IDENTIFICATION AND LOCATION OF HOT SPOTS IN PROTEINS USING THE SHORT-TIME FOURIER TRANSFORM

Parameswaran Ramachandran,[1*] Andreas Antoniou,[1] and P. P. Vaidyanathan[2]

[1]Department of Electrical and Computer Engineering
University of Victoria, BC, Canada V8W 3P6
pramacha@ece.uvic.ca aantoniou@ieee.org

[2]Department of Electrical Engineering
California Institute of Technology, Pasadena, CA
ppvnath@systems.caltech.edu

[*]Point of Contact

Technical Area: D6 - Bioinformatics

Abstract

A new approach for the identification and location of hot spots in proteins based on the short-time Fourier transform is proposed. In the new approach the short-time Fourier transform of the protein numerical sequence is first computed and its columns are then multiplied by the discrete Fourier transform coefficients. By performing this step, the hot spot locations can be clearly identified as distinct peaks, thus overcoming the ambiguities involved in the conventional Fourier transform approach.

1. Introduction

Proteins are the building blocks of all life. They consist of linear chains of subunits called amino acids. There are 20 different types of amino acids, and all proteins are made of combinations of these molecules. The different regions of a linear protein chain interact among themselves and fold into a complex three-dimensional structure. This folding gives a protein molecule its ability to bind very selectively to certain other macromolecules. Proteins perform their function by virtue of this ability. The regions of the protein chains where the selective binding occurs are called *hot spots*.

An approach for the prediction of hot spots described by Cosic [1] involves converting the protein character strings into numerical sequences, computing their discrete Fourier transform (DFT) to determine their unique characteristic frequencies, and changing the amplitudes of the characteristic frequencies in the Fourier spectra to cause a corresponding variation in the time-domain samples of the numerical protein sequence. By comparing the modified sequence with the original, an approximate estimate of the hot spot locations can be obtained. A drawback of the approach has to do with the fact that a change in the amplitude of a single frequency in the Fourier spectrum affects every

sample in the protein sequence. As all the samples change in this way, this method is somewhat unreliable.

2. Identification of Hot Spots in Proteins

It has been shown over the years that the specificity of protein interactions is due to certain periodicities in the distribution of the energies of the free electrons in the protein molecules. One way of representing the energy distribution of free electrons in the amino acids is to use a measure called the electron-ion interaction potential (EIIP) [1] which is a measure of the average energy states of all valence electrons in a particular amino acid. EIIP values can be calculated for each of the 20 amino acids and, consequently, the protein character string can equivalently be represented by a numerical sequence of these values. By computing the DFT of such a sequence, the energy-distribution periodicities can be observed in terms of the frequency components in the Fourier spectrum. It has been pointed out in [1] that the Fourier spectra of protein sequences having similar biological functions have a common frequency component. This component can be identified by taking the product of the amplitude spectra of the protein sequences belonging to a particular functional group. Such a product for a pair of cytochrome C proteins from different organisms is plotted in Figure 1.

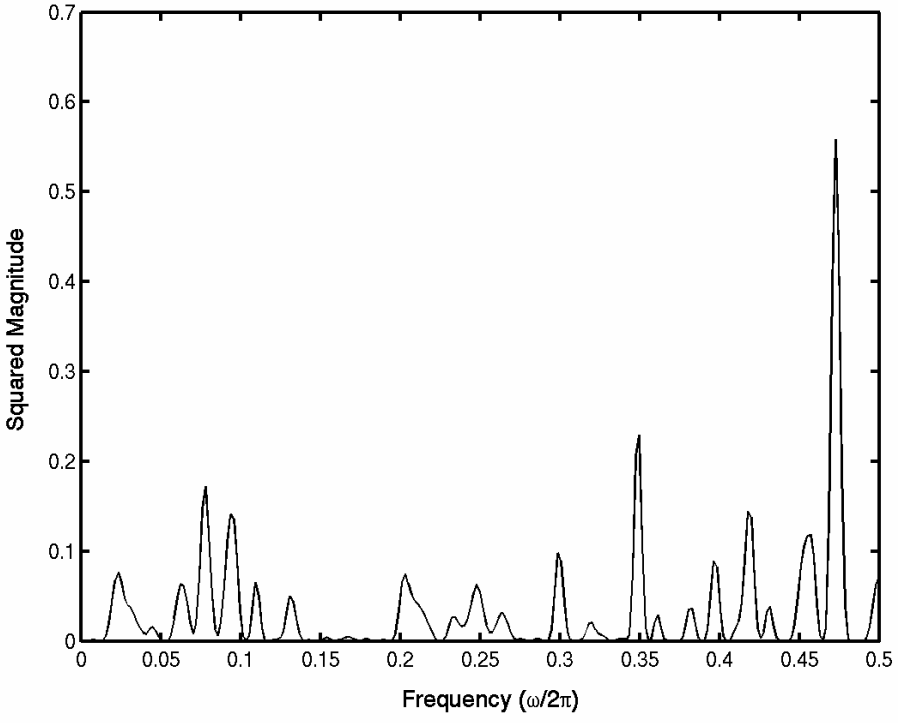

Figure 1: Product DFTs of cytochrome C proteins.

The common frequency component is found to be unique for a specific functional group, and thus, can be used to represent it. Hence, it has been referred to as the *characteristic frequency* associated with the functional group. There are certain regions in a protein sequence, where its characteristic frequency is dominant. These regions play an important role in protein binding, and hence, are referred to as 'hot spots'. Knowledge about the locations of the hot spots of a protein facilitates the understanding of its functionality. Though the Fourier spectrum of a protein sequence reveals its characteristic frequency, it does not give any information regarding the location of hot spots where the characteristic frequency is actually dominant. Thus, in order to efficiently locate the hot spots, we need to resort to time-frequency analysis methods. An attempt in this regard has been made in [2] using wavelets. Wavelets can be used to analyze a signal at different resolutions. They give very good time resolution for high frequencies (such as sudden spikes), but poor time resolution for low frequencies.

Since the characteristic frequency is already known from the Fourier spectrum, a hot-spot identification method does not require very sharp frequency resolution anywhere in the time-frequency plane. Hence, we can afford to compromise on the frequency resolution to a certain extent, in order to gain good time resolution everywhere. Such a compromise is provided by the short-time Fourier transform (STFT) which has uniform time resolution everywhere in the time-frequency plane.

3. New Approach – The Use of the Short-Time Fourier Transform

In this paper, we will describe the use of STFT as a tool in identifying the hot spots associated with the characteristic frequencies. Our approach involves four steps:

1) Conversion of a number of protein character strings of a particular functional group into suitable numerical sequences using EIIP values.
2) Computation of the DFTs of the sequences, followed by their pointwise multiplication to determine the characteristic frequency.
3) Computation of the STFT of the protein sequence in question, using a suitable window.
4) Multiplication of each column of the STFT by the DFT product obtained in step 2.

Step 4 above is crucial, and is an important element of novelty in the proposed method, as explained next.

The STFT gives the frequency content of a signal over short intervals of time. So, if the frequency content varies with time, it can be pictured effectively using the STFT. By simply taking the STFT of a protein sequence (i.e., by just performing step 3 above), we do get an estimate as to where in the protein sequence the characteristic frequency is dominant. However, this is accompanied by disturbingly large amplitudes of many other unwanted frequencies. By performing step 4 as above, we find that the unwanted frequencies vanish, thus highlighting only the characteristic frequency. In this way, it becomes relatively easy to identify the hot spots as distinct peaks.

To demonstrate the power of the proposed approach, a three-dimensional plot of the STFT of cytochrome C protein of tuna heart is shown in Figure 2. The distinct peaks at different regions of the plot give a very clear idea about the location of the hot spots.

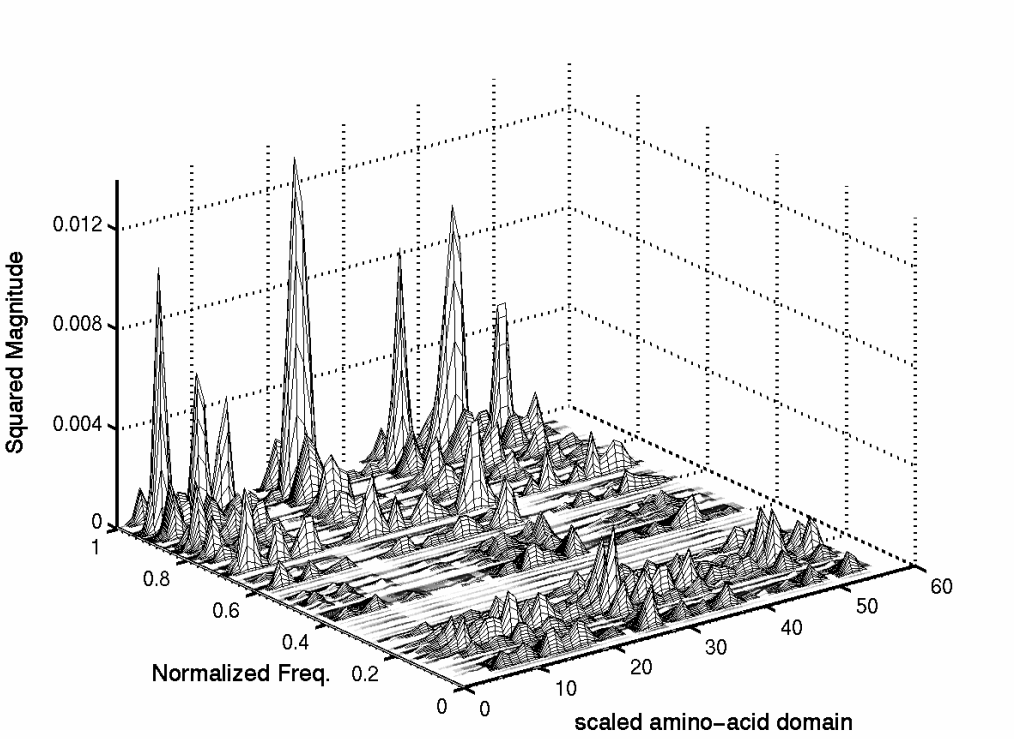

Figure 2: STFT of tuna heart cytochrome C protein, showing peaks as the hot spots.

4. Conclusions

The hot spot regions identified by our approach were found to match well with other available published data. Tests using a number of other protein sequences are now carried out to verify the usefulness of the method, and the results will be detailed in the final version of the paper.

References

[1] I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules? – theory and applications," *IEEE Trans. on Biomedical Engr.*, vol. 41, no. 12, pp. 1101-1114, Dec. 1994.

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