31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

A new Approach to Revealing Functional Residues from Analysis of Protein Primary Structure

Vuk Vojisavljevic, Elena Pirogova, Dragomir Davidovic, Irena CosicSenior *Member, IEEE*

Abstract. A protein's biological function is encrypted within its primary structure. Nevertheless, revealing protein function from analysis of its primary structure is still unsolved problem. In this article we present a new methodology for determining functionally significant amino acid residues in proteins sequences, which is based on time-frequency signal analysis and Smoothed Pseudo Wigner Ville distribution (SPWV). This investigation is the extension of the Resonant Recognition Model (RRM) approach designed for structure-function analysis of proteins and DNA. The RRM is based on the finding that there is a significant correlation between spectra of the numerical presentation of amino acids and their biological activity. The RRM assumes that the selectivity of protein interactions is based on the resonant electromagnetic energy transfer at the specific frequency for each interaction. In this study Cytochrome C, Glucagon, and Hemoglobin proteins were used as the protein examples. By incorporating the SPWV distribution in the RRM, we can define the active regions along the protein molecule. In addition, it was also shown that our computational predictions are corresponding closely with the experimentally identified locations of the active/binding sites for the selected protein examples.

INTRODUCTION

With the rapid expansion of protein databases, the prediction of the biological function of newly sequenced proteins or the determination of their relationships with defined functional families becomes a real problem. The physical nature of a protein's biological function is based on its ability to interact selectively with particular targets (other proteins, DNA regulatory segments or small molecules). So far, the rules governing such selectivity have not been elucidated. Proteins play various biological roles such as catalysis of innumerable chemical reactions, they are active as carrier and storage molecules, responsible for immune protection, and also generation and transmission of nerve impulses. Proteins can be biologically "active" only by achievement of a certain active native conformation as a three-dimensional structure (3-D). It is generally accepted that three dimensional structures of proteins are fully predetermined by their primary structures. Consequently a protein's biological function is encrypted within the protein's primary structure, i.e. the sequence of amino acids. The RRM model [1]-[8] is able to determine a protein's functional and structural information by the analysis of its primary sequences using signal processing methods, Fourier and Wavelet Transforms. It is assumed that the selectivity of protein interactions is based on the resonant electromagnetic energy transfer at the specific frequency for each interaction. One of the main applications of this model is the prediction of active/binding site(s) location within protein primary structures [1],[2]. The main disadvantage of the signal analysis based on standard Fourier Transform is that the information about frequency characteristic along the series is hidden, and we can obtain only an averaged time and frequency content of the analyzed signal. In the last 20 years the time-frequency distribution methods have become powerful alternative tools for signal analysis. A time-frequency transform presents energy distribution of a signal over the time and frequency domains. In this study we applied the time-frequency signal processing technique to structure-function analysis of the selected proteins analyzed, aiming to demonstrate how the signal's energy is distributed over the two dimensional time-frequency space. By incorporating the Smoothed Pseudo Wigner Ville distribution (SPWV) in the standard RRM approach we intend to overcome the problem of non-localization events currently present in the model and improve the RRM predictive capabilities and accuracy for investigation of a proteins' physical characteristics.

Methodology

The Resonant Recognition Model

The RRM, central to this study, involves a transformation of the amino acid sequences into numerical sequences and then analysis of these sequences by appropriate digital signal processing methods, Fourier and Wavelet [1]-[8]. In the RRM, a protein's primary structure is presented as a numerical series by assigning to each amino acid a physical parameter value relevant to the protein's biological activity. Our previous investigations [1]-[8] as well as studies of other authors [9], have shown that the best correlation is achieved with parameters related to the energy of delocalized electrons from each amino acid. These findings can be explained by the fact that electrons delocalized from a particular amino acid have the strongest impact on the electronic distribution of energy in the entire protein. In this study the energy of delocalized electrons, calculated as the Electron Ion Interaction Potential (EIIP) [10] of each amino acid residue, was used. By assigning to each amino acid the EIIP value we convert the original protein sequence into a numerical sequence. This resulting numerical series represents then the distribution of free electron energies along the protein molecule. The numerical sequences obtained are analyzed using Discrete Fourier Transform (DFT) in order to extract information pertinent to the

Manuscript received April 23, 2009.

Vuk Vojisavljevic is with the Royal Melbourne institute of technology ph: 61399253077; fax: +61399252007; vuk.vojisavljevic@rmit.edu.au.

Elena.Pirogova, is with the Royal Melbourne Institute of Technology e-mail: Elena.pirogova@lrmit.edu.au.

Irena.Cosic is with the Royal Melbourne Institute of Technology e-mail: I.Cosic@rmit.edu.au.

Dragomir .Davidovic with Institite of Nuclear science – Vinca. Department of Radiation and Environmental Protection, ddavidovic@rt270.vin.bg.ac.yu

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

biological function. As the average distance between amino acid residues in a protein sequence is about 3.8 Å, it can be assumed that the points in the numerical sequence derived are equidistant. For further numerical analysis the distance between points in these numerical sequences is set at an arbitrary value d=1. Peak frequencies in the amplitude cross-spectral function define common frequency components of the two sequences analyzed. To determine the common frequency components for a group of protein sequences, we have calculated the absolute values of multiple cross-spectral function coefficients M, which are defined as follows [1]:

$$\left|M_n\right| = \left|X_{1,n}\right| \cdot \left|X_{2,n}\right| \cdots \left|X_{M,n}\right| \ldots n = 1,2,\ldots N/2 \qquad (1)$$

Peak frequencies in such a multiple cross-spectral function denote common frequency components for all sequences analyzed. The multiple cross-spectral function for a large group of sequences with the same biological function has been named "consensus spectrum". The presence of a distinct peak frequency in a consensus spectrum implies that all of the analyzed sequences within the group have one frequency component in common. This frequency is related to the biological function provided the following criteria are met:

1) One peak only exists for a group of protein sequences sharing the same biological function
2) No significant peak exists for biologically unrelated protein sequences
3) Peak frequencies are different for different biological functions.

In our previous research [1]-[8] the above criteria have been implemented and the following fundamental conclusion was drawn: each specific biological function of a given protein or DNA is characterized by a single frequency. It has been found in previous research that proteins with the same biological function have a common frequency in their numerical spectra and shown that each specific biological function of protein or regulatory DNA sequence(s) is characterized by a single frequency [1]. The results of our previous studies with a number of different protein families revealed that proteins and their interacting targets (receptors, binding proteins, inhibitors) display the same characteristic frequency in their interactions. However, it is obvious that one protein can participate in more than one biological process, i.e. revealing more than one biological function. Therefore, it has been postulated that the RRM frequency characterizes a particular biological process of interaction between selected bio-molecules. Further research in this direction has lead to the conclusion that interacting molecules have the same characteristic frequency but opposite phases at that frequency [1]. Thus, the RRM characteristic frequencies represent a proteins general functions as well as a mutual recognition between a particular protein and its target (receptor, ligand, etc). As this recognition arises from the matching of periodicities within the distribution of energies of free electrons along the interacting proteins, it can be regarded as the resonant recognition. Once the characteristic frequency for the particular biological function or interaction is determined, it becomes possible to identify the individual "hot spot" amino acids that contributed most to this specific characteristic frequency and thus, possibly to the observed biological behavior of the protein.

Time-frequency analysis

The Wigner quasi-distribution was initially introduced to replace the classical phase-space distribution in statistical physics with corresponding quantum analogue [11] Von Neumann [12] established a method where two non-simultaneously measurable quantum mechanical quantities, such as the coordinate and momentum, can be measured simultaneously with a limited precision. He also showed that all measurements, with limited accuracy, can be replaced by the absolutely accurate measurements of other quantities, which are related to their non-simultaneously measurable quantities. Although due to the uncertainty principle, the concept of phase space in quantum mechanics is somewhat problematic, various functions which bear some resemblance to true phase-space distribution functions of non-quantum world were introduced. They proved to be useful not only as calculation tools, but also provided insights into the relations between classical and quantum mechanics. The first of such functions was introduced by Wigner [11] to study quantum corrections in classical statistical mechanics. It is now known as the Wigner function. It may be shown [11],[12] that the phase space distribution, which is produced in simultaneous measurements of position and momentum, can be represented as a convolution of the Wigner function of considered quantum state and the Wigner function of the filter state, which represents a measuring apparatus.

In general, Wigner-Ville distribution (WVD) describes the frequency content changes as a function of time. The distribution is the actual energy intensity of various frequency components of the signal at a given position along the protein assuming that average distance between amino acid is set at an arbitrary value d=1. In practical calculations, convolution of the signal generates the cross term that represents interference of the signals, and consequently decreases significantly the resolution of the signal. A number of methods have been developed to reduce the cross-term [14]. In this investigation we replaced the WVD by the Smoothed Pseudo Wigner-Ville distribution (SPWVD), where some window functions are convolved with the WVD to restrain and decrease the effect of the interference terms.

Supposing *EIIP[i], i=1,2..N* is the numerical sequence of the Electron Ion Interaction Potentials of amino acids along the polypeptide chain, then the SPWVD of *EIIP[i]* is given by [13]:

$$S(t,f) =$$
$$\int_{-\infty}^{\infty} h(\tau) \int_{-\infty}^{\infty} g(s-t) z(s+\tau/2) z(s+\tau/2)^{*} ds \, e^{-j2\pi\nu\tau} d\tau$$

In discrete form the SPWVD can be calculated as:

$$W(n,m) =$$

$$\frac{1}{2}N \sum_{k=-N+1}^{N+1} \left|h(k)\right|^{2} \sum_{p=-M+1}^{M-1} g(p)z(n+p+k)z^{*}(n+p-k)e^{-\frac{2i\pi km}{M}}$$

Where *h(k)* and *g(p)* represent an independent frequency and time smoothing, respectively. In this study as the smoothing functions we used the Gauss filters, which are defined as:

$$h(k) = e^{(-k^{2}/2\sigma)/(\sigma\sqrt{2\pi})}$$

$$g(p) = e^{(-p^{2}/2\sigma)/(\sigma\sqrt{2\pi})}$$

where s is standard deviation , and k and p are the average values in the frequency and distance sets. The resulting SPWVD could be shown in a t-f plane as a contour plot according to the values of *S*(*t*,*f*) (Figures 1-3), which represents the distribution of the signal energy in the space domain. By choosing the standard deviation of the Gaussian functions h and g, we are practically balancing between the resolution in frequency and space domain interferences.

Results and Discussion

The selected protein groups, i.e. Cytochrome C, Haemoglobin and Glucagon, have been analyzed using both procedures, the standard RRM approach, and the SPWV transformation, aiming to determine the most functionally important regions within the protein sequences, i.e. the protein's active/binding sites. Furthermore, these computationally predicted regions have been compared with published experimental findings [15]-[16].

Glucagons

Glucagon is a small highly conserved peptide of 29 amino acids, which is responsible for maintaining normal concentrations of glucose in blood [16]. Recent structure-function studies of Glucagon demonstrated that a single replacement or deletion of either His (1) or Asp (9) amino acids in Glucagon may cause a 20- to 50-fold decrease in its activity, whereas these same changes made in tandem caused virtually complete loss of activity, with decreases of 10(4) to 10(6)-fold [16]. Glucagon characteristic frequency was identified at $f_{RRM} = 0.34$. Our calculations performed using the SPWV showed that the amino acids mostly contributing to the glucagon characteristic frequency $f_{SPWV} = 0.3±.0.05$ are located at 1-4, and this region covers the functionally important His (1) amino acid. A second frequency occurs at $f_{RRM} = 0.089$ ($f_{SPWV} = 0.095±.0.06$), which correspond to the amino acids region around Asp (9) (figure 1a, 1b).

Cytochrome C

The RRM was applied to structure-function analysis of the Cytochrome C proteins. The Cytochrome C is a soluble heme protein acting as the electron acceptor in respiratory electron transport chain in the mitochondria of eukaryotic organisms. Cross-spectral analysis of twenty nine Cytochrome C proteins from different origins revealed the common frequency component at the $f_{RRM} = 0.486$. All Cytochromes C proteins contain a cluster of highly conserved leucine amino acids at the positions 9, 68, 85, 94, and 98, which are located in the hydrophobic heme pocket of the protein. The protein structure-function mutagenesis studies demonstrated the great importance of these amino acids for maintenance of Cytochrome C overall structural integrity and electron transport activity [14]. Previous studies revealed that Leu (85) and Leu(94) are very important for the stability of hydrophobic interactions of the Cytochrome C complexes with the electron transfer partner. Our analysis performed on turkey Cytochrome C indicated the functionally important amino acids to be at the positions of 1-20 and 70-100. These regions are closely related to the identified RRM specific frequency $f_{RRM} = 0.486$. The SPWV spectra reveals the high energy area in the frequency interval $f_{SPWV} = $=0.46±0.02, which correspond to the amino acids 80-101. This finding accords well with the experimentally determined position of the heme cavity for Cytochrome C protein sequences (figure.2a, 2b).

Haemoglobins

Hemoglobins are polymers usually made from two units structurally and evolutionary related to each other. Each unit of haemoglobin non-covalently bind a single heme group. We analyzed the group of 19 $\tilde{\alpha}$-chains and compared results with those experimentally obtained. From crystallography data it is known for human haemoglobin that His (87) is attached to the heme group. Leu (83), Leu (91), His (51) and Lys (61) are in close proximity to the heme group and they play a major role in inhibiting its oxygenation [15]. SPWV distribution in the spatial frequency domain has an area of high energy located in the frequency interval corresponding to the RRM frequency $f_{SPWV} = 0.29±.0.02$ and include amino acids region of His (51)- Leu (91) (Figure 3a, 3b).

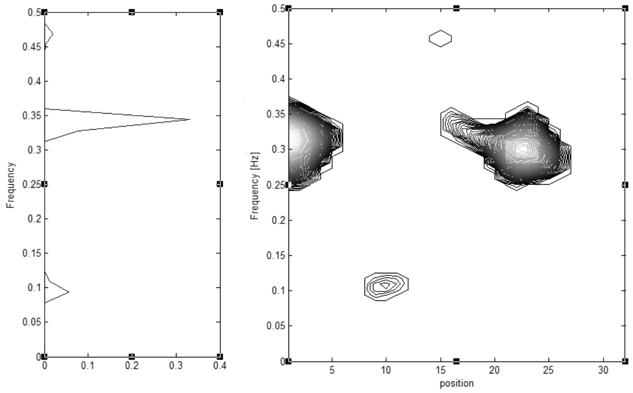

Figure 1. a) The RRM cross-spectral function calculated for 16 Glucagons, b) SPWV distribution calculated for chicken Glucagon. Amino acids His

(1) and Asp (9) are critical for Glucagon activity. It was also found that Ser residues at the positions of 2, 8, 11, and 16 are highly conserved.

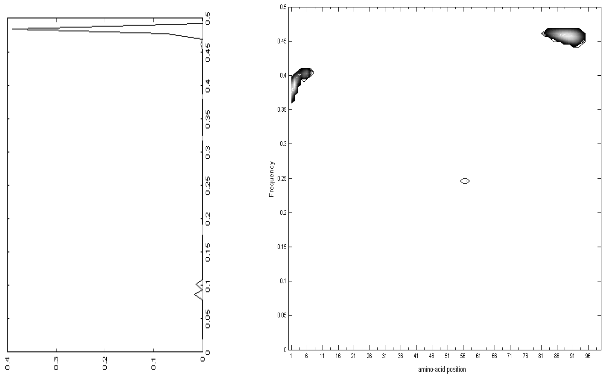

Figure 2. a) The RRM cross spectra calculated for the group of 29 Cytochromes C, b) SPWV distribution calculated for human Cytochrome C, c) All cytochromes C proteins contain a cluster of highly conserved leucine amino acids at 9, 68, 85, 94, and 98, which are located in the hydrophobic heme pocket of this protein

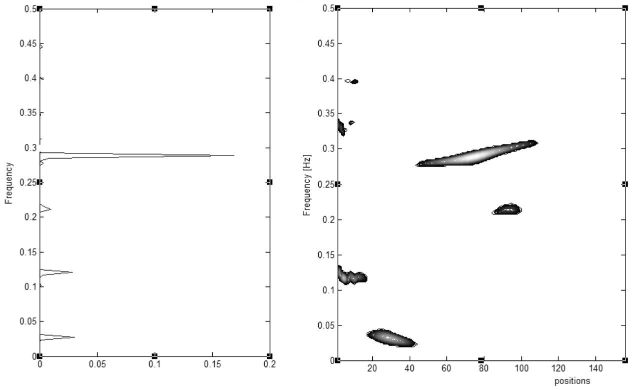

Figure 3. a) The RRM cross spectra calculated for 17 Haemoglobins alpha chains, b) SPWV distribution calculated for alpha chain of human Haemoglobin. From crystallography data it is known that His (87) is attached to heme group, and that Leu (83), Leu (91), His (51) and Lys (61) are in close proximity to heme group and play a major role in inhibiting hemes' oxygenation.

Conclusion

It has been shown previously that the RRM approach, based on the digital signal processing methods, can be successfully applied to structure-function analysis of different proteins. Due to limitations of the classical non-localised spectral transformations, in this study we attempted to use the SPWV distribution instead of the one-dimensional Fourier transform within the RRM approach. This new tool has been tested on selected proteins Cytochrome C, Glucagon and Haemoglobin. The results obtained have demonstrated that incorporation of the SPWV distribution in the RRM methodology improved both the accuracy and efficiency of the RRM predictive capabilities for protein active/binding sites allocation. Through use of SPWV distribution in the RRM we can not only predict the functionally important amino acids (as done in the standard RRM using the IFFT), but also define the active regions along the protein molecule. Another advantage of the SPWV is that we can also reduce the number of analyzed proteins that are required for accurate analysis. In particular, we can calculate the RRM frequency by using a limited number of protein sequences (from one to three proteins sequences). In addition, it was also shown that our computational predictions are corresponding closely with the experimentally identified locations of the active/binding sites for the selected protein examples.

Bibliography

[1] I. Cosic, *The Resonant Recognition Model of Macromolecular Bioactivity: Theory and Applications.* Basel:Birkhauser Verlag, 1997.

[2] I. Cosic, "Macromolecular Bioactivity: Is it Resonant Interaction between Macromolecules? - Theory and Applications," IEEE Trans. on Biomedical Engineering, 41, pp 1101-1114, 1994

[3] "Virtual Spectroscopy for Fun and Profit," Biotechnology, 13, pp. 236-238, 1995.

[4] E. Pirogova, V. Vojisavljevic, J. Fang and I. Cosic, "Computational analysis of DNA photolyases using digital signal processing methods," *Molecular Simulation* 32 (4) pp. 1195–1203, 2006.

[5] E. Pirogova, G.P. Simon, I. Cosic, , "Investigation of the applicability of Dielectric Relaxation properties of amino acid solutions within the Resonant Recognition Model," *IEEE Transactions on NanoBioscience* 2(2) pp. 63-69, 2003.

[6] E. Pirogova, Q. Fang, M. Akay, I. Cosic, "Investigation of the structure and function relationships of Oncogene proteins," *Proceeding of the IEEE* 90(12) pp. 1859-67, 2002.

[7] E. Pirogova, M. Akay and I. Cosic, "Investigating the Interaction Between Oncogene and Tumor Suppressor Protein," *IEEE Transaction on information technology in biomedicine*, 13(1) pp. 10-5, 2009.

[8] I. Cosic, "The Resonant Recognition Model of Bio-molecular Interactions: possibility of electromagnetic resonance," Polish Journal of Medical Physics and Engineering 7(1) pp. 73-87, 2001.

[9] V. Veljkovic, N Veljkovic, J. Aüther, U. Dietrich "Application Of The EIIP/ISM Bioinformatics Concept in Development of New Drugs." *Curr Med Chem* 14, pp 441-53, 2007.

[10] V. Veljković and M. Slavić, "General model of pseudo potentials," *Physical Review Letters* 29 pp. 105-10, 1972.

[11] E.P. Wigner, On the Quantum Correction For Thermodynamic Equilibrium, *Phys. Rev*. 40 pp. 749, 1932.

[12] D. Lalovic, D.M. Davidovic, and N. Bijedic, "Quantum mechanics in terms of non negative smoothed Wigner functions," *Phys. Rev. A* 46, pp. 1206–1212, 2003.

[13] B.Boashash, *Time-Frequency Signal Analysis and Processing* Prentice Hall PTR, 2005

[14] T. P. Lo, M.E. Murphy, J.G. Guillemette, M. Smith, and D. Gary. "Brayer Replacements in a conserved leucine cluster in the hydrophobic heme pocket of cytochrome *c," Protein Science* **4** pp. 198-208, 1995.

[15] D. Voet, G. Voet, *Biochemistry,* John Willey and son, 2004

[16] T.J. Kieffer, and J.F. Habener, "The glucagon-like peptides," *Endocr Rev.* 20(6) pp. 876-913, 1999.