DOI: http://dx.doi.org/10.26483/ijarcs.v9i3.6108

ISSN No. 0976-5697

Volume 9, No. 3, May-June 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

IDENTIFICATION OF HOTSPOTS IN PROTEIN SEQUENCES USING CPNR AND DWT

G. Anitha Mary
Asst. Prof, Research scholar (Ph.D. Part - time)
Dept.MCA,Loyola AcademyDegree,PGCollege
(Sri Venkateswara University, Tirupati,A.P,India)
Old Alwal, Secunderabad-10
Telangana, India

G. Anjan Babu
Professor & Head
Dept. of Computer Science
Sri Venkateswara University
Tirupati, Chittoor(Dt)A.P, India

G. Aparna Raghava Rao, ResearchScholar
Dept.of Electronics &Communication
Engineering,University College of Engg
(UCOE),Osmania University, Hyderabad
Telangana, India.

***Abstract*:** Protein-protein interactions controls most of all the biological activities and therefore the key functions of proteins. The interaction layout of protein rely on the organic compound sequence. Computationally assessing the purposeful affinities between proteins is a crucial job of bioinformatics analysis. It will favor molecular biologists share information on few proteins to others and thus scale back the quantity of tedious and valuable bench work. Hence, identification of protein performance from its primary sequence may be an important and tiring task in bioinformatics.Identification of the amino acids (hotspots) that ends up in the characteristic frequency signifying a particular biological perform is de facto associate in nursing annoying task in Genomic signal process. Since experimental procedures of protein hotspot identification are still financially very rigorous and time taking, there's a wrench to supply enough reliable process procedures for this specific task. Signal process demands the sequence to be in numerical illustration, therefore protein sequences are mapped (encoding) into digital signal. Amino acids are the building blocks of proteins plays a major role achieve this job.For signal processing the sequences need to converted into numerical sequence, for which mapping is necessary.From literature experimented results incontestable among EIIP and CPNR mapping, CPNR achieves better performance than EIIP. So, CPNR mapping is taken into account. With the recent advances in Genomic signal process (GSP) domain, researchers are applying digital signal process (DSP) techniques in raw genomic information for extracting the hidden features among proteins. Wavelet transformation has been a really novel strategy for the analysis and process of non-stationary signals like bio signals within which each time and frequency data area unit required. In this paper, we have incorporated complex prime numerical representation (CPNR) used for mapping of protein sequence into digital signal form and discrete wavelet transform (DWT) strategies to spot the hotspots. A new approach using CPNR & DWT is introduced.

Keywords - Complex prime numerical representation (CPNR), discrete wavelet transform (DWT), protein sequence, and Genomic signal process (GSP).

I. INTODUCTION

One of the foremost vital challenges within the past genomic amount is that the characterization and analysis of protein-protein interactions in living organisms, as these necessary for shaping of normal and abnormal behaviors in cells. The foremost majority of proteins bind one another in their lifetime so as to perform completely different functionality. Protein - protein interactions play a very important role in cellular perform to create the backbone of the many biological activities. Though the principles existing for protein interactions aren't fully understood, it's accepted that majority of the energy in associate interaction is given by little portion of the overall range of amino acids. These amino acids are termed as hotspots that seems to be clustered in tightly packed regions within the protein interfaces, and area unit discovered to be necessary for preserving protein perform and maintaining the stability of protein association. The protein-protein interaction or hotspot identification provides a baseline to spot and analyze the precise residues liable for majority of abnormalities. The identification of protein hotspots may be a difficult job for researchers in engineering, biology and computing. Therefore, reliable and economical techniques for distinctive hotspots positions in proteins are needed.From literature it's understood that the raise of DSP techniques within the hotspots in protein sequences are booming, however the potency of the techniques got to be raised. In current work, protein sequences are analyzed to resolve the matter of locating the hotspots, to seek out the hotspots in an exceedingly protein sequence DSP techniques are used, initially protein sequence got too mapped into numerical sequence to convert it into signal. DWT is applied for feature extraction of hotspot[1].

II. OVERVIEW OF PROTEINS

Proteins are cluster of proteins synthesized from ribonucleic acid by the process known as translation. They're to cause for finishing up most of the cellular activities and most of the metabolic activities of cells. The protein sequences are long chains of amino acids joined by amide bonds observed as peptide chains. There are twenty amino acids and are drawn in a very protein sequence as a string of alphabetical symbols with typical lengths ranging from one hundred to ten thousand.Proteins have an inclination to fold into 3 dimensional (3D) structures. The 3D structure of protein is vital as a result of the structure is joined with the biological operate. The method of folding is extremely complicated, in which a polypeptide chain attains a stable 3D structure through short and long vary chemical interactions between amino acids that are near and in numerous parts of the molecule, severally [2].During this folding method, the peptide chain twists and bends till it achieves a state of minimum energy that maximizes the steadiness of the ensuing structure. This 3D form permits the protein to act with different molecules referred to as targets at specific sites that are remarked active sites of protein. These active sites have specific shapes in order that they'll match into the target molecules throughout their

interaction during a method analogous to a hand fitting into a glove as shown in Figure 1. In and around these active sites are sub regions referred to as hotspots that are chargeable for each the chemical stability of active sites in addition as provision the energy for the protein-target interactions. A hotspot could contains one or a lot of amino acids organized in a very distinctive pattern within the protein sequence. Because the hotspots play a vital role in sanctioning proteins to perform their functions, a thorough data concerning their locations is important for understanding protein perform.

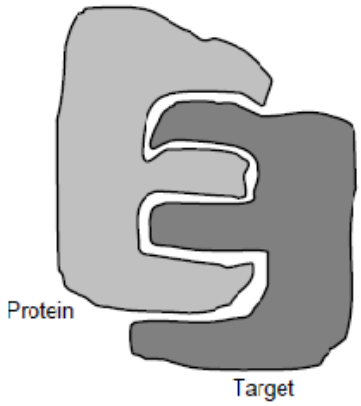

Figure 1: Illustration of a protein fitting into its target

The seek for protein functions provides the identification and characterization of every protein still as in-depth data concerning their interaction with alternative proteins and deoxyribonucleic acid molecules. The protein target interaction is extremely specific in nature. The protein binds to the target in an identical manner as a key fits to the corresponding lock. Figure 2 shows the interaction interface of protein A and protein B. The teams of amino acids at this interaction interface measure referred to as hotspots. It's well established that the hotspots exhibit a characteristic frequency adore their functional operation. Knowing the characteristic frequency of a selected hotspot, new similar hotspots are often foretold in alternative unannotated protein sequences.

Often, cancer causing mutations cluster in hotspots wherever tumors from completely different patients harbor a similar continual mutation. Some hotspot mutations could occur often, whereas others measure rare. In cancer, corporeal driver mutations usually target hotspots of paralogous residues across evolutionarily connected members of one protein family. These hotspots usually represent the foremost drug gable genetic alterations.

III. PROBLEM FORMULATION

The real life protein sequences measure typically characterized by clattery signals and therefore, the employment of existing filtering strategies to such sequences does not offer correct results. In these applications, the time-frequency analysis and filtering measure needed to extend accuracy. Since wavelet transform possesses superior time-frequency resolution in addition as frequency detection capability, the motivation of this work is to propose discrete wavelet filtering to spot the corresponding hotspot locations. To apply suitable DSP strategies for the analysis, the character string of protein have to be compelled to be regenerate to an appropriate numerical sequence. This is often achieved by distribution of assigning a numeral to every organic compound that forms the protein.

It is fascinating to possess the conversion formulas protein and target share a similar characteristic frequency, however measure supported some physical properties that square measure relevant to the protein's biological functioning. Once numerical sequences obtained, DWT filtering is applied to discover hotpots. It's proposed to use discrete wavelet filtering to do this task. In protein-target interaction, each the opposite in section. This is believed to produce a resonant recognition within the binding method. Hence, the energy such as this frequency are going to be most. This is the fundamental conception that has been used for hotspot identification in protein sequences [3][4].

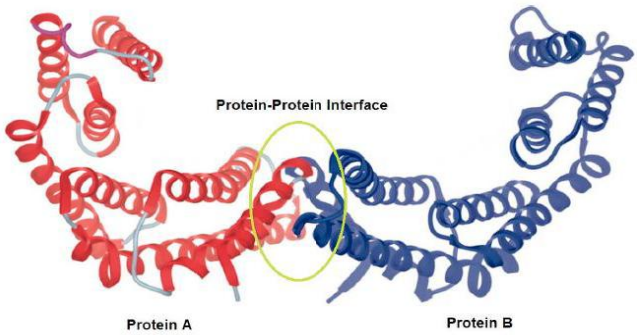

Figure 2: Illustration of Protein-Protein interaction render to hotspots

IV. METHODOLOGY

The wavelet provides energy distribution throughout the sequence. It's better-known that the sequences similar to characteristic frequency turn out larger magnitude coefficients in wavelet domain. Hence, the separate wavelet filtered output produces completely different intense energy areas in the time-frequency plot. The whole procedure for DWT based hotspot identification system is disclosed in Figure 3. The full step by step procedure for identification of hotspots is as follows:

Step 1: The input protein sequences are obtained from the standard databases such as NCBI website [5].

Step 2: The protein characters in that sequences are converted into numeric sequences using CPNR mapping scheme.

Step 3: The numeric sequences are applied to narrow bandpass filters. This is done to get the energy components in characteristic frequencies.

Step 4: Filter the protein numerical sequence of interest by using discrete wavelets. The wavelet coefficients have the information about the signal energies available in hotspots.

Step 5: The peaks in the plot represent the positions of the hotspots.

A. ALGORITHM

Figure 3 depicts the flow chart for evaluation of the new approach**.**

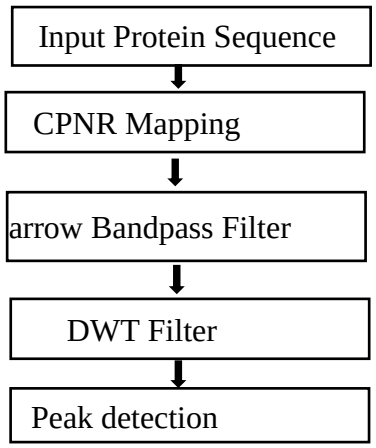

Figure 3: Flow diagram showing the DWT based hotspots identification scheme

B. CPNR BASED MAPPING

Protein sequence consists of unique twenty amino acids. Correct illustration of amino acids is vital to extract correct comparison of protein functions. Compared to alphabetical illustration of amino acids, numerically illustration of amino acids will massively expand the power to research them. Among the progressive procedures, electron-ion interaction potential (EIIP) could be a default numerical illustration approach littered with degeneracy problems that two functionally completely different proteins could get terribly similar or maybe constant illustration. The degeneracy issue is its style during which some amino acids maybemapped to same numbers. To tackle this challenge in literature a replacement procedure, Complex prime numerical Representation (CPNR) of amino acids is evolved.CPNR is taken into account supported the amount of codons of amino acids. They discovered a good similarity between the amount of codons of amino acids and a differential pattern of prime numbers [6]. To form the illustration a lot of biologically important, prime numbers are mapped into complicated domain. Compared with EIIP, CPNR has no degeneracy drawback, and also the numbers in CPNR are comparatively dependent from one another, which implies that one range can't be generated from another by addition, multiplication, or involution with a correct number shown in table1.

Table 1: Mapping of Amino Acids using CPNR

S.NO	AMINO ACIDS	CPNR
1.	M	1
2.	W	2
3.	F	3
4.	Y	5
5.	P	7
6.	C	11
7.	T	13
8.	H	17
9.	V	19
10.	L	23
11.	Q	29
12.	S	31
13.	A	37
14.	N	41
15.	G	43
16.	R	47
17.	I	53
18.	D	59
19.	E	61
20.	K	67

For example Human Fibroblast Growth Factor (HFGF) protein sequence is considered as input to the procedure with length l=146 and its CPNR shown in figure4**.**

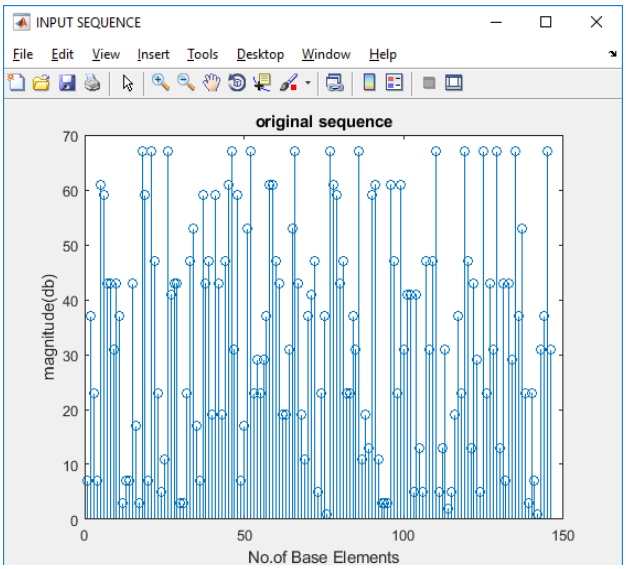

Figure 4: CPNR representation of Human Fibroblast Growth Factor

C. FEATURE EXTACTION TECHNIQUE

An advanced evolving discipline in bioinformatics known as Genomic signal process (GSP), that method with logical analysis of genomic information by involving digital signal process (DSP) techniques. Genomic information such protein sequence that exist in character type mapped into digital type known as genomic signal. Bioinformatics is associate degree knowledge base space that contains of process procedures for quick and correct illustration of biological information [7]. Wavelet transform has been a really reliable methodology for the analysis and process of non-stationary signals like bio-signals within which each time and frequency details area unit needed. Protein sequences area unit associate array of twenty alphabets representing twenty amino acids with random prevalence of every alphabet. So as to use DSP techniques, the initial task is to convert protein sequences into numeric values, next step to proceed with band pass filter and then application of DWT to extract energies of hotspots.

Wavelet transform is a reliable method for theanalysis and processing of non-stationary like biosignalsin which time and frequency information isrequired. The wavelet transform gives solution to the problem of time& frequency resolution because of multi resolution strategy of wavelet transform [8]. Wavelets are well localized inboth time & frequency domain. Wavelet analysis can do compress or de

noise a signal without appreciablereduction. Wavelet analysis is the dividing of a signalinto shifted and scaled versions of the original (or mother)wavelet. Wavelet transform is given in continuous & discrete domain.

The parameters a and b represents dilation and translations respectively. Ψthe mother wavelet. The DWT is computed by successively passing discrete time domain signal through band pass and stop band filters. DWT coefficients will have energies of hotspot signals required.

V. DATASET

For testing proposed algorithm the dataset used shown in table2 are collected from PDB database

Table 2: Details of protein sequences used for analysis

Organism	Protein Name	PDB ID	Sequence Length	Characteristic Frequency
Human	Fibroblast growth factor(FGF)	4fgf	146	0.904
Human	Human Growth Hormone(HGH)	3hhr	190	0.270
Bacteria	Barstar	1brs	89	0.321
E.Colo	Colicin-E9 Immunity IM9	1bxi	86	0.190
Bacteria	Barnase	1brs	110	0.321
Bacteria	TRAP	1wap	75	0.247
Human	Human alpha hemoglobin	1vwt	142	0.023

VI. RESULTS

Prediction of hotspots using CPNR & DWT for Human fibroblast growth factor protein sequence is given in figure5 and details of hotspots identified for the dataset used are shown in table3. We have considered the hotspots obtained fromthe alanine scans (ASEdb) as the reference for comparison from literature[9].

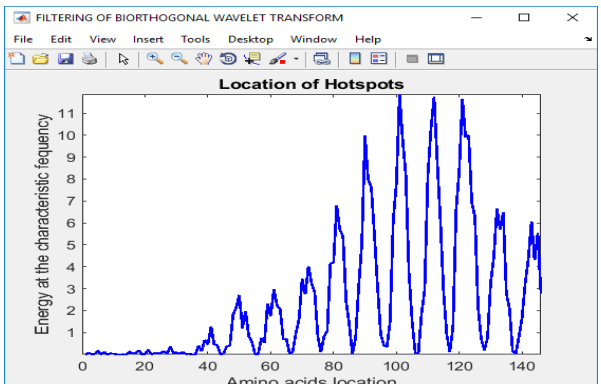

Figure 5 Output of the CPNR&DWT for Human fibroblastgrowth factor protein

Mathematical formula Continuous wavelet transform CWT) describe as

$$C(a,b) = <\psi_{a,b}(t),x(t)> = \int_{-\infty}^{+\infty} x(t)\,1/\sqrt{a}\,\psi((t-b)/a)dt$$

Table 3 Details of inputs and identified hotspots using CPNR&DWT basedscheme

Input Sequence	Hotspots Identified using Proposed method
Fibroblast growth factor	5,17, 24,28,41,50,61,72,81,96,103,111,121,134,140
Human Growth Hormone	4,11,125,129,133,141,145,148,151,155,159,163,168,171, 172,178179,188
Barstar	3,6,9,12,15,18,22,25,29,31,35,39,42,44,47,50,53,56,59,62,66,69,72,76,78,81,84,87
Colicin-E9 Immunity IM9	5,9,15,21,25,33,34,41,45,50,51,55,64,68,71,79
Barnase	3,7,10,16,23,27,32,35,41,43,46,52,58,60,63,67,70,73,77, 80,83,87,90,93,96,99,102 106,108
TRAP	4,7,13,16,20,25,29,33,37,40,45,49,56,58,61,65,69,73
Human alpha hemoglobin	14,18,20,22,28,36,56,59,63,95,98,100,103,117,126,139, 141

VII. CONCLUSION

For the current work,MATLAB is used to develop the proposedalgorithm using bioinformatics and signal processing toolboxes.A discrete wavelet primarily based filtering apprroach has been proposed for the identification of the hotspots. This methodology doesn't need information regarding the structure of protein advanced and hence it is effectively used in hotspot identification wherever the interface region is not noted. Furthermore this finds applications in the growth of enhancement of medicines.

VIII. REFERENCES

[1] George TP, Thomas T, "Discrete wavelet transform de-noising in eukaryotic gene splicing," BMC Bioinformatics, 2010.

[2] Rafale C. Gonzalez, Richard E Woods's _Digital image processing, Second Edition page 410.

[3] Tuncbag, N, Gursoy, A &Keskin, O 2009, 'Identification ofcomputational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy',Bioinformatics, vol. 25, no. 12, pp. 1513–1520

[4] Yadav, Y &Wadhwani, S 2011, 'Determination of characteristicfrequency for identification of hot spots in proteins', InternationalJournal of Electrical and Electronics Engineering (IJEEE), vol.1, no.1,pp. 1-4

[5] National Centre for Biotechnology Information (NCBI).Available: http://www.ncbi.nlm.nih.gov/.

[6] DUO CHEN, JIASONG WANG, MING YAN, and FORREST SHENG BAO, 'A Complex Prime Numerical Representation of Amino Acids for Protein Function,

[7] Shu-ching Chen, "Wavelet analysis in current cancer Genome Research: A Survey", IEEE /ACM transactions on computational biology and bioinformatics, vol. 10, pp. 567-570, 2013.

[8] P. P. Vaidyanathan and B. J. Yoon, "The role of signal processing concepts in genomics and proteomics," Journal of the Franklin Institute, Vol. 341, pp. 111-135,

[9] Sahu, SS & Panda, G 2011, 'Efficient localization of hot spots inProteins using a novel S-transform based filtering approach',IEEE/ACM Transactions on computational biology and bioinformatics,vol.8, no.05, pp. 1235-1246