5-1-2002

Analysis of Protein Sequences Using Time Frequency and Kolmogorov-Smirnov Methods

Kobby Essien
Dartmouth College

Dartmouth College Computer Science Technical Report TR2002-431

Analysis of Protein Sequences Using Time Frequency and Kolmogorov-Smirnov Methods

Kobby Essien
Kobby.Essien@alum.dartmouth.org
Student Honors Thesis
Advisor: Metin Akay
Dartmouth College Computer Science

Abstract

The plethora of genomic data currently available has resulted in a search for new algorithms and analysis techniques to interpret genomic data. In this two-fold study we explore techniques for locating critical amino acid residues in protein sequences and for estimating the similarity between proteins. We demonstrate the use of the Short-Time Fourier Transform and the Continuous Wavelet Transform together with amino acid hydrophobicity in locating important amino acid domains in proteins and also show that the Kolmogorov-Smirnov statistic can be used as a metric of protein similarity.

I. Detection of Active Sites in Protein Sequences

A. Introduction

The Human Genome Project is releasing a wealth of information about the genetic code of humans however much of the information is yet to be interpreted. The code contains instructions that make up the three-dimensional proteins that constitute and maintain the human body. The three-dimensional structure of a protein is important because protein structure is linked protein function. In order to effect a function on a cell,

it is not uncommon for a protein to have to dock a specially shaped section of its three-dimensional structure into a specifically shaped receptor on a target cell. In short, many proteins have a site on them that initiate, mediate or terminate a particular biological action.

The genetic code is a sequence and as sequences are related to series one might wonder if the genetic code could be made amenable to time series analysis techniques. White et al and [13] used signal-processing techniques to predict the protein tertiary structure from amino acid sequences. Peng et al. [14] used random walks to study correlations in nucleotide sequences. Irena Cosic has taken signal processing even further and established the Resonant Recognition Model [RRM] which is a physicomathematical model for protein analysis.

The basis of the RRM is Cosic's discovery that there exists a significant correlation between the spectra of numerical representations of amino acids and their biological activity [15]. More specifically, the biological function of a protein is characterized by certain frequencies of its signal representation [16].

RRM analysis first involves converting the amino acids that constitute a protein into a "discrete time series." The position of an amino acid in the sequence can be thought of as the time. The datum associated with each time in our study is hydrophobicity which is a measure of an amino acid's tendency to avoid water. After the conversion of the amino acid sequence is made into the protein time (space) series signal (which we call a "protein signal") the signal is analyzed to locate the dominant frequencies. It has been shown that a particular function in a protein is represented by one RRM characteristic frequency that can be determined by Fourier analysis [15].

In this study we take advantage of the RRM framework and use the Continuous Wavelet Transform (CWT) and the Short-Time Fourier Transform (STFT) to perform time-frequency analyses of proteins. These two transforms have the added advantage of presenting information about space (time) that in our case is associated with corresponds to a particular amino acids location in a protein and we are consequently able to identify the active amino acids contributing to the characteristic frequencies of the proteins. With this information, we are able to determine some of the most functionally important amino acids in the protein [34].

B. Genetics Background

1. From DNA to Gene

The human genetic code is found in DNA (deoxyribonucleic acid) which is organized into cellular structures called chromosomes. DNA is a double helix of nucleic acids (see Figure 1). Each nucleic acid consists of units called nucleotides each of which is composed of a nitrogenous base, a 5-carbon sugar and a phosphate group. There are four bases present in DNA – adenine (A), cytosine(C), thymine (T) and guanine (G).

The deoxyribose sugar and the phosphate group of each nucleotide form the sugar-phosphate backbone of each strand of DNA. Within the backbone, the nucleotides are joined by covalent bonds between the phosphate group of one nucleotide and the deoxyribose sugar of another nucleotide.

Across the helix, nucleotides are linked by hydrogen bonds between the nitrogenous bases. In DNA, hydrogen bonding of bases produces only two combinations of amino acid pairs across the helix. Adenine must always be paired with thymine.

Cytosine must always be paired with guanine. This specificity of the amino acid pairing is termed complementary base pairing.

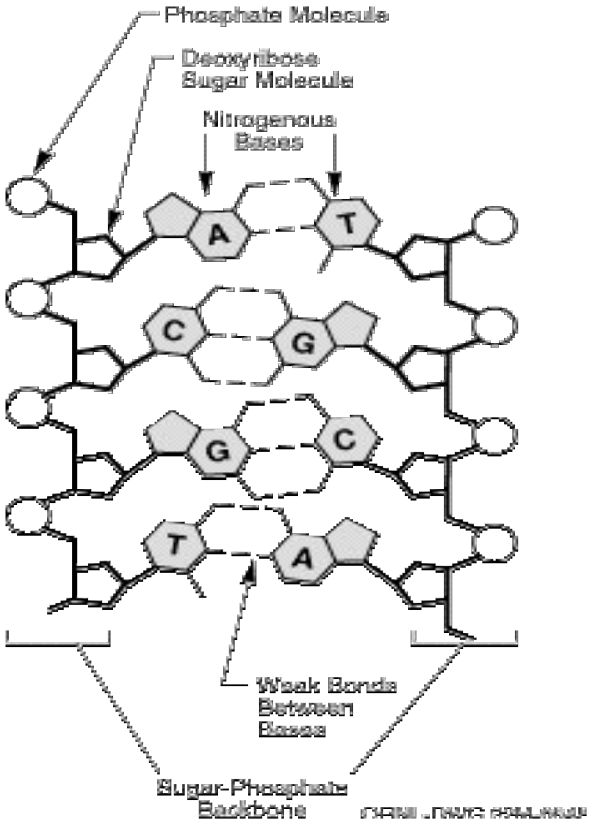

Figure 1 - The Structure of DNA [1]

The bases and the order in which they appear in a strand of DNA are of particular interest as they comprise the DNA sequence of a human. Within the DNA sequence are genes, which are specific sequences of bases that specify the information necessary for creating a particular protein. The human DNA sequence contains between 30000 – 40000 genes [3]. Despite the high number of genes, only about 10% of genome consists of meaningful code or is expressed as proteins [3]. These expressed regions are known as

exons. The remaining 90% of the intervening regions of the genome (introns) are not useless but are believed to have regulatory roles in cells [4].

2. From Gene to Protein

Ribonucleic acid (RNA) is another important nucleic acid found in living organisms. It is similar to deoxyribonucleic acid except that instead of a deoxyribose sugar, RNA contains a ribose sugar. There is also a difference in nitrogenous bases between RNA and DNA. As mentioned earlier, DNA has the bases adenine, thymine, guanine and cytosine. In DNA, adenine and thymine pair together in complementary fashion and so do cytosine and guanine. RNA on the other hand has the bases adenine (A), uracil (U), guanine (G) and cytosine (C). Uracil and adenine compliment each other as do cytosine and guanine.

There are 3 types of RNA involved in protein synthesis – messenger RNA (mRNA) which is the template of constituents of a protein; ribosomal RNA (rRNA) which makes up the ribosome, the site of protein synthesis; transfer RNA (tRNA) which brings amino acids to the ribosome

Transcription is the first major part of protein synthesis. During transcription, a molecule called RNA polymerase parts the DNA double helix in regions with genes. RNA polymerase then moves along one particular strand of the helix, known as the coding strand, and creates a strand of messenger RNA (mRNA) in complementary base pair fashion (an A in the DNA coding strand results in a U in mRNA; a T results in an A in mRNA; a C results in a G in mRNA; a G results in a C in mRNA). See **Figure 2** for a pictorial view of this process. As RNA polymerase moves along the coding strand, the

transcribed portions of the coding strand become associated once more with their complementary strand of DNA.

After transcription, mRNA leaves the nucleus of the cell and begins sliding through an organelle known as the ribosome. Each triplet of bases on the strand of mRNA is called a codon as it codes for a particular amino acid. The second major step of protein synthesis is called translation and during this step, codons are translated into amino acids, which are the building blocks of amino acids. **Figure 3** shows the generic structure of an amino acid. There are 20 amino acids that make up proteins. Structurally each amino acid consists of a centrally located alpha carbon, an amino end, a carboxyl end and a side chain. The side chain (also called the R-group) is the variable part of the amino acid that distinguishes one amino acid from another.

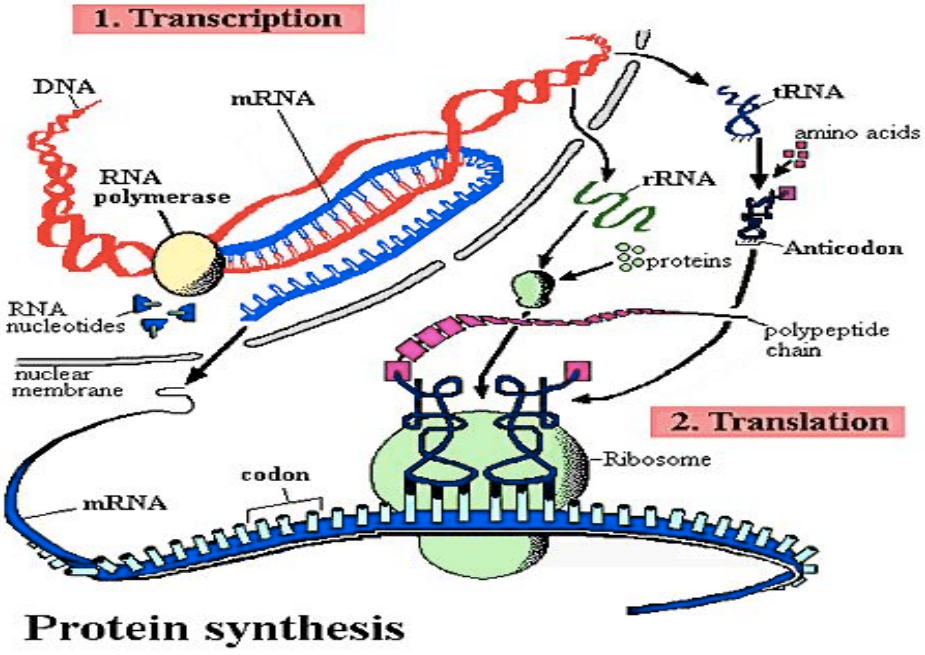

Figure 2 [5]

Transfer RNA (tRNA) transports specific amino acids from the cytoplasm to the ribosome. Each tRNA molecule has an anticodon that is complementary to a particular codon in mRNA. As the mRNA slides through the ribosome, a molecule of tRNA docks into the ribosome with an anticodon complementary to the mRNA's codon. Once this happens the amino acid attached to the tRNA molecule is appended to the string of amino acids that are needed to synthesize the protein specified by the strand of mRNA. Such a string of amino acids is called a polypeptide chain and proteins consist of polypeptides in a particular conformation.

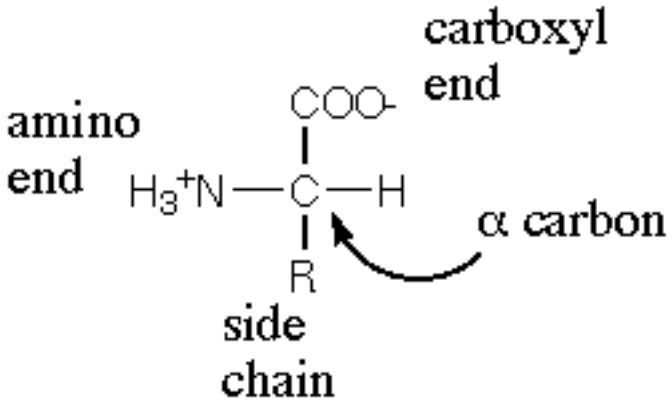

Figure 3 – A generic Amino acid [9]

Proteins like many other things in nature are three-dimensional. Owing to the structural complexity of proteins their structure is studied at four different levels known as primary, secondary, tertiary and quaternary structures.

The primary structure of a protein is its linear arrangement of amino acids. In many proteins there are repeating structural motifs resulting from hydrogen bonding between amino acids in a polypeptide chain. These motifs form the secondary structure of the protein. Two common motifs are alpha helices and beta pleated sheets. The level of organization right above the secondary structure is tertiary structure. Tertiary structure

is the consequence of proteins contorting themselves to form a more compact three-dimensional structure. Quaternary structure on the other hand refers to the structure of the protein formed when two or more proteins with tertiary structure, called subunits, come together to form a protein.

3. Determining the Three-Dimensional Structure of Proteins

Two major techniques for determining the three-dimensional structure of proteins are X-ray crystallography and Nuclear Magnetic Resonance Spectroscopy [7].

In x-ray crystallography, the protein to be studied must first be crystallized. X-rays are then passed through the resulting crystal. The x-rays diffract and form an image known as a diffraction pattern on photographic film or on a radiation counter [7]. From the diffraction pattern, a map showing the electron density of the protein can be constructed. Through a combination of an electron density map and protein primary structure information, the location of each atom in the protein can be determined [4]. Unfortunately, this technique only works for proteins that can be crystallized.

In solution magnetic resonance, an aqueous solution of the protein is made. The spins of the protons in the protein are then aligned by placing the solution in a magnetic field. The application of radiofrequency pulses excites the protons and results in them emitting signals with frequencies depending on the molecular environment of the protons [7]. The signals emitted can be used to come up with estimates of the distances between protons in the protein being studied. Using the distances and information such as bond angles, three-dimensional ensembles of models of the protein can be computed [7,8]. A collection of possible structures is obtained as opposed to a one unique structure.

C. Materials

1. Hemoglobin

Respiration in living cells requires oxygen. Oxygen enters the human body through the pulmonary system and has to be carried to cells all over the body. Hemoglobin is the protein found in red blood that carries oxygen to most of the cells the body.

In human adults hemoglobin is a protein with a quaternary structure composed of two sets of two identical subunits. The two alpha subunits are each made up of 141 amino acids and the beta subunits are made up of 146 subunits each. The most important part of each subunit is its heme group. The heme group is a cofactor (a non-protein compound required for the proper functioning of certain proteins) with a central iron atom. An oxygen molecule binds reversibly to each subunit via its heme group; consequently, each molecule of hemoglobin can carry four molecules of oxygen. Upon reaching its destination, hemoglobin unloads the oxygen molecules. In Figure 4 the heme group is the dark structure within the polypeptide chains

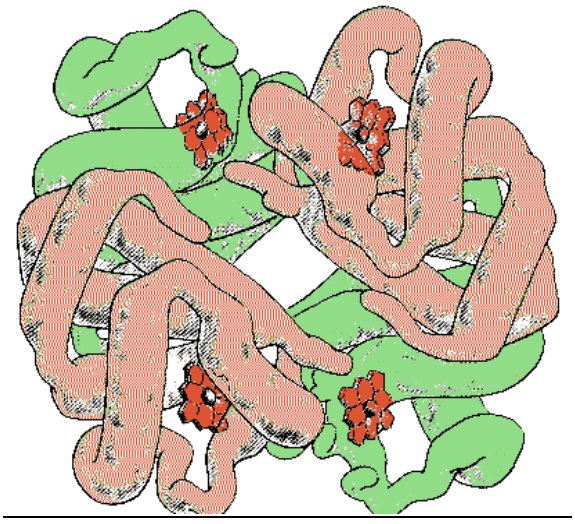

Figure 4- Hemoglobin [9]

2. Myoglobin

Myoglobin is similar in function to hemoglobin except that its function is to store oxygen in muscle cells. Diving animals like seals and whales often spend extended periods of time underwater and as such are not able to take in oxygen through their nostrils and into their pulmonary systems as they normally do on land. In times like these diving animals rely heavily on myoglobin for oxygen. Myoglobin is also found in the skeletal and cardiac muscle of non-diving animals [10]. Human myoglobin consists of one peptide chain consisting of 154 amino acids. It only has one heme group (see Figure 6) an consequently can only store one molecule of oxygen

Figure 6-Myoglobin [10]

3. Prolactin

Prolactin is a hormone produced in the anterior pituitary gland and has a diversity of effects in humans and in other animals. In humans, its primary role is the stimulation

of milk production in the mammary gland. Structurally, human prolactin is a polypeptide consisting of 227 amino acids. The most important site on most hormones is the binding site. Hormones like prolacatin attach themselves via their binding sites to a receptor on their target cells and effect their action.

4. Glucagon

Glucagon is a hormone produced by the alpha cells of the Islets of Langerhans in the pancreas. Glucose is the body's main energy source and the body stores excess glucose as a compound called glycogen in the liver and as triglycerides in fat cells. When blood glucose levels are low, the alpha cells secrete glucagon which causes the liver to break down glycogen into glucose and fat cells to break triglycerides into glucose which the body can burn to produce energy.

5. Cytochrome C

Energy production in animal cells occurs in a cellular structure called the mitochondrion. In most animal cells, the currency of energy is a molecule known as adenosine triphosphate (ATP) which is the end product of the energy-creating process of metabolism. Part of the metabolic process involves the transfer of high-energy electrons between certain proteins in the inner mitochondrial membrane. Cytochrome c carries out part of the electron transfer by shuttling electrons from a complex known as the cytochrome b-c1 complex to another complex known the cytochrome oxidase complex.

The lone heme group within the cytochrome c is the carrier of the electron.

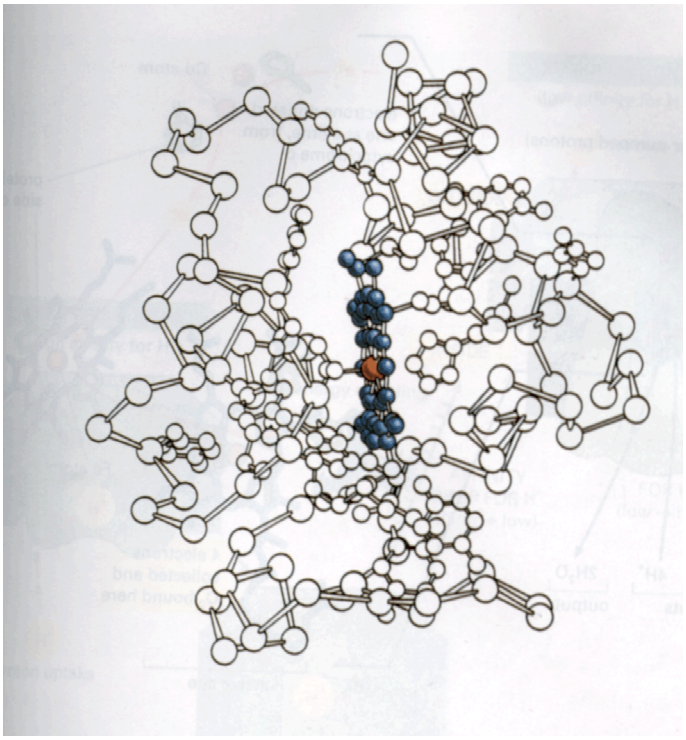

Figure 7-Cytochome C [12]

6. Lysozyme

Enzymes are substances that speed up reactions without taking part in the reactions themselves. Lysozyme is an enzyme whose primary role in humans is protection against bacteria. It effects this action by making possible a hydrolysis reaction in which a water molecule is use to break polysaccharide molecules in bacterial cell walls. The most important site of the lysozyme molecule is the active site in which parts of bacterial cell walls lodge before they are hydrolyzed. In humans, lysozyme is found in several secretions including tears where it aids in keeping the eyes bacteria free and mucus for protecting the inner lining of the nose. We examine rat and human lysozyme in this study.

6. Epidermal Growth Factor

Epidermal growth factor is protein that unlike most proteins is not confined to a particular site of synthesis or a specific target cell type. Its varied effects include accelerating incisor eruption and eye-lid opening in new-born mice [25], the proliferation of epidermal cells and the inhibition of gastric juice secretion. [11].

D. Methods

To obtain our series, the proteins were converted into a "time series" of the hydrophobicities of consecutive of amino acids in the protein. In the application of signal processing techniques to the sequences, the sampling rate can be assumed to be 1 since the distance between amino acids is about 3.8Å.

The Fourier Transform expands a protein signal into the frequencies that characterize it. Given a discrete signal x(nT), with N samples, its Fast Fourier Transform (FFT) is

$$X(w) = \sum_{n=0}^{N-1} x(n) e^{\frac{-jwn}{N}} \qquad (1)$$

However, the Fast Fourier Transform gives no time information so nothing is known about which parts of the signal contribute to those frequencies. Consequently, if there are two signals composed of similar frequencies but it is the case that these frequencies occur at very different locations in time, their Fourier Transforms may appear highly alike. Also the Fourier Transform assumes that the signals being analyzed are stationary.

One way to alleviate these problems is to use the Short-Time Fourier Transform (STFT) also known as the Windowed Fourier Transform. With the STFT, windows of

different sections of the signal are analyzed with the Fourier Transform. As the windows have specific sizes so time information is not lost. We chose a window length of 12 points as this is close to the average length of an alpha helix, one of the common secondary structure conformations in proteins. We had and overlap of 11 points between windows. Overlap is important as it dictates the amount of frequency information lost due to the splitting of the signal into windows. The STFT is defined as

$$X_{STFT}(t, f) = \int x(t)h^{*}(t-\tau)\exp(-j2\pi f\tau)d\tau \qquad (2)\ [33]$$

where h(t) is a windowing function in the time domain and * represents complex conjugation.

The Wavelet Transform is a powerful signal processing technique that has much facility in multi-resolution analysis and feature extraction from signals [16]. It has been utilized in applications as diverse as fingerprint compression, image recognition [31] and the analysis of heart rate variability data [32]. The Continuous Wavelet Transform is defined as

$$\mathrm{CWT(a, b)} = \int x(\tau)\psi_{a,b}^{*}(\tau)d\tau \qquad (3)$$

where * signifies complex conjugation, a is a scaling factor, b is the time. $\psi_{a,b}(t)$ is obtained scaling a function $\psi(t)$ by time b and scale a

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right) \qquad (4)\ [33]$$

$\psi_{a,b}(t)$ is one of several wavelet functions. In this study we use the Morlet wavelet which is defined as

$$\psi(t) = C\exp\left(-\frac{t^2}{2} + jw_0 t\right) \qquad (5)\ [16]$$

The Continuous Wavelet Transform provides a time-frequency representation of a signal. This representation captures the evolution of frequencies within a signal over time. The Fourier Transform, which is one of the main transforms utilized in signal processing, is unable to capture time information and only presents frequency information. In this study, our time series consist of hydrophobicities of consecutive amino acids in a protein. Consequently, the "time" in our time series actually represents the location of an amino acid within a protein sequence. In order to obtain the time-frequency version of the Continuous Wavelet Transform, one has to make the substitution $a_0 = f_0/f$ in (1) [16]. The time frequency representation of the CWT is sometimes referred to as the scalogram.

In applying the Continuous Wavelet Transform to the "hot spot" identification problem one has to look for the local energy maxima in the space-frequency representation of the protein signal [16] as the amino acids being searched for are those that contribute the most frequency-wise.

4. RESULTS

1. Identification of Critical Parts of Proteins ("hotspots").

a. Human Hemoglobin.

The oxygen-carrying heme prosthetic group has an iron atom that is critical to its function. This atom is attached to the 87^{th} residue in the alpha chain. Lehmann and Huntsman [18] also point out that there exists a gap between the iron atom in the heme group and the alpha chain's 58^{th} residue. This gap accommodates an oxygen molecule [18]. Several important residues that have been implicated in the functioning of

hemoglobin are located between the 40th and 90th residues. The 64th, 70th, 79th and 83rd amino acid residues are of particular importance [35].

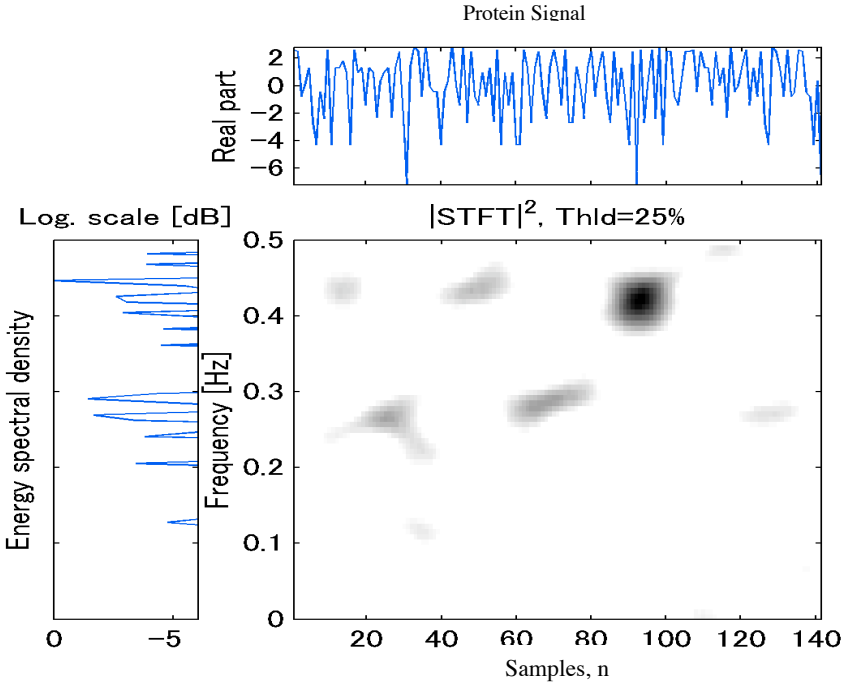

Figure 8 - STFT of human hemoglobin alpha chain

An examination of the Short-Time Fourier Transform of the alpha chain of hemoglobin (Fig. 8) reveals that most of the high-energy areas are located between the 40th and 90th residues, with the most significant region occurring in the vicinity of the 87th amino acid residue. There is also a hot spot at the 63rd amino acid. Residues 18, 22, 36, 43 and 59 are also within hot spot regions. These amino acids are significant as Cosic et. al [26] report that even though they are not part of the active site they "group in a dome-like fashion over the heme group."

The scalogram of the alpha chain of hemoglobin (Fig. 9) almost mirrors the STFT of the protein, the only difference being the reduced intensity of the high-energy regions below the 95th amino acid residue.

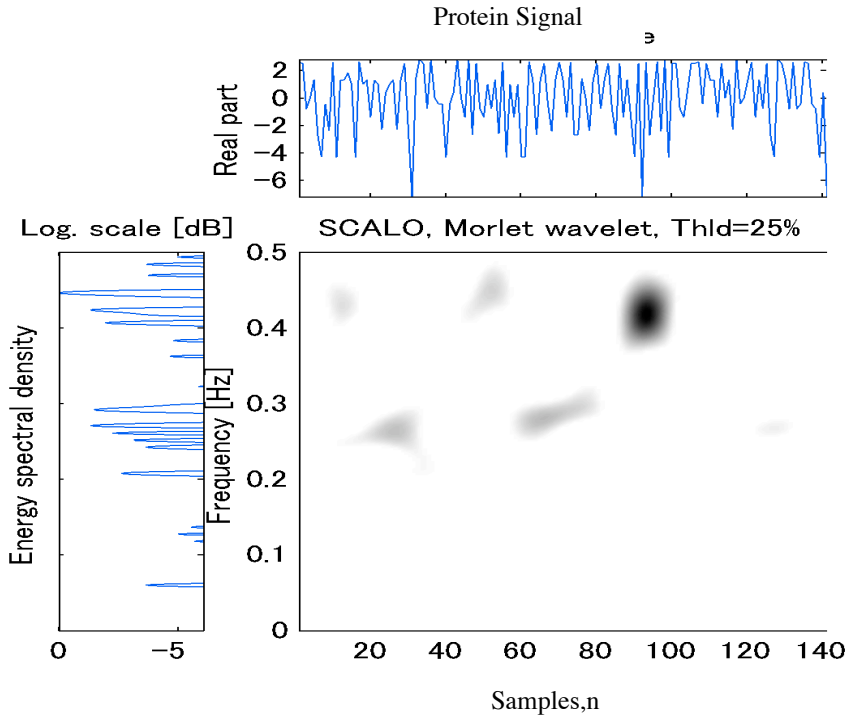

Figure 9 - Scalogram of hemoglobin's alpha chain

In the beta chain, the heme group is linked and the 92nd residue and the oxygen-accomodating-gap is located at the 63rd residue [18]. Most of the important residues are found between the 35th to the 90th residues. The 94th and 100th residues are also thought to be important [35].

As expected, the STFT of the beta chain of hemoglobin (Fig. 10) shows that most of the high-energy regions occur between the 35th and 90th amino acids. There are two regions of high energy at the 63rd residues and a region of extremely high energy at the 92nd residue capturing both the oxygen-accomodating space and the heme-linked residue. Cosic et al [26] note that residues 42, 45, 64 and 69 form a dome over the heme group and all these residues fall within the high-energy regions. There were regions of unaccounted for energy below the 35th residue and above the 92nd residue which warrant experimental investigation.

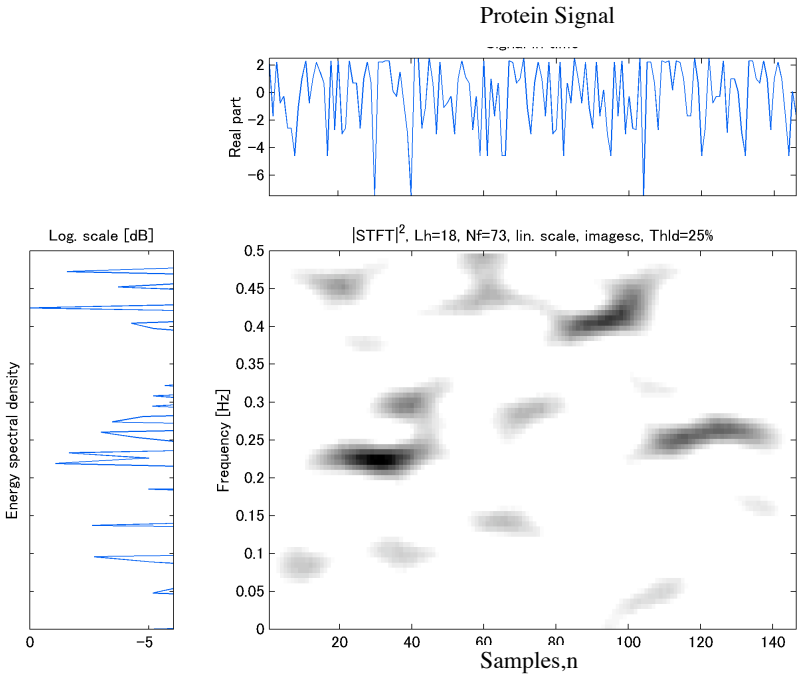

Figure 10 - STFT of the beta chain of hemoglobin

The Continuous Wavelet Transform of the beta chain (Fig. 11) presents basically the same information as the Short-Time Fourier Transform. It too has regions of high energy whose role in the function of hemoglobin was not found in the literature.

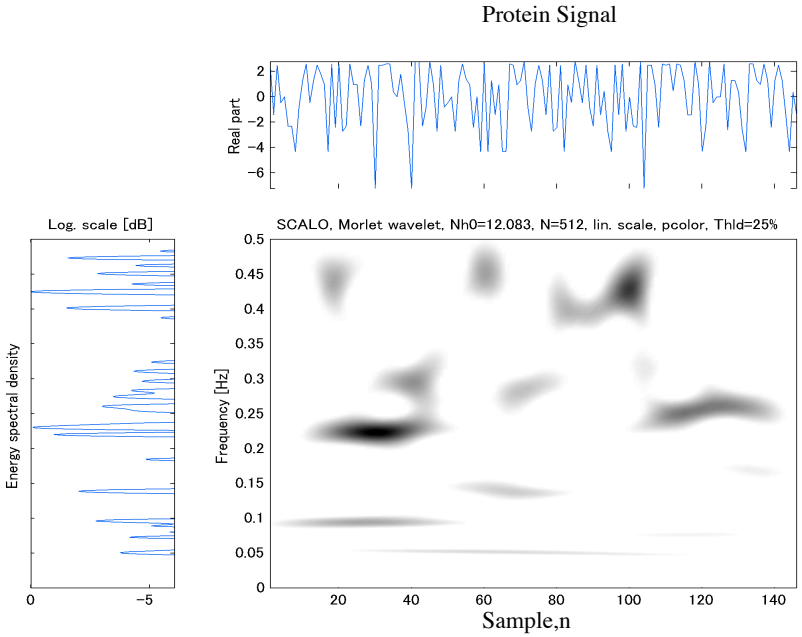

Figure 11. Scalogram of the beta chain of hemoglobin

b. Myoglobin

Residues 64 and 93 together with the heme group form myoglobin's active site [26]. Other important residues and residue ranges are the 138th amino acid and 29th to 45th, 64th to 72nd and 89 to 107th residue ranges [18].

The regions referred to earlier are covered by the high-energy regions in both the Short-Time Fourier Transform (Figure 12) and the scalogram (Figure 13). There are high-energy regions falling outside the aforementioned ranges. Cosic et al [26] point out that the 25th, 33rd, 46th and 65th residues, which are all captured by the STFT and the CWT, are part of the dome-shaped region covering the heme group.

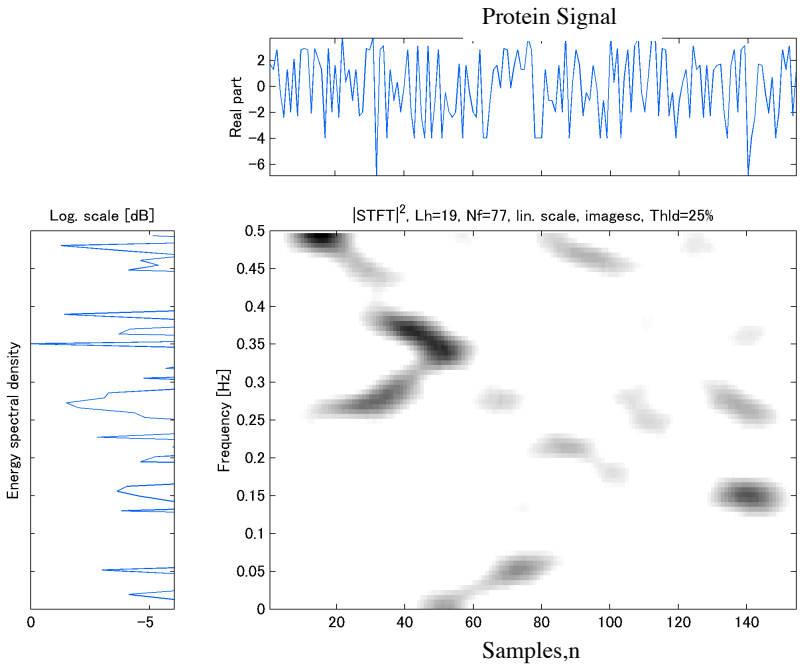

Figure 12 - STFT of myoglobin

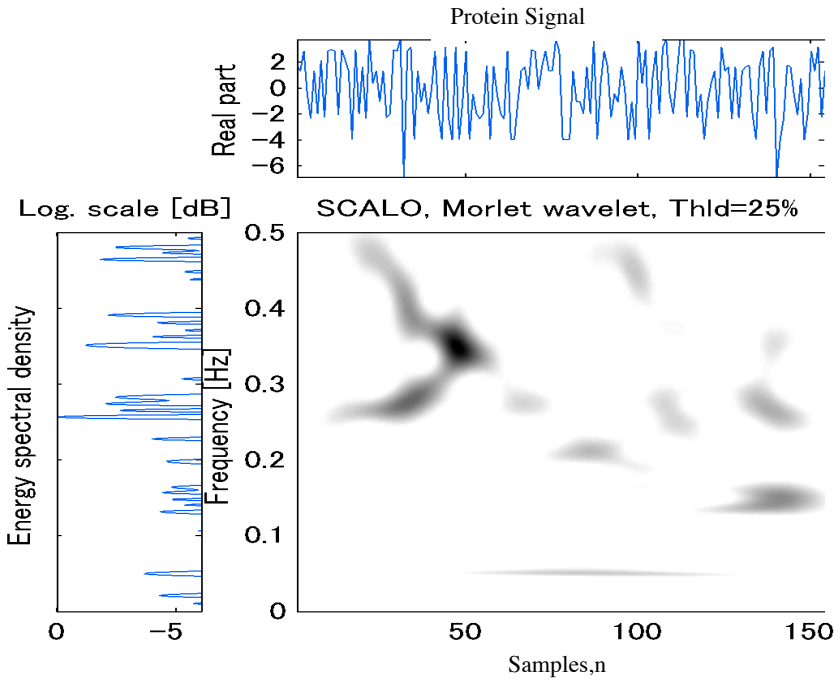

Figure 13 - Scalogram of Myoglobin

c. Cytochrome C

The lone heme group of cytochrome C is bonded to the cysteine residues at positions 14 and 17. Tzagaloff et. al [19] report that amino acids important in the association of cytochrome c to its heme group can be found in the ranges 10-17, 30-35, 64-71,82-85 and 94-98.

Both the STFT (Figure14) and CWT (Figure 15) capture the “hotspots” of cytochrome c. However, the intensity of the “hot spots” produced by the STFT is significantly less than that of those produced by the CWT. The scalogram appears to have a lot of excess high-energy spots whose possible role in the function of cytochrome c merites experimental exploration

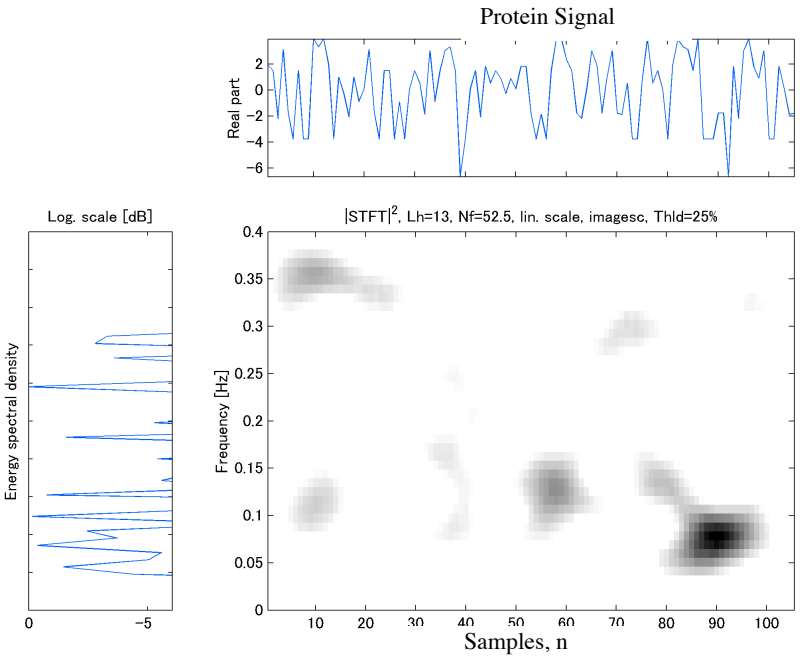

Figure 14 - STFT of human cytochrome c

.

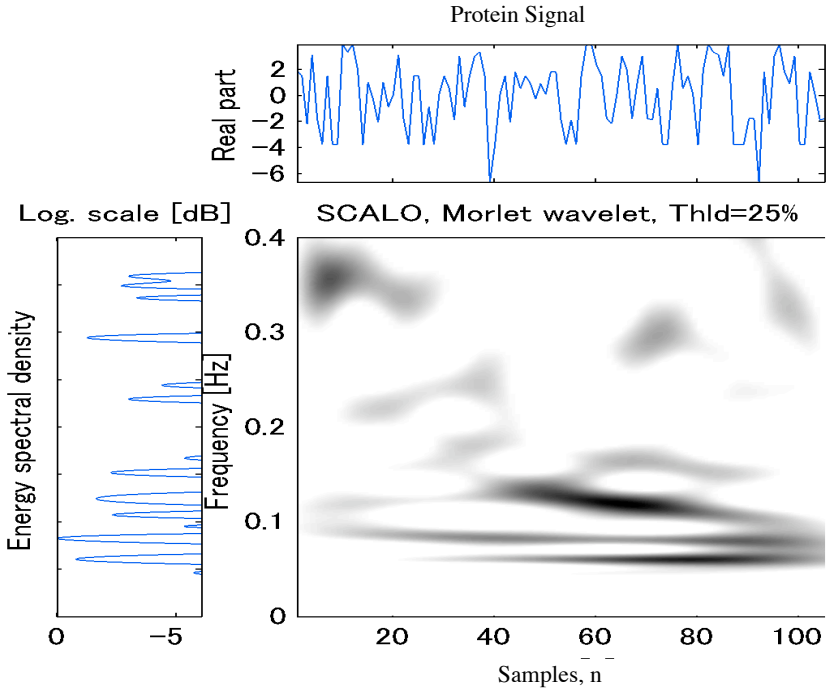

Figure 15 - Scalogram of cytochrome c

e. Prolactin

Prolactin has two binding sites. Binding Site I consists of amino acids in the ranges 19 - 37 and 169 -180 [20]. Site II consists of residues between the 21st to the 32nd amino acid residues and residues between the 110th to the 122nd amino acids.

The experimentally determined hot spot ranges fall within the ranges of high energy in both the STFT and CWT of prolactin. The brightest regions in the STFT (Figure 16) and CWT (Figure 17) occur around amino acids 10 – 28 and 28-90, 100-210. The 10-28 region is near prolactin's N-terminus which Clapp and Weiner [26] claim is of biological importance. In the 28 – 90 range the section of highest energy is located in the vicinity of the 37th amino acid which de Trad et al [16] cite as a binding determinant for human prolactin. The 58th to 74th amino acids are also known to be bioreactive [16]. The 129th amino acid residue is believed to be important in maintaining the geometry of the second binding site of prolactin [16]. The last region of high energy indicated is the 100th – 210th amino acids. Within this range the zone of highest energy radiates from the 129th amino acid. Residues 176 and 177 in the 100 – 200 are also binding determinants [16].

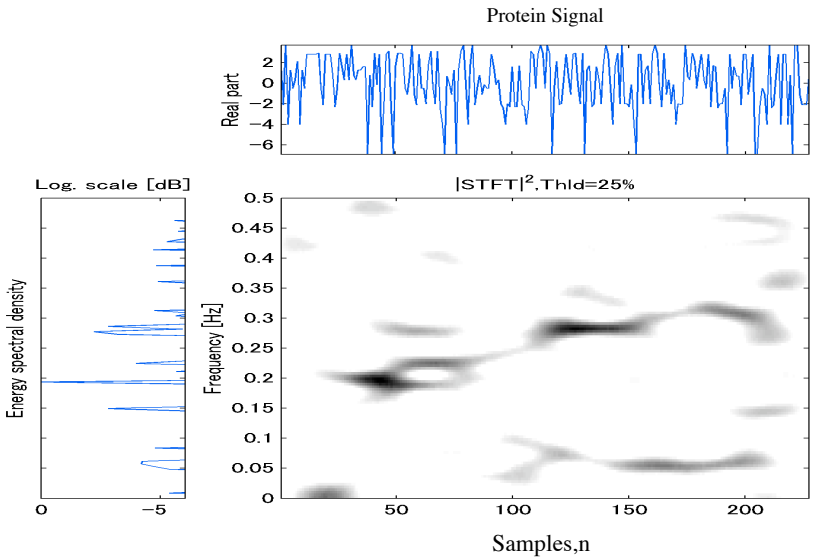

Figure 16 - STFT of prolactin

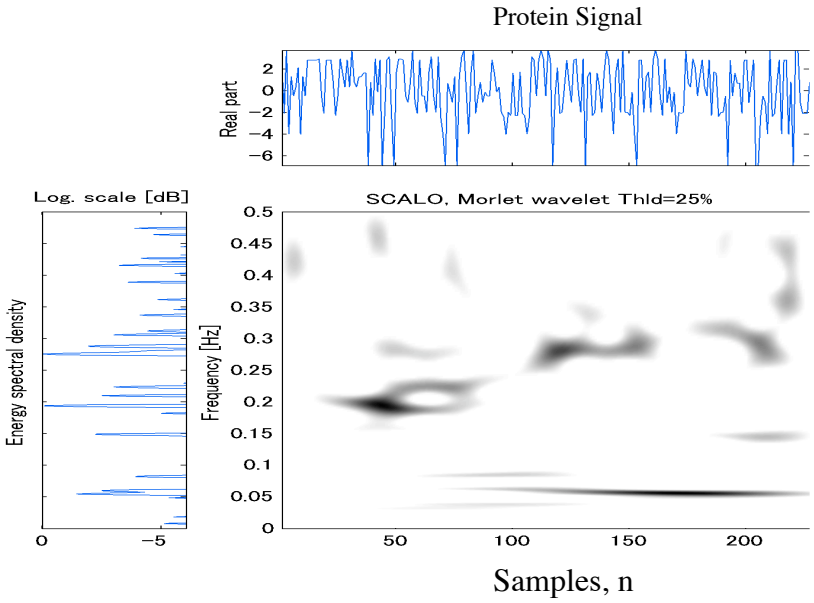

Figure 17– Scalogram of Prolactin

f. Lysozyme

Owing to its similarity with chicken / hen egg white lysozyme, human lysozyme is member of the chicken-type (c-type) lysozyme family. The c-type lysozymes have an active site divided into six subsites. The subsites are spread over the 35th to the 63rd and the 98th to 114th amino acid residues. The residues Glu-35 and Asp-52 while not part of any subsite play a catalytic role [23].

Figures 18 and 19 show that both the CWT and STFT capture the experimentally predicted "hot spots." In both figures, the region of highest-energy is the 98th to 114th amino acid zone. Owing to the excess energy appearing in the STFT image it appears that this is a case where the Continuous Wavelet Transform is better suited than the Short-Time Fourier Transform for protein analysis.

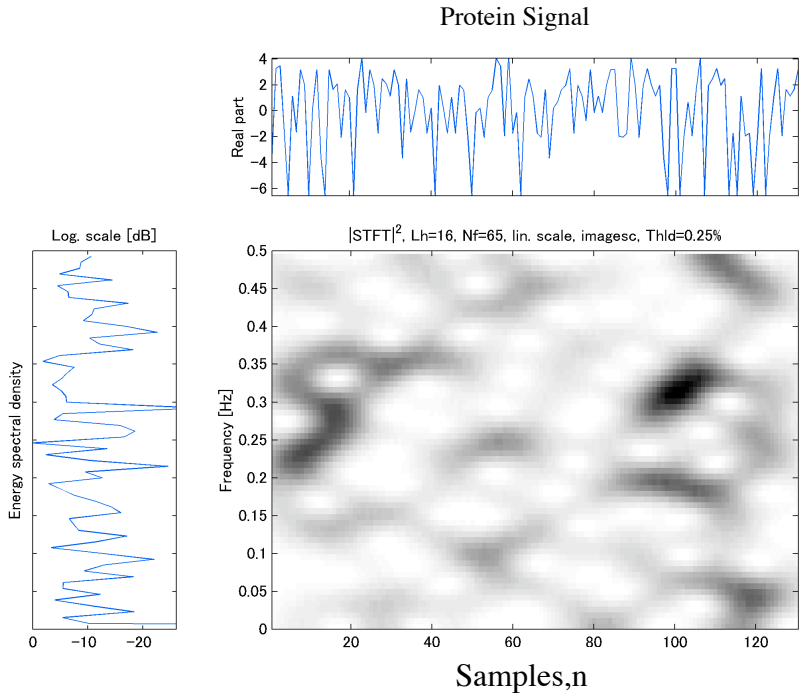

Figure 18– STFT of lysozyme

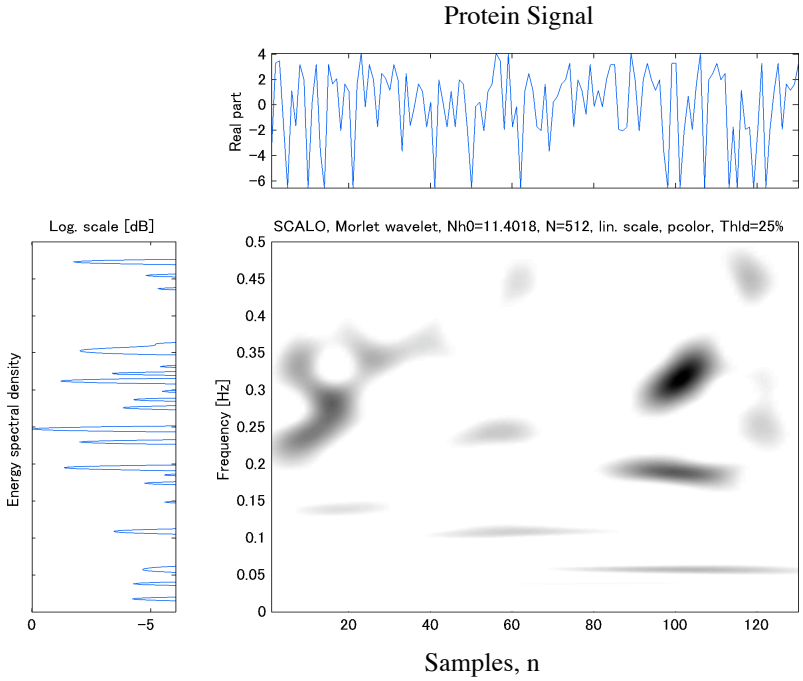

Figure 19 – Scalogram of human lysozyme

g. Epidermal Growth Factor

In their structure analysis of Epidermal Growth Factor, Groenen et al. [25] report binding affinity studies of fragments of EGF suggest that amino acid residues 4 to 48 are required for strong binding to the EGF-receptor.

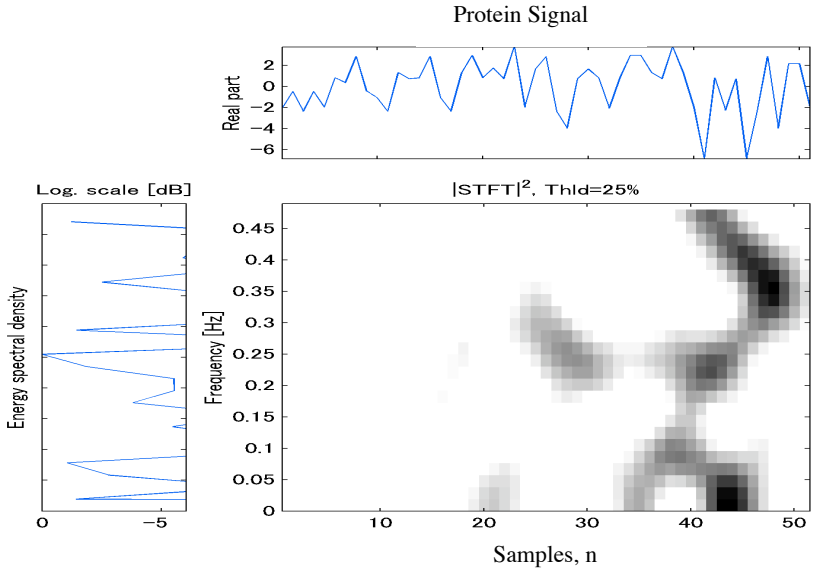

Figure 20 - STFT of human EGF

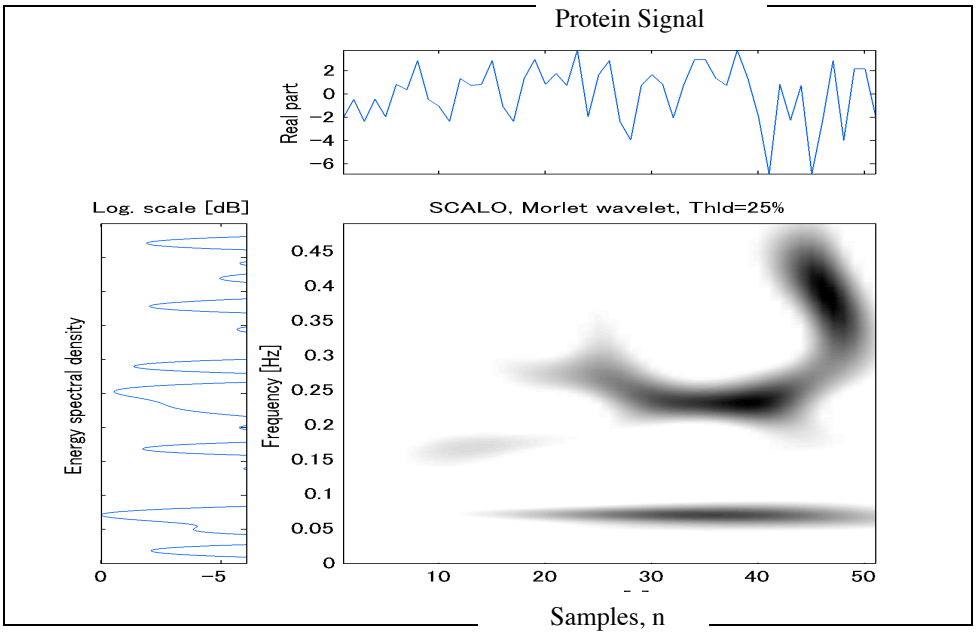

Figure 21– scalogram of human EGF

The STFT of human EGF (Figure 20) does not follow the Groenen et al's experimental results. However there are high-energy areas located in the vicinities of residues 37, 41,43 and 47 which Lu et al. [24] report to be among the most important residues for the function of human epidermal growth factor.

The scalogram in Figure 21 reveals that high-energy regions span the region from the 10th amino acid residue to slightly beyond the 48th amino acid residue. The lack of energy between the 4th to 10th amino acids might be due to the thresholding incorporated in the calculation of the scalogram. The 37th, 41st, 43rd and 47th amino acids appear in the regions with highest energy.

h. Glucagon

The hormone glucagon is an interesting molecule as all its 29 amino acid residues are critical to its role in inducing the breakdown of fatty acids and glycogen into glucose [20,21].

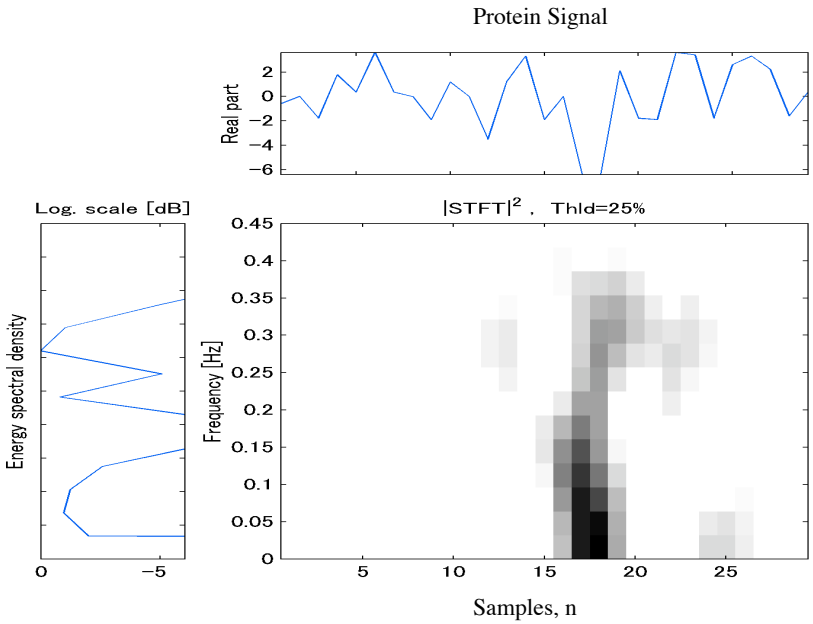

Figure 22 STFT of glucagon

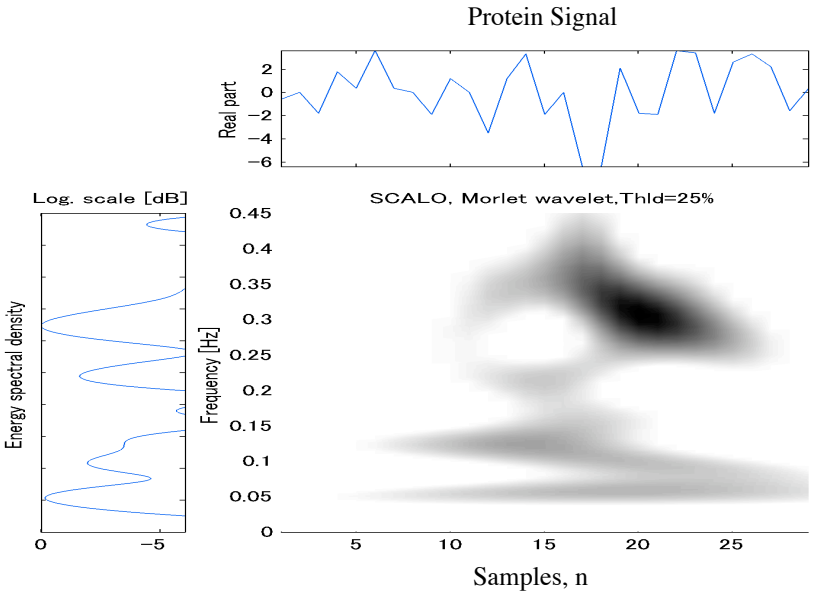

Figure 23 – Scalogram of Glucagon

The STFT (Fig. 22) fails drastically in this case as it signifies that only residues 15 to 25 are of importance. The CWT gives a much better result suggesting that the 7th right through to the 29th amino acids are important.

E. Conclusions

In Part I of this study we constructed protein signals based on amino acid hydrophobicity and analyzed them with the Short-Time Fourier Transform and the Continuous Wavelet Transform. Our objective was to locate the amino acids ("hot spots") critical to the function of these proteins in the resulting spectrogram and scalogram. Our results show that both the STFT and the CWT are able to locate a protein's "hot spots" given a protein signal based on amino acid hydrophobicities. This shows that hydrophobicity is a viable datum to be used in "hot spot" localization via Time-Frequency analysis. Our results also reveal that neither the Short-Time Fourier Transform nor the Continuous Wavelet Transform is significantly better than the other in

localizing “hot spots.” Future work includes pinpointing specific amino acids that contribute to protein activity instead of looking at residue ranges.

II. Investigating the Similarity Between Protein Sequences

A. Introduction

Protein structure is often highly related to protein function. A natural consequence of this observation is that proteins that perform related or similar functions often tend to be similar at some level of their structure. As major efforts are now under way to interpret genomic data and use the results in biology and medicine, protein comparison is an important task as it provides some information about possible functions of newly identified proteins.

The traditional method for protein similarity studies is sequence alignment that entails a cross-pairwise comparison of the amino acid sequences that make up a protein and the computation of a numerical similarity measure [30]. Other analysis techniques include searching proteins for common structural motifs and analysis with the Discrete Wavelet Transform[28]. We propose a novel approach to estimate the similarity between protein sequences based on the Kolmogorov-Smirnov test. Our results are also discussed in "Investigation of Protein Similarity Using the Kolmogorov-Smirnov Test" [36].

B Method

A cumulative distribution function, CDF(x) provides probability values as a function of random variables x. The function can be conceptualized as follows:

$$CDF(x) = \int_{-\infty}^{x} p(x)dx \qquad (6)$$

where *p(x)* is the probability density function. In CDF(x) the percentage is normalized to 1.

The two-sample Kolmogorov-Smirnov test computes a statistic that measures the similarity between two CDF(x). Given two cumulative distribution functions, the Kolmogorov-Smirnov statistic (K-S statistic) is defined as the absolute value of the maximum difference between the two distributions. The statistic can be conceptualized as follows

Let F(x) be CDF(x)

and let G(x) be CDF(y)

The Kolmogorov-Smirnov statistic is

$$K = \max\left[\left|F(x) - G(y)\right|\right] \qquad (7)$$

As the percentages in a cumulative distribution function are normalized to 1, the maximum value of the K-S statistic is 1. The larger the K-S statistic the more different the two data samples.

The electron ion-interaction potential (EIIP) is a numerical description of the average energy states of all valence electrons in a particular amino acid [29]. In our study, for each protein considered, we constructed two CDFs- one based on EIIP and the other on hydrophobicity- and ran the 2-sample Kolmogorov-Smirnov test on selected pairs of the proteins. Our test proteins are the alpha chain of human hemoglobin (Hahu), the beta chain of human hemoglobin (Hbhu), the alpha chain of horse hemoglobin (Haho), sperm whale myoglobin (myo), pig cytochrome c (Ccp), rat lysozyme (Lzrt) and lupine leghemoglobin (Legh).

Establishing a reasonable numerical demarcation for classification purposes is always a challenge. We take a K-S statistic that is less than 0.1 to be strongly correlated as proteins less than 10% different are definitely highly similar.

C. Results

Proteins Being Compared	K-S statistic (hydro)	K-S statistic (EIIP)
Hahu and Haho	0.0192	0.0478
Hahu and Hbhu	0.0404	0.1053
Myo and Legh	0.0741	0.050
Hahu and Ccp	0.1705	0.1290
Hahu and Lzrt	0.1328	0.1467

Table 1 - K-S statistic for five pairs of proteins.

For both the hydrophobicity (Fig. 24) and the EIIP-based comparisons the K-S statistic suggests that the most similar pair consists of the alpha chain of human hemoglobin and the alpha chain of horse hemoglobin. This result makes sense as both proteins are not only oxygen-carrying proteins but are also homologous.

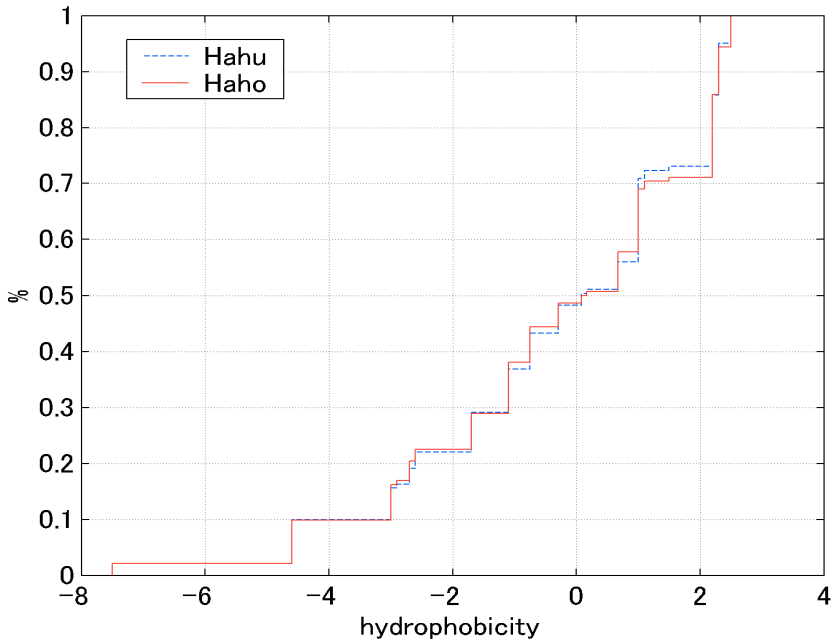

Fig. 24 – Hydrophobicity CDF of homologous and functionally-related Hahu and Haho

Hemoglobin is made up of two alpha subunits and two beta subunits. Each of these subunits serves the same function as they each carry one molecule of oxygen.

Comparing the alpha and beta subunits with the K-S test gives a K-S statistic of 0.0404 for the hydrophobicity-based comparison and 0.1053 for an EIIP based one. The K-S statistic for the hydrophobicity-based comparison of this pair of proteins is very low, rightfully suggesting a strong functional correlation. The EIIP-based K-S statistic exceeds 0.1 suggesting a weak correlation. This is the only case where the statistic obtained from the EIIP data appears unreliable as it predicts a weak correlation when a strong correlation is expected.

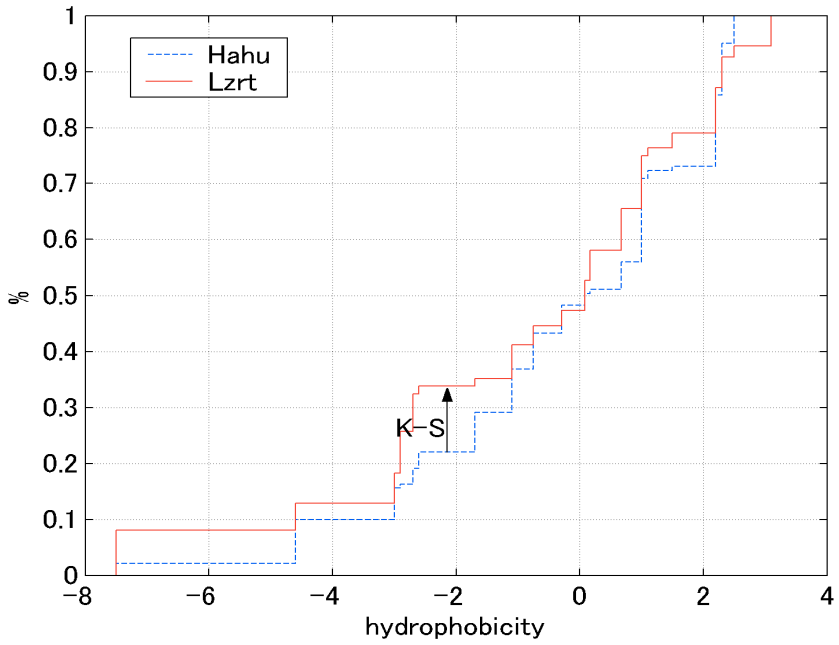

Figure 25 - Hydrophobicity CDF of functionally unrelated Hahu and Lzrt

The proteins considered so far have both been functionally related and also had fairly similar primary sequences. The sperm whale myoglobin – lupine leghemoglobin pair is distinguished by the fact that even though the two polypeptides in question serve similar purposes their primary sequences have only a 15% similarity [28]. The K-S statistic results for this comparison are 0.0741 (hydrophobicity) and 0.050 (EIIP). These

results are low to reflect the functional relationship between the two proteins. The statistics are not as low as the homologous alpha chains considered earlier suggesting a slightly weaker correlation between myoglobin and leghemoglobin. de Trad et al. point out that protein comparison techniques that obtain a strong correlation between these two proteins are laudable as the disparity between the primary structures makes it hard to detect a correlation via traditional sequence alignment methods [28].

The alpha chain of human hemoglobin and the antibacterial protein lysozyme (rat) are two unrelated polypeptides. As can be seen from Table 1 and Fig. 2 the K-S statistics of this pair are both above 0.1 predictably suggesting a weak functional correlation. A comparison of the alpha chain of human hemoglobin and the metabolic protein cytochrome C (pig) produced K-S statistic values of 0.1705 (hydrophobicity) and 0.1290 (EIIP) correctly predicting another weak correlation.

In most test cases, the K-S statistic correctly classified the similarity of protein pairs suggesting that it is a promising measure of protein similarity.

VI. Conclusions

We used the Kolmogorov-Smirnov test to compare protein sequences and classify them as weakly or strongly functionally related. Our data consisted of the hydrophobicity and the electron-ion interaction potential of the proteins. The results suggest that the Kolmogrov-Smirnov statistic is good measure of protein similarity. Future work involves determining a more rigorous demarcation for classifying proteins as weakly or strongly functionally related.

References

[1] U.S Department of Energy Primer on Molecular Genetics, http://www.ornl.gov/TechResources/Human_Genome/publicat/primer/toc.html

[2] Genome Management Information System: Genomics and its Impact on Society: A 2001 Primer ,

[3] CNN, Human Genome Map Ready in 2003 April 16, 2002 http://www.cnn.com/2002/TECH/science/04/15/china.genome.reut/index.html

[4] Campell N, *Biology,* Benjamin/Cumming Publishing Co., 1993

[5] http://members.tripod.com/beckysroom/protein_synthesis.jpg, April 17th 2002 11:00am

[6] http://esg-www.mit.edu:8001/esgbio/lm/proteins/aa/aminoacids.html, April 20, 2002 8:15pm

[7] http://cmgm.stanford.edu/biochem118/Papers/Protein%20Papers/Voet&Voet%20chapter 6.pdf

[8] http://www.rcsb.org/pdb/experimental_methods.html; April 21, 2002, 10:45pm

[9] http://tidepool.st.usm.edu/crswr/hemoglobin.html April 21, 2002; 11:15pm

[10] Garrett R. and Grisham C., *Biochemistry*, Fort Worth: Saunders College Publishing, 1999

[11] Carpenter, G and Cohen, S (1990) J Biol. Chem. 265, 7709-7712

[12] Alberts, B. et. al, *Essential Cell Biology*, New York: Garland Publishing, 1998

[13] White et al "Protein Classification by Stochastic modeling and Optimal Filtering of Amino Acid Sequences," Mathematical Biosciences 119:35-75(1994).

[14] Peng et al. “Long-range Correlations in Nucleotide Sequences,” Nature 356:168-170(1992)

[15] Pirogova E. and Cosic I, “Examination of Amino Acid Indexes Within the Resonant Recognition Model,” Proceedings of the IEEE Engineering in Medicine and Biology Society Conference: Biomedical Research in 2001

[16] de Trad C, Fang Q., Cosic I , “The Resonant Recognition Model (RRM) Predicts Amino Acids in Highly Conserved Regions of the Hormone Prolactin (PRL),” Biophysical Chemistry 84 (2000)149-157

[17] de Trad C., Fang Q, Cosic I. “An Overview of Protein Sequence Comparisons Using Wavelets,” Proceedings of the IEEE Engineering in Medicine and Biology Society Conference: Biomedical Research in 2001

[18] Lehmann H. and R.G Huntsman. *Man’s Haemoglobins Including the Haemoglinopathies and Their Investigation* Philadelphia Lippincott, 1966

[19] Tzagaloff, A. *Mitochondria* New York: Plenum Press 1982

[20] Norman, A and Litwak, G., *Hormones*, San Diego: Academic Press 1997

[21] Chou, P. and Fasman, G.D “The Conformation of Glucagon: Predictions and Consequences” Biochemistry 14:2536-2541(1975)

[22] Goffin et al, “Use of a Model to Understand Prolactin and Growth Hormone Specificities ,” Protein Engineering 8:1215-1231(1995)

[23] Strynadka N. and James M. “Lysozyme: A Model Enzyme in Protein Crystallography” in *Lysozymes- Model Enzymes in Biochemistry and Biology* Jolles P. (ed) Basel; Boston: Birkhauser, 1996

[24] Lu, H. et al. "Crystal Structure of Human Epidermal Growth factor and Its Dimerization" Journal of Biochemistry 276:34913-34917(2001)

[25] Groenen, L et al. "Structure-Function Relationships for the EGF /TGF-alpha Family of Mitogens" Growth Factors, 11:235-257(1994)

[26] Clapp C. and Weiner , R.I Endocrinology 130 (1992) 1380-1386

[27] Cosic I. et al. "Resonant Recognition Model and Protein Topography," European Journal of Biochemistry, 198, 113-119 (1991).

[28] C. de Trad, Q.Fang, and I. Cosic,"An Overview of Protein Sequence Comparisons Using Wavelets" Proceedings of the IEEE Engineering in Medicine and Biology Society Conference: Biomedical Research in 2001

[29] C. de Trad et. al, "The Resonant Recognition Model (RRM) Predicts Amino Acids in Highly Conserved Regions of the Hormone Prolactin (PRL)" Biophysical Chemistry 84 (2000) 149-157

[30] G. Barton, "Protein Sequence Alignment and Database Scanning" in *Protein Structure Prediction - a Practical Approach*, M. J. E. Sternberg, (Ed). IRL Press at Oxford University Press, 1996,

[31] Walker J., *Wavelets and their Scientific Applications*, Chapman and Hall, 1999

[32] Fischer R. and Akay, M. "Fractal Analysis of Heart Rate Variability" in Time Frequency Analysis and Wavelets in Biomedical Signal Processing, Akay M. (ed.), IEEE Press, 1998

[33] Akay, M. and Mello C. "Time-Frequency and Time-Scale (Wavelets) Analysis Methods: Design and Alogorithms" Smart Engg. Sys. Des. Vol. 1, pp 77-84

[34] Essien K, Akay M and Sekine M. "Time Frequency Analysis Based on Hydrophobicity," IEEE EMBS Asia/Pacific Conference 2002

[35] Cosic I, de Trad C., Fang Q and Akay M. "Protein Sequencing the RRM Model and WT methods. A comparative Study" Proc. of the IEEE EMBS Asia/Pasific Conference on BME, Sept. 2000.

[36] Essien K. Akay M., Sekine M. "Investigation of Protein Similarity Using the Kolmogrov-Smirnov Test," First International Summer School on Molecular Diagnosis and Applications, Izmir, Turkey, May 26-29, 2002.