sequences is very important in the success of the overall framework and is an active research area. Different approaches are described in detail in the background section so that researchers in this domain could refer to these methods. different state-of-the-art wavelet analysis methods are introduced in this Section. The applications of wavelets in solving many critical biological problems have motivated us to apply the same in our classification problem using the wavelet coefficients.

According to the foundations built in previous sections, a detailed description of applying wavelet analysis in protein classification is given in Section 2.4 to illustrate its usage in cancer research. These information show that a proper combination of the wavelet coefficient-based features and protein physico-chemical property-based features enhances the classification accuracy. In the future, the most vital task will be to enhance the numerical representation of the protein sequence and the scheme of applying the wavelet transform. Other wavelet transforms, such as Meyer, Haar, Morlet, and Mexican Hat can be considered and the detailed comparison of the performance of using different wavelet-based features can be accomplished. In addition, another research direction is to integrate the rising importance of p53 pathway and interaction set for cancer research related protein sequence information processing.

REFERENCES

1. Khalid Raza, “Application Of Data Mining In Bioinformatics”, “ Indian Journal of Computer Science and Engineering Vol 1 No 2, pp. 114-118, 2010.
2. Chad Haynes. “Intrinsic Disorder is a Common Feature of Hub Proteins from Four Eukayotic Interactomes”, PLOS Computational Biology, Vol 2, Issue 8, pp. 23-27, 2006.
3. Rual JF. “Towards a proteome scale-map of the human protein-protein interaction network”, Nature, 437, pp. 1173-1178, 2005.
4. Kota Kasahara et al, “Ligand-binding Site Prediction of Proteins Based on known Fragement-Frgament Interactions”, Structural Bioinformatics, Vol 26, pp 12-18, 2010.
5. Michael Hsing, Kendall Grant Byler and Artem Cherkasov, “The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks”, BMC Systems Biology, Issue 2, pp.110-120, 2008.
6. Jingkai Yu, Russell L. Finley “Combining Multiple Positive Training Sets to Generate Confidence Scores for Protein- Protein Interactions”, Bioinformatics, Vol. 25, pp. 105-111, 2009.
7. Ideker T, Sharan R “Protein Networks in Disease”, Genome Res, 18, pp. 644-652, 2008.
8. Peter Murray-Rust, “Bioinformatics and Drug Discovery”, Current Opinion in Biotechnology, Volume 5, Issue 6, pp. 20-29, 1994
9. Alexei Vazquez, Elisabeth E Bond, Arnold J Levine, Gareth L Bond. “The genetics of the p53 pathway, apoptosis and cancer therapy”, Nature Reviews Drug Discovery Volume: 7, Issue: 12, Publisher: Nature Publishing Group pp. 979-987, 2008.
10. Sriganesh Srihari et al, “Detecting Hubs and Quasi Cliques in Scale-free Networks”, IEEE, Vol2 , pp. 201-212, 2008
11. Barabasi, A. L. and Oltvai Z N, “Network biology: understanding the cell's functional organization”, Nat Rev Genet 5(2): pp. 101-113, 2004.
12. Albert R., “Scale-free networks in cell biology”, J Cell Sci., Vol 21: pp. 4947-4957, 2005.
13. Carol Prives1, Peter A. Hall, “The p53 pathway”, The Journal of Pathology, Special Issue: Molecular and Cellular Themes in Cancer Research, Vol 5, pp. 112–126, 1999.
14. Olch et al, "The patter of p14ARF expression in primary and metastatic human endometrial carcinomas : correlation with clincicopathological features and TP53 pathway alteration", International Journal of Gynecological Cancer; Issue 6: pp. 993-999, 2010
15. Sandra L Harris and Arnold J Levine, “The p53 pathway: positive and negative feedback loops”, Oncogene – Nature, Vol 24, pp. 2899–2908, 2005,
16. Tao Meng, Ahmed T. Soliman, Mei-Ling Shyu, Yimin Yang, Shu-Ching Chen, S.S. Iyengar, John S. Yordy, and Puneeth Iyengar, “Wavelet Analysis in Current Cancer Genome Research: A Survey “,IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 10, pp. 201-211, 2013
17. S.G. Mallat, “A Theory for Multiresolution Signal Decomposition: The Wavelet Representation,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pp. 674-693, 1989.
18. I. Daubechies, Ten Lectures on Wavelets, series CBMS-NSF Regional Conf. Series in Applied Math. Soc. for Industrial and Applied Math., Vol 9, pp. 20-29, 1992.
19. D. Gabor, “Theory of Communication,” IEEE Radio Comm. Eng. J.,vol. 93, pp. 429-441, 1946.
20. M. Sifuzzaman, M.R. Islam, and M.Z. Ali, “Application of Wavelet Transform and Its Advantages Compared to Fourier Transform,” J. Physical Sciences, vol. 13, pp. 121-134, 2009.
21. Sumeet Agarwal et. al., “Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks”, PLOSComputational Biology, Volume 6, Issue 6, pp 134-142, 2010.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

22. Kim PM, Lu LJ, Xia Y, "Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights", Science, Issue 6, pp. 1938–1941, 2006.
23. Ramon Aragues et. al., "Characterization of Protein Hubs by Inferring Interacting Motifs from Protein Interactions", PLOS Computational Biology, September, Volume 3, Issue 9, pp. 147-156, 2007.
24. Qi, Y, Bar-Joseph, Z. and Klein-Seetharaman, J., "Evaluation of different biological data and computational classification methods for use in protein interaction prediction",Proteins, Vol 6, pp. 490-500, 2006.
25. P53 Interactions, Retrieved September 2011, from http://en.wikipedia.org/wiki/
26. Rolf Apweiler, Amos Bairoch et. Al., "UniProt: the Universal Protein knowledgebase ", Nucleic Acids Research, Vol. 32, D119-D119, 2004.
27. D.A. Benson, I. Karsch-Mizrachi, K. Clark, D.J. Lipman, J. Ostell, and E.W. Sayers, "Genbank," Nucleic acid research, vol. 39, pp. D32-D37, 2011.
28. Sajeev J., Dr. T. Mahalakshmi, "Web tool form Protein Sequence extraction and Feature using Uniprot", Juornal of Management and Innovative Information Technology, Vol 1, pp. 1-4, 2016.
29. Rakesh Kumar Agrawal et.al., "A novel approach to predict protein-protein interaction using protein sequence data", Bioinformatics Trends, Volume 1, Issue 1, pp. 14-25, 2006.
30. J.K. Meher, M.K. Raval, P.K. Meher, and G.N. Nash, "Wavelet Transform for Detection of Conserved Motifs in Protein Sequences with Ten Bit Physico-Chemical Properties," Int'l J. Information and Electronics Eng., vol. 2, no. 2, pp. 200-204, 2012.
31. S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "AAindex: Amino Acid Index Data, Progress Report 2008," Nucleic Acids Research, vol. 36, pp. D202-D205, 2008.
32. C. Chang and C. Lin, "Libsvm : A Library for Support Vector Machines," ACM Trans. Intelligent Systems and Technology, vol. 2,no. 27, pp. 1-27, 2011.

BIOGRAPHY

Sajeev J is a Associate Professor in the Department of Computer Applications, Sree Narayana Institute of Technology, Kerala. He received Master of Computer Application (MCA) degree in 2002 from Bharathiar University, Coimbatore, TamilNadu, India. His research interests are Bioinformatics, Web Mining, Text Mining and Social Network Mining.

Dr. T. Mahalekshmi is Professor in Department of Computer Applications and Principal in Sree Narayana Institute of Technology, Kollam, Kerala. She received Ph. D. from Kerala University. Her research interests are Bioinformatics, Medical Informatics and Data Mining.