Determination of Hot Spots in Protein Sequence using Digital Filter

Ashwini T. Walekar[1], **A. S. Shirsat**[2]

[1]PG Student, E&TC Department, Smt. Kashibai Navale College of Engineering, Vadgaon (Bk), Pune, India

[2]Assistant Professor, E&TC Department, Smt. Kashibai Navale College of Engineering, Vadgaon (Bk)
Pune, India

Abstract: ***The method is used to identify the hot spots from the protein sequence which gives the characteristic frequency. Protein sequence belonging to the functional group of interest is converted into equivalent numerical sequence by assigning EIIP value to each amino acid in the protein sequence. DFT's of the numerical sequence is computed to obtain their consensus spectrum which leads to the characteristic frequency. This characteristic frequency is selected by filtering the numerical sequence using a specialized narrowband band-pass digital filter. Energy peak is computed from the filter output to select the dominant characteristic frequency from the numerical sequence. These energy peaks from the filtered proteins numerical sequence gives the locations of hot spots. MATLAB is used for implementation of this technique.***

Keywords: Protein, characteristic frequency, amino acids, FFT, EIIP, RRM, hot spots, digital filter.

1. Introduction

Proteins are nothing but the most complex chemical entities in living organisms. No other class of molecule exhibits the irregularity in size, variety, shape, mobility and texture that can be found in protein as protein structures are inherently complex. Protein has different levels of structures having progressively greater complexity. The simplest level of the protein is nothing but its Primary structure. Amino acid sequence itself is the primary structure of a protein. Proteins are not just letters printed on a page, but in reality, they are formed by different combinations of amino acids covalently connected together by peptide bonds. Hence, amino acid chain is called as polypeptide chain [2].

An amino acid is a combination of carboxylic acid group, an amino group and a variable side chain, all these attached to a central carbon atom which is known as α- carbon. From these, side chain is the only parameter that varies from one amino acid to another amino acid which is responsible for the chemical variety of the amino acids. Theoretically number of different amino acids is possible but only 20 of them are commonly found in proteins of all living organism [3]. Proteins are represented in a certain active three-dimensional (3D) to express their biological function [4]. The order of amino acid in a protein sequence itself gives its primary structure. 3D structure of a protein molecule is obtained by using its primary structure. It is difficult to code the biological function and the 3D structure and of a protein into its primary structure. After solving this problem, it will be possible to create artificial proteins by carefully combining the amino acids which has desired functions. These artificial proteins are used to cure diseases such as paralysis, heart ailments, tumor and cancer [3].

The organization of paper is as follows. Section II gives literature survey of various methods for location of hot spots from protein sequence. Section III describes proposed technique and various terms used in the technique. Choice of filter and online databases are given in section IV and V respectively. Results and discussion for the given technique are described in Section VI, followed by conclusion and future scope in Section VII and Section VIII respectively.

2. Literature Survey

Different techniques are used to locate the hot spots from the protein sequence. These techniques are explained bellow.

2.1 Site-Directed Mutagenesis

This is the most widely used experimental technique for location of hot spots form the protein sequence. In this method, the amino acid at specific locations in a predefined protein sequence is replaced by the other type of amino acid. These replacements are known as mutations. The biological properties of protein may change due to these mutations, if the specific amino acid location plays a very crucial role in a protein's biological function. Protein biological function affects considerably due to these mutations. Hence, we can mark this specific location of amino acid as hot spot. W0e can obtain all the hot spots from the protein sequence by repeating this procedure for each location of suspected amino acid [3].

Generally alanine amino acid is used for the replacement. Hence, the name is *Alanine Scanning Mutagenesis* (ASM). Alanine contains methyl group (-CH3) as the side chain which makes it simplest amino acid with respect to molecular structure, hence it is chosen for the replacement. Methyl group does not involve directly in the proteins function as it is non- reactive in nature. This technique is not appropriate for a suspected hot spot alanine amino acid. This is major limitation of the technique [3].

Licensed Under Creative Commons Attribution CC BY

2.2 Short Time Discrete Fourier Transform

One of the methods for identification of hot spots in protein sequence is Short-Time Discrete Fourier Transform (STDFT). The protein numerical sequence is computed using this technique used its columns is multiplied by the coefficients of the discrete Fourier transform [5].

A numerical sequence f(k) can be obtained from any one-dimensional sequence of amino acids having length 1 by assignment of numerical values to the amino acids. For given numerical sequence F(n) denotes discrete Fourier transform sequence, where k is the position along numerical sequence and n gives the frequency which is given as,

$$F(n)=\sum_{k=1}^{l} f(k)e^{2\pi i kn/l} \qquad (1)$$

Usually F(n) is complex. This F(n) is the combination real cosine series and imaginary sine series as follows:

$$F(n)=C(n)+iS(n) \qquad (2)$$

$$F(n)= \sum_{k=1}^{l} f(k)\cos\left(\frac{2\pi kn}{l}\right)+i\sum_{k=1}^{l} f(k)\sin(2\pi kn/l) \qquad (3)$$

Locations of hot spots are identified using this method in terms of distinct peaks in the spectrogram, and achieving good location in the amino acid domain [5].

3. Proposed technique

3.1 Location of Hot Spots in Protein Using Digital Filter

Hot spots in a protein sequence are the regions in the corresponding numerical sequence where the characteristic frequency is dominant. Short Time Discrete Fourier Transform (STDFT) technique is very effective but computationally very expensive. When there is a requirement of a implementation of the hot spot location technique as a hardware system then improved computational efficiency is highly desirable. Here, digital filters are desirable to improve computational efficiency instead of the STDFT. The filter based technique deals with the designing of specialized digital filter to select the characteristic frequency from protein sequence [3].

As compared to other two methods, digital filter based technique is most suitable technique as it is more suitable for hardware implementation, computationally much more efficient and less expensive. Hence, it is necessary to use digital filter based technique for identification of hot spots from protein sequence. The steps for implementation of this technique are given as [3]:

1. Numerical sequence is obtained by converting several protein sequences of functional group of interest by assigning EIIP values to them.
2. Draw their spectrum after obtaining the DFT's of the numerical sequence so as to determine the characteristics frequency.
3. Design the digital filter which is used to select the characteristic frequency.
4. Then, compute the energy peaks of the filter output to determine the regions in the protein sequence where the characteristic frequency is dominant.
5. Finally mark the hot spots in protein sequence by locating the energy peaks in the filtered signal.

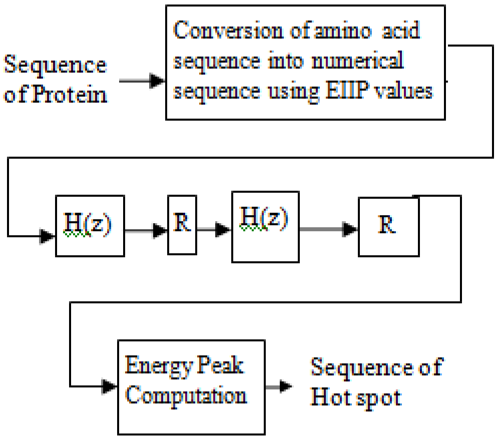
Sequence of Protein
Conversion of amino acid sequence into numerical sequence using EIIP values
H(z)
R
H(z)
R
Energy Peak Computation
Sequence of Hot spot

Figure 1: Complete hot spot location system based on digital filter

3.2 Hot Spots in Protein

The amino acid groups which dominate the proteins function are known as hot spots. Hot spots are region in the protein that plays very crucial role in the functioning of the proteins. For protein target interactions hot spots contribute to the binding energy [1]. Hot spots are the residues which shows a change in the binding free energy by less or more than 1 kcal/mol when it is replaced by alanine (ΔΔGbind). Correctly identified hot spots are residues with an observed and predicted ΔΔGbind value which is larger than or equal to 1 kcal/mol [6]. Specific regions in the target molecules and protein are called as active sites. In its configuration, an active site of a protein needs to remain stable. This group of amino acids which surrounds the active site is usually referred to as hot spots in a protein function. Therefore, efficient and reliable techniques for location of hot spots from protein sequence are required [7].

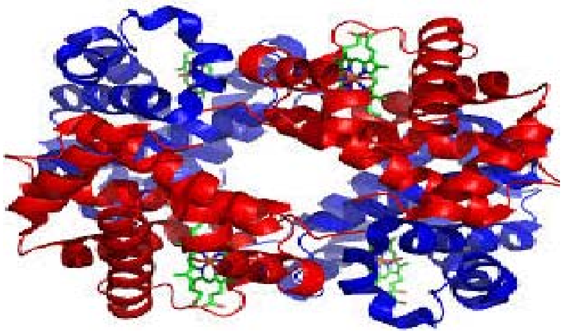

Figure 2: Interaction of protein with hot spots [7].

3.3 Resonant Recognition Model

The Resonant Recognition Model (RRM) is used to derive the proteins structural and functional information from the analysis of amino acid sequences. The RRM is a physical and mathematical model which obtains the linear information of a protein sequence using signal analysis method. In the RRM method, the proteins primary structure is represented as a numerical series by assigning a physical parameter value to each amino acid in the protein sequence which is relevant to biological activity of the protein. The RRM concept is used to define the significant correlation between biological activity of the amino acid and their spectra of the numerical presentation. Proteins having some biological function share a common frequency in their numerical spectra. This frequency is nothing but the characteristic frequency. After that it is possible to use the RRM which marks the amino acids in the protein sequence which predominantly contributed to this frequency. One RRM characteristic frequency represents particular interaction or biological function [8]. This frequency is the biological function which meets the following criteria [9]:

3.2.1 One peak only exists for a group of protein sequences which shares the same biological function.
3.2.2 There is no significant energy for biologically unrelated protein sequences.
3.2.3 For different biological functions there are different peak frequencies exist.

In all calculations, the assignment of a numerical value for each amino acid in the protein sequence is a very crucial part. The set of these numerical values should have a physical meaning which is related to the proteins biological function. The RRM method involves two main steps for calculation. The first step is to obtain numerical sequence from amino acid sequence by assigning Electron-Ion Interaction Potential (EIIP) values to each amino acid in a protein sequence which describes the average energy states of all valence electrons which presents in a given amino acid [8]. For structural and functional analysis of a protein EIIP is used which is the most suitable and known amino acid property [10]. EIIP values for each amino acid is calculated using following general model of pseudo potentials [4].

$$(\vec{K}+\vec{q}|w|k) = 0.25Z\sin(\pi\ 1.04Z)/(2\pi) \qquad (4)$$

Here, w gives the potential and q denotes the change of momentum of the delocalized electron in the interaction, where as

$$Z = (\sum Zi)/N \qquad (5)$$

Where, Zi gives the number of valence electrons of the ith component of each amino acid and N is nothing but the total number of atoms in the amino acid. The EIIP values of the 20 amino acids are given in Table I [10]. Each amino acid is represented by a unique number irrespective of its position in a protein sequence. Once obtain the numerical series, next step is to analyze this series using digital signal analysis method in order to extract information which is relevant to the biological function. Frequency domain sequence is obtained from original numerical sequence using Discrete Fourier Transform (DFT) [8].

Table 1: EIIP values of 20 amino acids

Amino acids	*Alphabet*	*EIIP Value*
Alanine (Ala)	A	0.0373
Cysteine (Cys)	C	0.0829
Aspartic acid (Asp)	D	0.1263
Glutamic acid (Glu)	E	0.0058
Phenylalanine (Phe)	F	0.0946
Glycine (Gly)	G	0.0050
Histidine (His)	H	0.0242
Isoleucine (Ile)	I	0
Lysine (Lys)	K	0.0371
Leucine (leu)	L	0
Methionine (Met)	M	0.0823
Asparagine (Asn)	N	0.0036
Proline (Pro)	P	0.0198
Glutamine (Gln)	Q	0.0761
Arginine (Arg)	R	0.0959
Serine (Ser)	S	0.0829
Threonine (Thr)	T	0.0941
Valine (Val)	V	0.0057
Tryptophan (Trp)	W	0.0548
Tyrosine (Tyr)	Y	0.0516

4. Choice of filter

4.1 Choice between IIR and FIR Filter

Various factors need to be considering while choosing the type of the digital filter that can be used for our application to assure minimal computational effort as well as accurate locations of the hot spots [3].

4.1.1 Linear Phase Response

The choice between an infinite-duration Impulse Response (IIR) and an Finite-duration Impulse Response (FIR) digital filter is dependent on whether linear phase response is required or not. This is because in FIR filters, linear phase response is easily achieved. If the requirement is a linear phase response and an application is real-time then the FIR filter must be used. But for the non-real time application IIR filters are preferred since they offered several advantages over FIR filters. Zero filtering is used to eliminate the filter delay.

4.1.2 Low Filter Order and High Selectivity

The transient response is long whenever the order of filter is high. Hence, order of filter should be as low as possible to reduce the transient response. Our aim is to attenuate all the other frequency components to an insignificant level and selects only the characteristic frequency component from the protein sequence. Hence, high selectivity is desirable. As IIR filters satisfies all the above requirements simultaneously. Also only a small amount of computation is required for their filtering due to much lower order of the IIR filters when would lead to an efficient implementation of the hot-spot location system. Hence, IIR filter is preferred for the application.

4.2 Choice among Different Types of IIR Filters

Various analog filter approximations which are most frequently used are as the Bessel-Thomson, Butterworth, Chebyshev, inverse-chebyshev, narrowband band-pass notch

Licensed Under Creative Commons Attribution CC BY

(BPNs) and the elliptic approximations. But for our application, narrowband band-pass notch (BPNs) filter is the most preferable choice as it has a minimum ripple in their pass-band amplitude response, has a good selectivity and low computational efforts involved.

5. Online Databases

Protein sequence data is nothing but the stings of alphabets, where each alphabet represents an amino acid. At various web databases, these alphabets are freely available. The protein data bank (PDB) [11], [12] and Swiss-port [13], [14] are the most important online databases of this class. The PDB deals with the 3-D structural information proteins in the form of atomic coordinates whereas Swiss-port deals with the detailed about the amino acid sequence.

6. Results and Discussion

6.1 Characteristic Frequency

Input is protein sequence of growth factor for Chicken which is taken for evaluation. Figure 4 shows maximum energy at 0.90616.This peak represents the characteristic frequency in protein sequence.

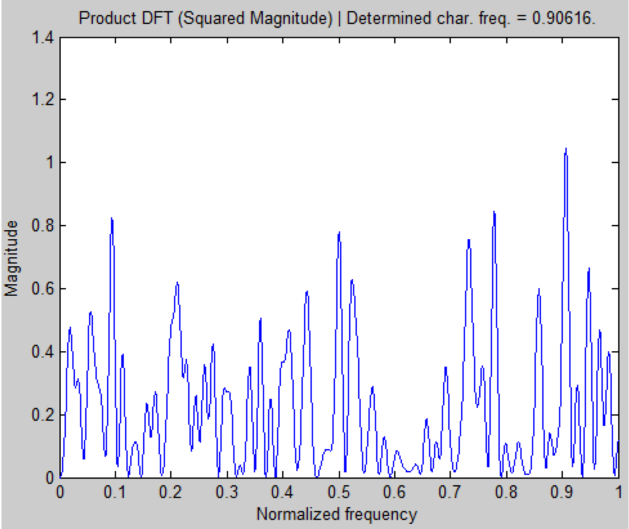

Figure 4: Energy peak in the protein sequence

6.2 Hot Spot Locations

Figure 5 shows the hot-spot locations corresponding to the Chicken. The threshold level at the average value is marked. If we set t_p=1, then only the peaks which are above the average threshold level will be designated as hot-spot locations. If the user wishes to consider all the peaks, then t_p can be set to zero.

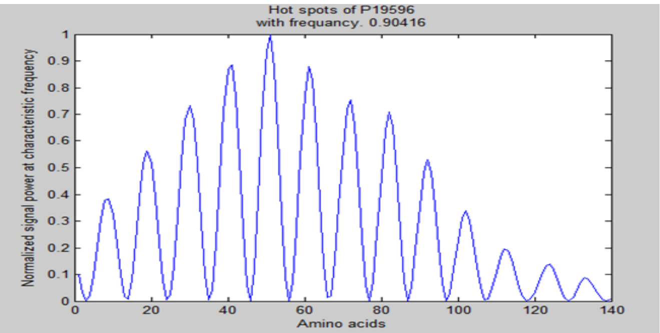

Figure 5: Hot Spot Locations Corresponding to Protein of Chicken

7. Conclusions

This technique is computationally much more efficient than the transform based technique due to low-order IIR digital filter. Effectiveness of this technique is illustrated using several examples of protein sequences. The best accuracy is achieved using this methodology as EIIP transformation is used to obtained binary nucleotide sequence. Fast identification of hot spot from given protein sequence is achieved using Fourier transform.

8. Future scope

More intensive trials can be performing by using a larger and more comprehensive set of the test sequence. The improvement in reliability of the filter based technique is achieved by developing effective strategies which are used to reduce the initial transients that occur during filtering process. The proposed work can be extended for location of hot spots in DNA molecules. One can change filter technique, and vary filter parameters as filter based method is very flexible.

Acknowledgment

It is my pleasure to get this opportunity to thank my beloved and respected Guide Prof. A. S. Shirsat who imparted valuable basic knowledge of Electronics specifically related to Biomedical Signal Processing. We are grateful to Elec. & Comm. Dept. of Smt. Kashibai Navale College of Engineering, Pune for providing us infrastructure facilities and moral support.

References

[1] P. Ramachandran and A. Antoniou, “Filter-Based Methodology for the Location of Hot Spots in Proteins and Exons in DNA,” IEEE Transactions On Biomedical Engineering, Vol. 59, No. 6, pp. 1598-1609, Jun. 2012.
[2] Andreas D. Baxevanis and B.F. Francis Ouellette, “Bioinformatics,” Wiley, 2011.
[3] P. Ramachandran and A. Antoniou, “Identification of hot-spot locations in proteins using digital filters,” IEEE J. Sel. Topics Signal Process. vol. 2, no. 3, pp. 378–389, Jun. 2008.
[4] Cosic, “Macromolecular bioactivity: Is it resonant interaction between macromolecules?—Theory and

Licensed Under Creative Commons Attribution CC BY

applications," IEEE Trans. Biomed. Eng., vol. 41, no. 12, pp. 1101–1114, Dec. 1994.

[5] Christopher J.R. Illingworth, Kevin E. Parkes, Christopher R. Snell,

[6] Philip M. Mullineaux, Christopher A. Reynolds, "Criteria for confirming sequence periodicity identified by Fourier transform analysis: application to GCR2, a candidate plant GPCR?," Biophysical Chemistry 133, 1-3 (2008) 28, version 1-12 Jul 2010.

[7] T. Kortemme and D. Baker, "A simple physical model for binding energy hot spots in protein-protein complexes," Proc. Nat. Acad. Sci., vol. 99, no. 22, pp. 14116–14121, Oct. 2002.

[8] P. Ramachandran and A. Antoniou, "Localization of hot spots in proteins using digital filters," in Proc. IEEE Int. Symp. Signal Processing and Information Technology, Vancouver, BC,, Canada, pp. 926–931, Aug. 2006.

[9] E. Pirogova,Q. Fang, M. Akay, and I. Cosic, "Investigation of the structural and functional relationships of oncogene proteins," Proc. IEEE, vol. 90, no. 12, pp. 1859–1867, Dec. 2002.

[10] Cosic, E. Pirogova, and M. Akay, "Application of the resonant recognition model to analysis of interaction between viral and tumor suppressor proteins," in Proc. 25th Annu. Int. Conf. IEEE EMBS, Cancun, Mexico, pp. 2398–2401, Sep. 17–21, 2003.

[11] Anu Sabarish.R and Tessamma Thomas, "A Frequency Domain Approach to Protein Sequence Similarity Analysis and Functional Classification," Signal & Image Processing: An International Journal (SIPIJ) Vol.2, No.1, pp. 36-49, March 2011.

[12] Protein Data Bank (PDB), Research Collaboratory for Structural Bioinformatics (RCSB). [Online]. Available: http://www.rcsb.org/pdb/
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," Nucl. Acids Res., vol. 28, no. 1, pp. 235–242, 2000.

[13] Swiss-Prot Protein Knowledgebase. Swiss Inst. Bioinformatics (SIB). [Online]. Available: http://us.expasy.org/sprot/

[14] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger,M. J.Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003," Nucl. Acids Res., vol. 31, no. 1, pp. 365–370, 2003.

Author Profile

Ashwini Walekar received the B.E. degree in Electronics and Tele-communication Engineering from Smt. Kashibai Navale College of Engg., Pune, Maharashtra, India. Now, she is pursuing M.E. degree in Signal Processing from Smt. Kashibai Navale College of Engg., Pune, Maharashtra, India.

Licensed Under Creative Commons Attribution CC BY