American Journal of Computing Research Repository, 2014, Vol. 2, No. 2, 38-43
Available online at http://pubs.sciepub.com/ajcrr/2/2/3

DOI:10.12691/ajcrr-2-2-3

Determination of Characteristic Frequency for identification of Hot spots in Proteins using Computational Simulations: a Review

Sidhartha Sankar Sahoo*, Malaya Kumar Hota

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

Received July 08, 2014; Revised July 16, 2014; Accepted July 20, 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 far various computational simulations have been used for determination of characteristics frequency. In this paper we have applied all computational simulations used till now and also use comparative study based on computational time and Signal to Noise Ratio parameter to stress the best suitable technique. All computational simulation works in this paper are done using MATLAB.

Keywords: *proteins, amino acids, Electron Ion Interaction Potential (EIIP), consensus spectrum, resonant recognition model (RRM), characteristic frequency, Discrete Fourier Transform (DFT), Chirp Z-Transform (CZT), Discrete Cosine Transform (DCT)*

Cite This Article: Sidhartha Sankar Sahoo, and Malaya Kumar Hota, "Determination of Characteristic Frequency for identification of Hot spots in Proteins using Computational Simulations: a Review." *American Journal of Computing Research Repository*, vol. 2, no. 2 (2014): 38-43. doi: 10.12691/ajcrr-2-2-3.

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 computational simulations [4] can be used to extract these characteristic frequencies in the protein sequences which are primarily based on the sequence information only. Previous successful attempts have been made for determination of characteristics frequency using Discrete Fourier Transform (DFT) [5-9], Chirp Z-Transform (CZT) [10] and Discrete Cosine Transform (DCT) [11-12] which are discussed in this paper. The rest of the paper is organized as follows. Section 2 gives brief definition of

computational simulations i.e. DFT, CZT and 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 each approach are presented in section 6 and 7 respectively.

2. Computational Simulations Used

1. Discrete Fourier Transform (DFT)

The DFT is a frequency spectrum analysis which takes the discrete signal in time domain and transforms that signal into its discrete frequency domain. If that sequence is to be represented is of finite duration, the transform used is discrete fourier transform. Thus given a signal that can be represented by a sequence of numbers a spectral characterization of the signal can be obtained, which can also be represented by a sequence of numbers.

Let x(n) be a finite duration signal and its N point DFT is expressed by

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi nk/N} \quad (1)$$

where k= 0,1,2……..N-1 and n is the length of signal.

2. Chirp Z-Transform (CZT)

Chirp Z-transform is one of the popular computational algorithm for numerical evaluation of Z transform of N samples. This algorithm can be used to evaluate the Z transform at M points in z plane. The evaluation is based on the facts that the values of Z transform on circular spiral contour can be expressed as discrete convolution.

$$X_k=\sum_{n=0}^{N-1}\left[x(n).A^{-n}.W^{n^2/2}\right].(W^{k^2/2})W.^{-(k-n)^2/2} \quad (2)$$

where {x(n)}, 0≤ n ≤ N-1 is a given N point sequence, M is arbitrary integer and A and W are arbitrary complex number.

3. Discrete Cosine Transform (DCT)

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. DCT can linearly transform data into the frequency domain, where the data can be represented by a set of coefficients.

DCT is expressed by

$$f(k)=w(k)\sum_{n=1}^{N}x(n)\cos\cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right) \quad (3)$$

$$K=1,\ 2,\ 3\ldots\ldots\ldots N$$

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

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 computational simulations can be applied to the obtained numerical sequence of amino acid sequence.

5. Determination of Characteristic Frequency

To determine the characteristic frequency the first step is to select a protein functional group of interest. In a functional group the number of proteins may vary from case to case. Let we have K number of protein sequence in a functional group. Here the procedure is described using only DCT. The same procedure follows for 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.

$$S(k) = |f_1(k)||f_2(k)| |f_3(k)|.....|f_K(k)| \quad (4)$$

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

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	

Functional group of proteins were selected from Swiss-Prot Protein Knowledgebase [13] & Protein Data Bank [14] 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. For the review protein sequences have been obtained from these databases.

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.
4. Barnase

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.

Fibroblast growth factor (FGF)

The consensus spectrums for FGF using DFT, CZT, and DCT are shown in figure 1, 2 and 3 respectively. In each of the consensus spectrum the distinct peak indicates the characteristic frequency of FGF.

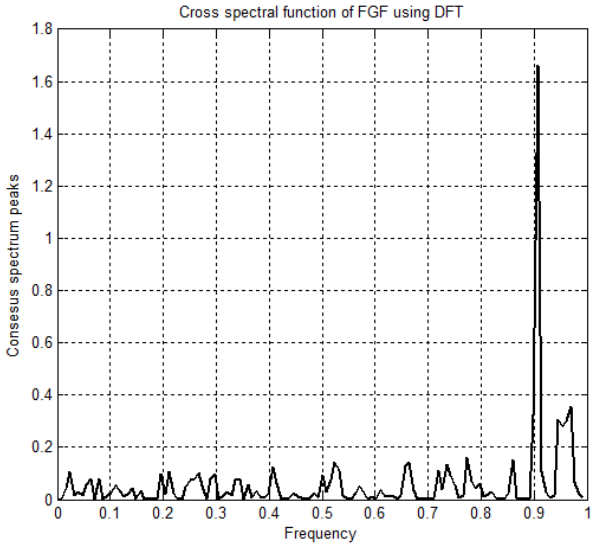

Figure 1. Consensus spectrum for FGF using DFT

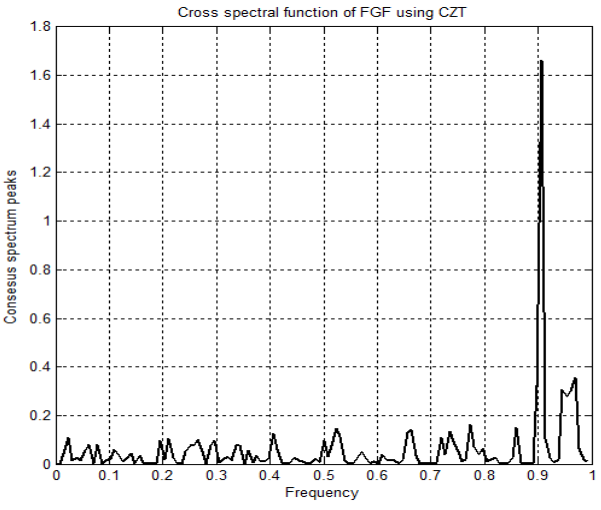

Figure 2. Consensus spectrum for FGF using CZT

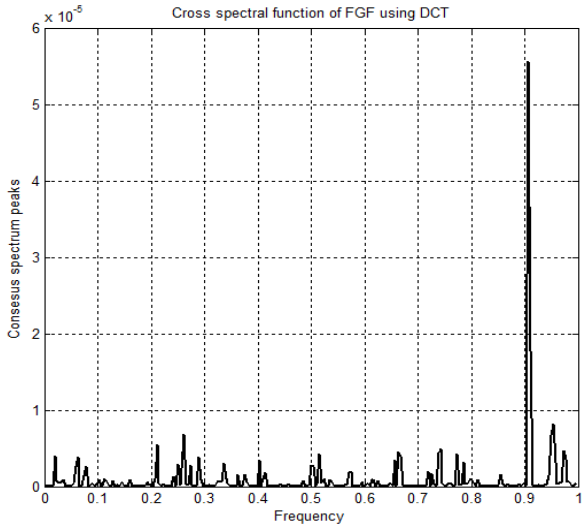

Figure 3. Consensus spectrum for FGF using DCT

Cytochrome C

The consensus spectrums for Cytochrome C using DFT, CZT, and DCT are shown in figure 4, 5 and 6 respectively. In each of the consensus spectrum the distinct peak indicates the characteristic frequency of Cytochrome C.

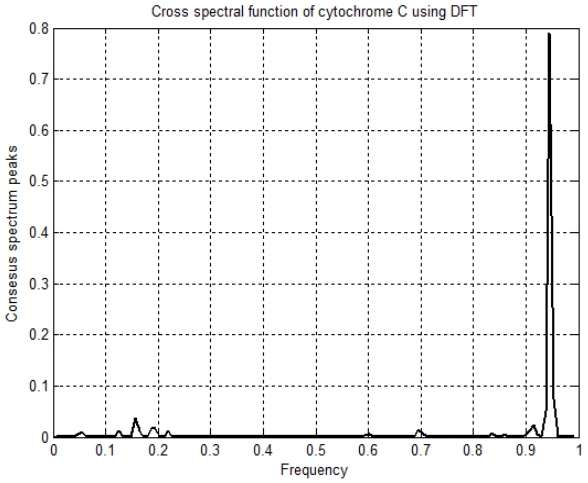

Figure 4. Consensus spectrum for Cytochrome C using DFT

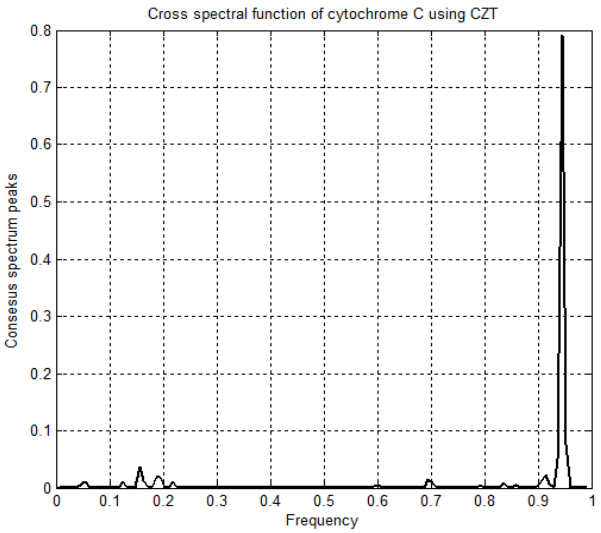

Figure 5. Consensus spectrum for Cytochrome C using CZT

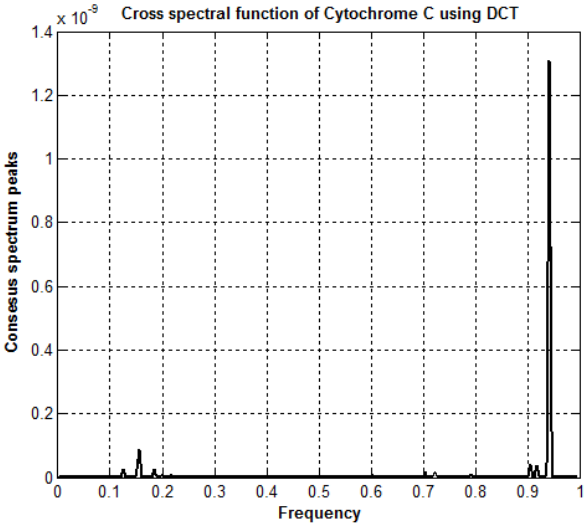

Figure 6. Consensus spectrum for Cytochrome C using DCT

Human Hemoglobin

The consensus spectrums for Hemoglobin using DFT, CZT, and DCT are shown in figure 7, 8 and 9 respectively. In each of the consensus spectrum the distinct peak indicates the characteristic frequency of Hemoglobin.

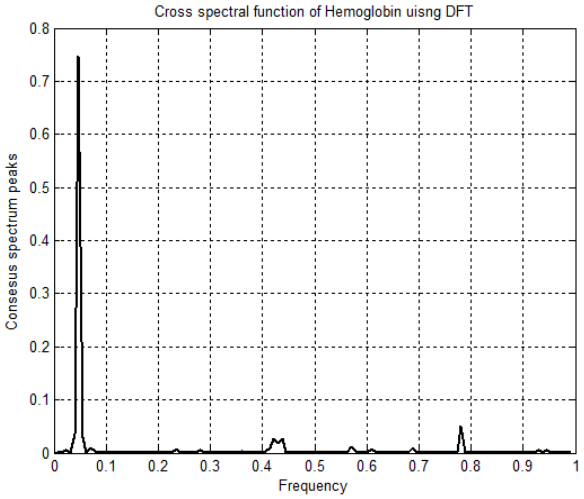

Figure 7. Consensus spectrum for Hemoglobin using DFT

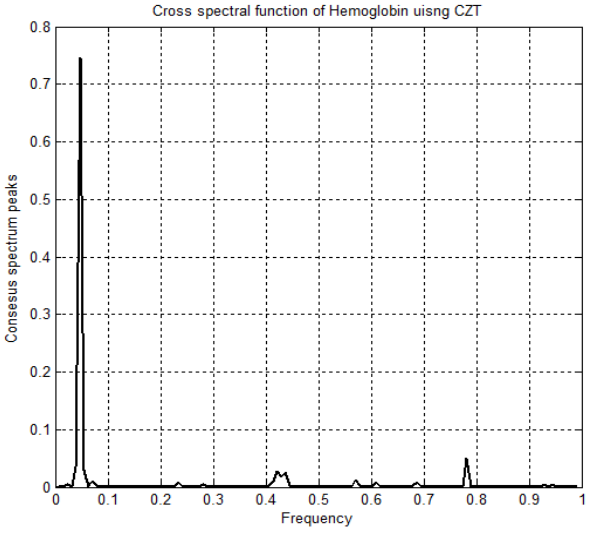

Figure 8. Consensus spectrum for Hemoglobin using CZT

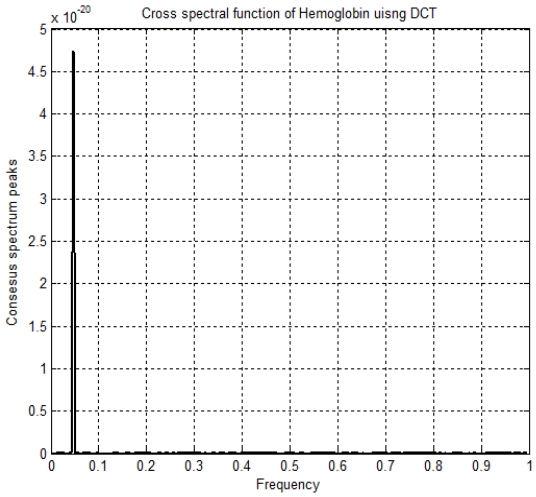

Figure 9. Consensus spectrum for Hemoglobin using DCT

Barnase

The consensus spectrums for Barnase using DFT, CZT, and DCT are shown in figure 10, 11 and 12 respectively. In each of the consensus spectrum the distinct peak indicates the characteristic frequency of Barnase.

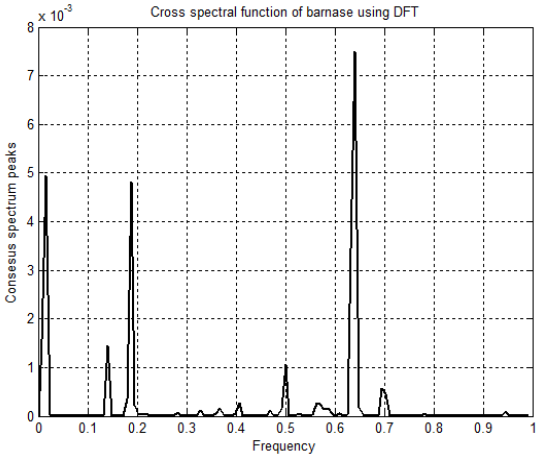

Figure 10. Consensus spectrum for Barnase using DFT

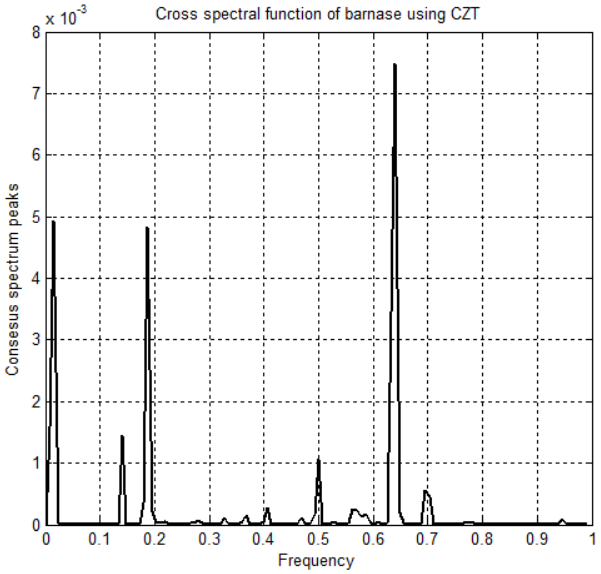

Figure 11. Consensus spectrum for Barnase using CZT

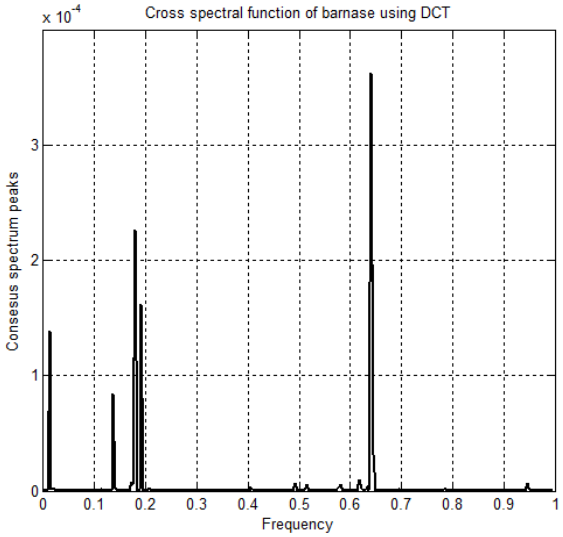

Figure 12. Consensus spectrum for Barnase using DCT

Table 3. Comparison of Computational Time

Protein Name	Characteristic Frequency	No. of sequence used	Sequence length	Computational Time in Seconds Over 1000 Iterations		
				DFT	CZT	DCT
Fibroblast Growth Factor (FGF)	0.9063	2	146	0.1272564	0.7853155	0.0756892
Cytochrome C	0.9414	4	107	0.2564825	1.3562355	0.1356954
Human Hemoglobin	0.0468	9	142	0.5245624	2.9256415	0.2256457
Barnase		8	90	0.4256175	1.8479624	0.1452648

Table 4. Comparison of SNR

Protein Name	Characteristic Frequency	No. of seq. used	Sequence length	Signal to Noise Ratio (SNR)		
				DFT	CZT	DCT
Fibroblast Growth Factor (FGF)	0.9063	2	146	27.787	27.787	58.448
Cytochrome C	0.9414	4	107	86.857	86.857	176.579
Human Hemoglobin	0.0468	9	142	126.411	126.411	255.081
Barnase	0.6406	8	90	32.6139	32.6139	83.5743

For comparative study of all computational simulations, the parameters known as computational efficiency and Signal to Noise Ratio (SNR) are used. The computational efficiency [7] of each simulation work is recorded as the average CPU times over 1000 runs for each protein sequence. SNR for each distinct peak is defined as measure of similarity between sequences analyzed [15]. SNR has been calculated as ratio between signal intensity at particular peak frequency and the mean value over whole spectrum. SNR of at least 20 is considered significant [3]. The Computational efficiency and SNR for DFT, CZT and DCT approaches are compared and analyzed in Table 3 and Table 4 respectively.

The results obtained clearly indicate 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, SNR of DCT is far better than DFT and CZT, also by investigating figure 1 to figure 9 we observed the unwanted peaks other than characteristic frequency in consensus spectrum are suppressed for DCT approach. Thus we can say that DCT approach is far better than DFT and CZT approach for determination of characteristic frequency in proteins.

8. Conclusion

In this paper three different computational simulations are discussed 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. For comparative study computational efficiency and SNR are used and it is observed that there is a considerable improvement in computational time and SNR in DCT approach compared to DFT and CZT. Hence DCT approach is 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] Ramachandran, P. and Antoniou, A., "Identification of Hot-Spot Locations in Proteins Using Digital Filters", *IEEE journal of selected topics in signal processing*, 2 (3). June 2008

[8] 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.

[9] 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.

[10] 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.

[11] Proakis, J.G. and Manolakis, D.G., "Digital Signal Processing: principles, algorithms and applications ", *published by Pearson Education*, Inc., © 2012

[12] Sahoo, S.S. and Hota, M.K., "A Computational Simulation of Determination of Characteristic Frequency for Identification of Hot Spots in Proteins." *American Journal of Systems and Software*, 2 (3). 81-84. 2014

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

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

[15] 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.