American Journal of Systems and Software, 2014, Vol. 2, No. 3, 81-84
Available online at http://pubs.sciepub.com/ajss/2/3/5

DOI:10.12691/ajss-2-3-5

A Computational Simulation of Determination of Characteristic Frequency for Identification of Hot Spots in Proteins

Sidhartha Sankar Sahoo, Malaya Kumar Hota*

Department of Electronics and Telecommunication Engineering, Synergy Institute of Engineering & Technology, Dhenkanal 759001, Odisha, India
*Corresponding author: mkhota.mnnit@gmail.com

Received June 09, 2014; Revised June 27, 2014; Accepted July 03, 2014

Abstract Proteins perform their functions by interaction with other molecules known as target. Protein-target interactions are very specific in nature and occur at predefined locations in proteins known as hotspots. For successful protein-target interaction both protein and target must share common spectral component known as characteristic frequency. Characteristic frequency is very importance since it forms basis for protein-target interactions, thus an approach for determination of characteristic frequency in proteins using discrete cosine transform (DCT) is illustrated in this paper. The performance of the proposed method is observed to be better than existing approaches and is illustrated using simulation examples.

Keywords: *proteins, Electron Ion Interaction Potential (EIIP), consensus spectrum, resonant recognition model (RRM), characteristic frequency, Discrete Cosine Transform (DCT)*

Cite This Article: Sidhartha Sankar Sahoo, and Malaya Kumar Hota, "A Computational Simulation of Determination of Characteristic Frequency for Identification of Hot Spots in Proteins." *American Journal of Systems and Software*, vol. 2, no. 3 (2014): 81-84. doi: 10.12691/ajss-2-3-5.

1. Introduction

Proteins are the probably the most important carrier and work force of every living organism. Proteins form the basis for major structural component of animal & human tissue. Proteins are the building blocks of life and are essential for growth of cells and tissue repair. Protein is natural polymer molecule consisting of amino acid unit. All proteins are made up of different combination of 20 compound called amino acids. Depending upon which amino acid link together proteins molecules form enzymes, hormones, muscles, organs and many tissues in the body [1].

Proteins are polymers of amino acid joined together by peptide bond. There are 20 different amino acids that make up essentially all the proteins on earth. An amino acid consists of a carboxylic acid group, an amino group and a variable side chain all attached to central carbon atom. The side chain is the only component that varies from one amino acid to another. Thus the characteristic that distinguish one amino acid from another is its unique side chain that dictates an amino acid chemical property [1]. Even though proteins can be imagined to be linear chain of amino acid, they are not present as linear chains in reality. They fold into complex three dimensional (3-D) structures and it is this folding ability that enables them to perform extreme specific functions. The information necessary to specify the three dimensional (3-D) shape of proteins is contained in its amino acid sequence. The 3-D structure of proteins is most stable form which a protein can attain and this 3-D structure is due to certain specialized regions in proteins known as hot spots [2]. Proteins perform their biological function by interacting with other molecules known as targets and the necessary binding energy for this protein-target interaction is provided by hot spots. Hot spots are small groups of amino acids which provide functional stability to proteins, so that protein can efficiently bind with a target and thus can perform its biological function.

The hot spots in proteins can be identified by the use of Resonant Recognition Model (RRM) [3], which correlates the biological functioning of the protein to the characteristic frequencies. These hot spots in proteins can be localized where the characteristic frequencies of the functional groups are dominant. The signal processing techniques [4] can be used to extract these characteristic frequencies in the protein sequences which are primarily based on the sequence information only. In the earlier reported works [5-9], Discrete Fourier Transform (DFT) and Chirp Z Transform (CZT) have been used to determine the characteristic frequency. In this work, determination of characteristic frequency using Discrete Cosine Transform (DCT) is proposed. The rest of the paper is organized as follows. Section 2 gives brief definition of DCT. Section 3 describes the resonant recognition model. Section 4 gives idea about the amino acids. Step by step procedure for determination of characteristic frequency is described in section 5. Illustrative examples and results using new approach are presented in section 6 and 7 respectively.

2. Discrete Cosine Transform

The Discrete Cosine Transform (DCT) algorithm has been one of the most popular algorithms in domain of digital signal processing. Discrete Cosine Transform is a computational algorithm for numerical evaluation of N samples. The DCT is closely related to the discrete Fourier transform.

DFT is very popular due to its computational efficiency but the strong disadvantages for some application are

It is complex.

It has poor energy compaction.

Since DCT has the very good energy packing property, It means, it contains much information with the less number of coefficients and as it is the real part of DFT, so computational complexity is also less in case of DCT. Because of these two properties, DCT is preferred over DFT. DCT can linearly transform data into the frequency domain, where the data can be represented by a set of coefficients.

DCT is expressed by

$$f(k) = w(k)\sum_{n=1}^{N} x(n)\cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right),$$

$$K = 1, 2, 3 \ldots\ldots N \qquad (1)$$

$$\text{Where} \quad w(k) = \begin{cases} \frac{1}{\sqrt{N}} & k = 1 \\ \sqrt{\frac{2}{N}} & 2 \le k \le N \end{cases}$$

3. Resonant Recognition Model

The RRM is a model which treats the protein sequence as a discrete signal. Certain frequencies in this signal characterize the protein biological function. The RRM was employed to determine the characteristic frequency and to identify amino acids ('hotspot') mostly contribute to the biological function. According to RRM, the hotspots of a particular protein are the amino acids correspond to the region in protein numerical sequence where the characteristics frequency is dominant [3].

For a successful protein target interaction both protein and target must share the same characteristic frequency but with opposite phase. Protein target interaction is highly selectivity and this selectivity depends upon matching of periodicities within the energy distribution of electrons of interacting molecules. Thus a peak in energy of a protein matches a trough in energy of its target and vice versa. The characteristic frequency provides recognition between a protein and its target and hence this model depicts the protein target interaction based on common characteristics frequency named as RRM.

4. Amino Acids

Protein is made up of different combinations of twenty compounds and these compounds are known as amino acids. Proteins perform their binding with other proteins with these amino acids. The protein is not available as a whole rather it is a linear chain of amino acid sequence [1]. The various regions of protein chains interact among themselves and fold into a 3D structure. The amino acid sequence is mapped into numerical sequence i.e. each amino acid is represented by a numerical value which is known as EIIP (Electron Ion Interaction Potential) value [4]. EIIP is a physical property which denotes the average energy of valence electrons in amino acids. The EIIP values for 20 different amino acids are listed in Table 1.

Table 1. EIIP Value for the 20 Amino acids

S. No	Amino Acids	Three letter Symbol	Single letter Symbol	EIIP Value
1	Alanine	Ala	A	0.0373
2	Arginine	Arg	R	0.0959
3	Asparagine	Asn	N	0.0036
4	Aspartic Acid	Asp	D	0.1263
5	Cysteine	Cys	C	0.0829
6	Glutamine	Gln	Q	0.0671
7	Glutamic Acid	Glu	E	0.0058
8	Glycine	Gly	G	0.0050
9	Histidine	His	H	0.0242
10	Isoleucine	Ile	I	0.0000
11	Leucine	Leu	L	0.0000
12	Lysine	Lys	K	0.0371
13	Methionine	Met	M	0.0823
14	Phenylalanine	Phe	F	0.0946
15	Proline	Pro	P	0.0198
16	Serine	Ser	S	0.0829
17	Threonine	Thr	T	0.0941
18	Tryptophan	Trp	W	0.0548
19	Tyrosine	Tyr	Y	0.0516
20	Valin	Val	V	0.0057

Thus each and every amino acid in sequence can be represented by a unique number. Now successfully all digital signal processing tools can be applied to the obtained numerical sequence of amino acid sequence.

5. Determination of Characteristic Frequency

Previous successful attempts have been made for determination of characteristic frequency using DFT [5,6,7,8] and CZT [9]. Here we have proposed a similar approach using DCT and corresponding results are compared with DFT and CZT.

Step by step procedure for determination of characteristic frequency for proteins using DCT is given below.

1. Select minimum no of two proteins from the functional group.

2. Convert protein character sequences into numerical sequences using EIIP values.

3. Determine DCT of numerical sequences obtained in step 2 and evaluate consensus spectrum or cross spectral function by multiplying them.

$$S(k) = |f_1(k)||f_2(k)| \; |f_3(k)| \ldots\ldots |f_K(k)| \qquad (2)$$

Where $f_1(k)$ is DCT of sequence 1, $f_2(k)$ is DCT of sequence 2 and so on. S (k) is cross spectral function of Kth frequencies.

4. If a distinct peak is observed in the consensus spectrum, S (k), observe the corresponding frequency as the characteristic frequency.

5. If the peak in the consensus spectrum is not distinct, increase a protein in steps 1 to 4 until a distinct peak is not available.

6. Illustrative Examples

Functional group of proteins were selected from Swiss-Prot Protein Knowledgebase [10] & Protein Data Bank [11] to demonstrate the performance of the proposed approach and some of the available protein functional group is given in Table 2. Both database are very helpful and reliable and strongly recommended by the biological community. The databases are updated if any existing sequences are altered or if new sequence information becomes available. In our work, the protein sequences have been obtained from these databases.

Table 2. Proteins of functional family used for computation of consensus spectrum

Organism	Protein Name	Swiss-Prot ID	PDB ID
Human	FGF		1fga, 1afc
Human	Glutathione		1aw9, 1axd
Tuna heart	Cytochrome C	P00025, P62894, P99999, P00008	
Human	Human Hemoglobin	P60524, P01958, P02062, P68871, P69905, P68050, P01942, P01946, P68048	
Human	Human Growth hormone	P10912, P16310, P16882, P19941, Q9JI97, Q9TU69, Q02092, O46600	
Bacteria	Barnase	B7M0V1, C4ZK78, C6UF64, C6XRM1, C9NF27, D0KFB0, P00648, P10912	
Bacteria	Barstar	P11540, A7FDT9, A9R1V5, B4TJT7, B5PWB6, C5BAW5, C7MPS8, Q2SZB1, Q62H00	
Anti Bacteria	Lysozome	P04421, P67977, P00698, P61626, P16973	

7. Results and Discussions

To demonstrate the characteristic frequency, we have chosen the following three protein sequences from the online database.

1. Fibroblast growth factor (FGF) of cow family.
2. Cytochrome C from tuna heart.
3. Human Hemoglobin.

For each of the above examples, the characteristic frequency has been determined from consensus spectrum of sufficiently large set of protein sequences belonging to same functional group as shown in figure 1, figure 2 and figure 3 respectively. In each of the consensus spectrum, the peak indicates characteristic frequency. All simulation works in this paper are done using MATLAB.

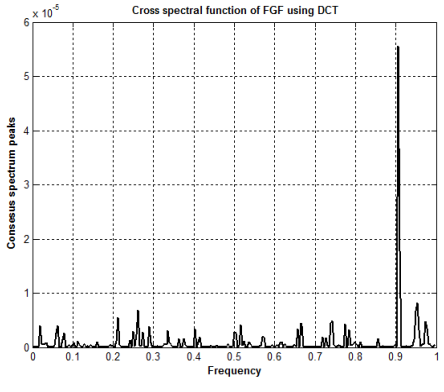

Figure 1. Consensus spectrum for FGF

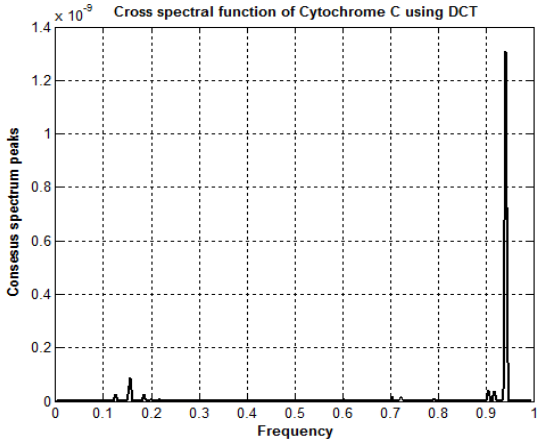

Figure 2. Consensus spectrum for Cytochrome C

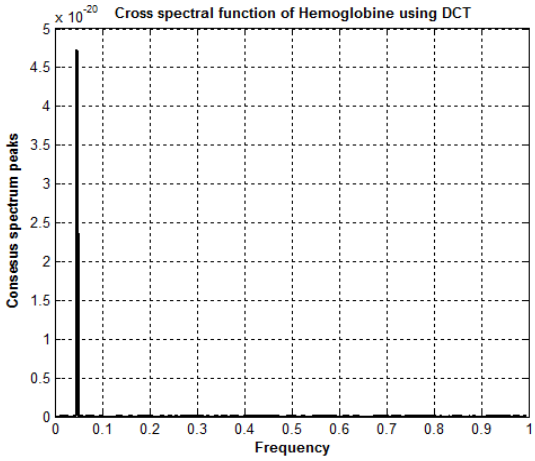

Figure 3. Consensus spectrum for Hemoglobin

The Computational efficiency and Signal to Noise Ratio (SNR) [12] for the proposed method in comparison with the existing methods is analyzed in Table 3 and Table 4 respectively.

Table 3. Comparison of Computational Time

Protein Name	Charact-eristic Frequency	Computational Time in Seconds Over 1000 Iterations		
		DFT	CZT	DCT
Fibroblast Growth Factor (FGF)	0.9063	0.128040	0.799665	0.064133
Cytochrome C	0.9414	0.249861	1.394902	0.124793
Human Hemoglobin	0.0468	0.571974	2.979954	0.292169

Table 4. Comparison of SNR

Protein Name	Charact-eristic Frequency	Signal to Noise Ratio (SNR)		
		DFT	CZT	DCT
Fibroblast Growth Factor (FGF)	0.9063	27.7873	27.7873	58.4480
Cytochrome C	0.9414	86.8571	86.8571	176.5795
Human Hemoglobin	0.0468	126.4115	126.4115	255.0814

We compared the computational efficiency of the proposed work by recording the average CPU times over 1000 runs for each protein sequence. From the result it is found that the computational time in Chirp Z transform is more, DFT is moderate and DCT is less. It is also found that time taken by DCT approach is reduced approximately by 50% compared to DFT approach.

Further, Signal to Noise Ratio (SNR) has been calculated as ratio between signal intensity at the particular peak frequency and the mean value over the whole spectrum. SNR of proposed approach for proteins are compared with existing approaches. DCT approach clearly indicates a considerable improvement in SNR over existing approaches.

8. Conclusion

In this paper, DCT based approach, as an alternative to the DFT or CZT transform method, has been suggested for determination of characteristic frequency. A significant peak exists at characteristic frequency which is obtained from consensus spectrum using a number of proteins sequences from same functional group. Further, there is a considerable improvement in computational time and SNR in DCT approach compared to DFT and CZT Transform. Hence this approach can be very useful for correctly identifying the characteristic frequency which can be useful for hot spots detection.

References

[1] Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter P., "Essential Cell Biology", *Garland Publishing*, New York, 1998.

[2] Bogan, A. A. and Thorn, K. S., "Anatomy of hot spots in protein interfaces", *Journal of Molecular Biology*, 280 (1). 1-9. 1998.

[3] Cosic, I., "Macromolecular bioactivity: is it resonant interaction between macro-molecules? – theory and applications", *IEEE Trans. on Biomedical Engr.*, 41 (12). 1101-1114. Dec. 1994.

[4] Vaidyanathan, P. P. and Yoon, B.J., "The role of signal-processing concepts in genomics and proteomics", *Journal of the Franklin Institute*, 341 (1-2). 111-135. 2004.

[5] Ramachandran, P., Antoniou, A. and Vaidyanathan, P. P., "Identification and location of hot spots in proteins using the short-time discrete Fourier transform", in *Proc. 38th Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA. 1656–1660. Nov. 2004.

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

[7] Sahu, S.S. and Panda, G., "Efficient Localization of Hot Spot in Proteins Using A Novel S-Transform Based Filtering Approach", *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, 8 (5). 1235-1246. 2011.

[8] Kasparek, J., Maderankova, D. and Tkacz, E., "Protein Hotspot Prediction Using S-Transform. *In Information Technologies in Biomedicine", Springer International Publishing.* 3. 327-336. 2014.

[9] Sharma, A. and Singh, R., "Determination of Characteristic Frequency in Proteins using Chirp Z-transform", *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2 (6). June 2013.

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

[11] Protein Data Bank (PDB), Research Collaboratory for Structural Bioinformatics (RCSB). [Online]. Available: http://www.rcsb.org/pdb/.

[12] Yadav, Y. and Wadhwani, S., "Identification of Characteristic frequency in Proteins using Power Spectral Density", *International Journal of Advances in Electronics Engineering*, 1 (1). 342-346. 2011.