Investigation of the Structural and Functional Relationships of Oncogene Proteins

ELENA PIROGOVA, QIANG FANG, MEMBER, IEEE, METIN AKAY, SENIOR MEMBER, IEEE, AND IRENA COSIC

Invited Paper

Proteins are the biomolecular workhorses driving the most biological processes in any living organism. These processes are based on selective interactions between particular proteins. So far, the rules governing the coding of the protein's biological function, i.e. its ability to selectively interact with other biomolecules, have not been elucidated. The resonant recognition model (RRM) is a novel physicomathematical approach established to analyze the interaction between a protein and its target. The RRM assumes that the specificities of protein interactions are based on the resonant electromagnetic energy transfer at the specific frequency for each interaction. One of the main applications of this model is to predict the location of a protein's biological active site(s) using digital signal processing. This paper incorporates the continuous wavelet transform (CWT) into the RRM to predict the active sites for a chosen protein example. We have investigated the oncogene functional group using digital signal analysis methods, in particular Fourier transform and CWT; determined oncogenes' characteristic frequency and functional active sites; and performed the design of the peptide analogous. The results obtained provide new insights into the structure–function relationships of the analyzed oncogene protein family.

Keywords*—Amino acids, characteristic frequency, digital signal processing, protein active site, protein function.*

I. INTRODUCTION

Proteins are one of the most complex and varied classes of macromolecules found in the cell. As enzymes, they catalyze innumerable chemical reactions that would otherwise occur slowly. They are active as carrier and storage molecules, in muscle contraction, and in mechanical support. As antibodies, they are responsible for immune protection; as receptors in the nervous system, they are responsible for the generation and transmission of nerve impulses. Proteins are polymers built up from amino acids. The great diversity and versatility of protein sequences are derived from the properties of 20 different amino acid side chains that may exist in a protein molecule, which are reflected in the wide range of bioactivity of the formed protein molecules. However, proteins are able to express their biological functions only by achievement of a certain active native conformation, the so-called three-dimensional (3-D) structure. Obviously, the particular function of a given protein and its active 3-D structure are determined by the sequence of amino acids forming this particular protein molecule. The protein's biological function is encrypted within the protein's primary structure, i.e., the sequence of amino acids. There have been many attempts to discover the main principles governing the functional behavior of proteins. Typical approaches are either homology characterization of specific features of the primary and secondary structure of proteins or molecular modeling of the protein's 3-D structure. Although such approaches permit a significant insight into the protein's structure and active site location, they still do not provide sufficient knowledge about informational, structural, and physicochemical parameters crucial to the selectivity of protein interactions that can be used for the *de novo* design of peptides or proteins analogous to the desired biological activity [1], [2].

The resonant recognition model (RRM) [1], [2] employed in this study essentially belongs to the approaches able to derive the protein's functional and structural information from the analysis of amino acid sequences and dopxyribonucleic acid (DNA). The RRM is a physical and mathematical model that interprets the protein sequence linear information using signal analysis methods. In the RRM, the protein's primary structure is represented as a numerical series by assigning to each amino acid in the sequence a physical parameter value relevant to the protein's biological activity.

Manuscript received April 30, 2002; revised September 8, 2002.

E. Pirogova, Q. Fang, and I. Cosic are with the School of Electrical and Computer Engineering, RMIT University, Melbourne 3001, VIC, Australia (e-mail: elena.pirogova@rmit.edu.au; irena.cosic@rmit.edu.au; john@mds.rmit.edu.au).

M. Akay is with the Thayer School of Engineering, Dartmouth College, Hanover, NH 03755-8000 USA (e-mail: makay@northstar.dartmouth.edu).

Digital Object Identifier 10.1109/JPROC.2002.805305

The RRM concept is based on the finding that there is a significant correlation between spectra of the numerical presentation of amino acids and their biological activity. It has been found through extensive research that proteins with the same biological function have a common frequency in their numerical spectra. This frequency was found to be a characteristic feature of a protein's biological function or interaction [1], [2]. Once the characteristic frequency for a particular protein function/interaction is identified, it is possible to use the RRM to predict the amino acids in the sequence which predominantly contributed to this frequency and consequently to the observed function. It also becomes possible to design peptides having the desired periodicities. Initially, these amino acids were identified using inverse Fourier transform (IFT) [1], [2]. Wavelet transform (WT), as a new signal-processing tool for multiresolution analysis and local feature extraction of nonstationary signals, has recently been incorporated into the RRM. In our recent studies, continuous wavelet transform (CWT) was successfully used to determine the functional active sites of different protein families [3]–[5]. The continuous scalogram of mouse epidermal growth factor (EGF), human β hemoglobin, prolactin, and tuna heart cytochrome c [3], [15] have been obtained using different wavelet functions, including Morley, Meyer, Daubechies, Simlets, Coiflets, and Mexican Hat. The combination of Fourier and WT methods has been proposed as a useful analytical tool in determining a protein's active site(s). Our preliminary studies suggested that the high-energy domains of EGF, hemoglobin, prolactin, and cytochrome c are sensitive to the wavelet type used in the analysis. Morley/Meyer wavelets are more successful in the identification of active sites or domains than other wavelet functions. However, better results could be obtained if a specific wavelet for the analysis of proteins were designed.

The focus of this study is therefore directed at solving the problem of functional and structural relationships of oncogene proteins using the RRM with the incorporated CWT method. Oncogenes are a specific group of growth factors that promotes uncontrolled cell growth and proliferation. These proteins are derived from normal cellular growth factors (so-called proto-oncogenes) by a limited number of modifications: mutations, insertions, or deletions. Because proto-oncogenes control the cell cycle, it is obvious that should a proto-oncogene be mutated, the potential for an unregulated cell cycle results. An unregulated cell cycle is the essence of cancer. Cells begin to divide uncontrollably, forming tissue masses, tumors, and the disease known as cancer. Here we have focused on the question whether there is a common characteristic of oncogene proteins that causes their ability to promote uncontrolled cell proliferation, and if so, whether it is possible to predict modifications in proto-oncogenes that cause their transformation into oncogenes.

II. Materials and Methods

With the rapid expansion of the protein databases, the identification of the biological function of newly sequenced proteins or the determination of their relationships with defined functional families becomes a real problem. Therefore, the introduction of additional information concerning the relationship between amino acids within the protein sequence would be helpful. The information encoded in the amino acid sequence ultimately determines the 3-D structure and biological function of a protein under physiological conditions. To understand empirical relationships between the amino acid sequence, structural patterns, and functional sites, the RRM has been invented. This model presents a completely new engineering approach to analysis of linear macromolecules: protein and DNA sequences [1], [2]. The physical nature of the biological function of a protein or DNA is based on the ability of the macromolecule to interact selectively with the particular targets (other proteins, DNA regulatory segments, or small molecules). According to the RRM, the information pertinent to the protein's biological function can be obtained by digital signal analysis of original amino acid sequences transformed into the numerical series representing the distribution of delocalized electron energies along the protein molecule. Taking into account the protein's conductive properties, a theoretical model of biologically relevant protein resonant frequencies was established. These frequencies were calculated and found to cover a very wide range including the infrared and visible light [2].

Here, we summarize the analytical methods (RRM and CWT) briefly.

A. The RRM Physicomathematical Basis

The RRM model incorporates digital signal-processing methods [1], [2]. It has been shown that certain periodicities (frequencies) within the distribution of energies of delocalized electrons along the protein molecule are critical for the protein's biological function (i.e., interaction with its target). Once the RRM characteristic frequency for a particular biological function or interaction has been determined, it is possible to identify the individual amino acids, the so-called hot spots, or domains that contribute most to the characteristic frequency and thus to the protein's biological function [1], [2]. The application of the RRM involves two stages of calculation. The first is the transformation of the amino acid sequence into a numerical sequence. Each amino acid is represented by the value of the electron-ion interaction potential (EIIP) describing the average energy states of all valence electrons in a given amino acid. The EIIP values for each amino acid were calculated using the following general model of pseudopotentials, [6] and are presented in Table 1:

$$\langle \overrightarrow{k+q}|\overrightarrow{w}|k\rangle = 0.25Z\sin(\pi 1.04Z)/(2\pi) \tag{1}$$

where q is a change of momentum of the delocalized electron in the interaction with potential w, while

$$Z = (\Sigma Z_i)/N \tag{2}$$

where Z_i is the number of valence electrons of the ith component of each amino acid and N is the total number of atoms in the amino acid. A unique number can thus represent

Table 1
EIIP Values of Amino Acids

amino acid	EIIP
Leu	0
Ile	0
Asn	0.0036
Gly	0.0050
Val	0.0057
Glu	0.0058
Pro	0.0198
His	0.0242
Lys	0.0371
Ala	0.0373
Tyr	0.0516
Trp	0.0548
Gln	0.0761
Met	0.0823
Ser	0.0829
Cys	0.0829
Thr	0.0941
Phe	0.0946
Arg	0.0959
Asp	0.1263

each amino acid or nucleotide, irrespective of its position in a sequence.

Numerical series obtained in this way are then analyzed by digital signal analysis methods in order to extract information relevant to the biological function. The original numerical sequence is transformed to the frequency domain using the discrete Fourier transform (DFT). As the average distance between amino acid residues in a polypeptide chain is about 3.8 Å, it can be assumed that the points derived in the numerical sequence are equidistant. For further numerical analysis, the distance between points in these numerical sequences is set at an arbitrary value $d = 1$. Then the maximum frequency in the spectrum is $F = 1/2d = 0.5$. The total number of points in the sequence influences the resolution of the spectrum only. Thus, for N point sequence, the resolution in the spectrum is equal to $1/N$. The nth point in the spectral function corresponds to the frequency $f = n/N$. To extract common spectral characteristics of sequences having the same or similar biological function, the following cross-spectral function was used:

$$S_n = X_n Y_n^* \qquad n = 1, 2, \ldots, N/2 \tag{3}$$

where X_n are the DFT coefficients of the series $x(m)$ and Y_n^* are complex conjugate DFT coefficients of the series $y(m)$. 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, the absolute values of multiple cross-spectral function coefficients M have been calculated as follows:

$$|M_n| = |X1_n|.|X2_n| \cdots |XM_n| \qquad n = 1, 2, \ldots, N/2. \tag{4}$$

Peak frequencies in such a multiple cross-spectral function denote common frequency components for all sequences analyzed (see, e.g., Fig. 1). Signal-to-noise ratio (S/N) for each

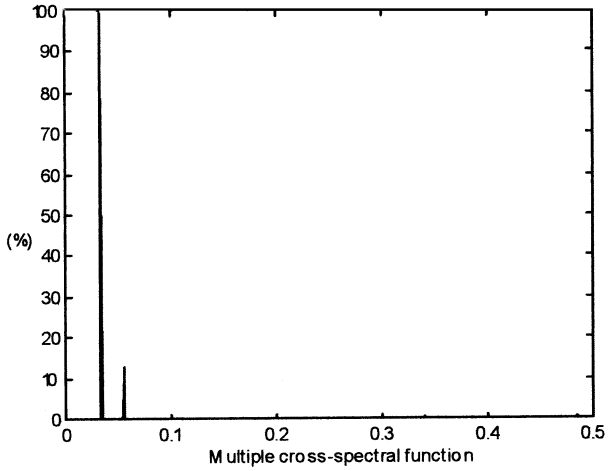


Fig. 1. Multiple cross-spectral function of oncogene proteins (46 sequences). The prominent peak(s) denote common frequency components. The abscissa represents RRM frequencies, and the ordinate is the normalized intensity.

peak is defined as a measure of similarity between sequences analyzed. S/N is calculated as the ratio between signal intensity at the particular peak frequency and the mean value over the whole spectrum. The extensive experience gained from previous research [1], [2], [7]–[9] suggests that an S/N of at least 20 can be considered as significant. The multiple cross-spectral functions for a large group of sequences with the same biological function have been named the consensus spectrum. The presence of a peak frequency with significant S/N 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) Only one peak 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 studies, the above criteria have been tested with more than 1000 proteins from 25 functional groups [1], [2]. The following fundamental conclusion was drawn from our studies: one RRM peak frequency characterizes one particular biological function or interaction [1], [2]. Therefore, those peaks are named as the RRM characteristic frequencies.

B. "Hot Spots" in Terms of the RRM and 3-D Protein Structures

It is known that proteins cannot express their biological function until they achieve a certain active 3-D conformation. By identifying the characteristic frequency of a particular protein, it is possible to predict which amino acids in the sequence predominantly contribute to the frequency and consequently to the observed function [1], [2], [7], [8]. Since the characteristic frequency correlates with the biological function, the positions of the amino acids that are most affected by the change of amplitude at the particular frequency can be

defined as hot spots for the corresponding biological function. The strategy for this prediction includes the following steps.

1) Determine the unique characteristic frequency for the specific biological function by multiple cross-spectral analysis for the group of sequences with the corresponding biological function.
2) Alter the amplitude at this characteristic frequency in the particular numerical spectrum. The criterion used for identifying the critical characteristic frequency change is the minimum number of hot spot amino acids that are least sensitive to further changes in the amplitude of the characteristic frequency.
3) Derive a numerical sequence from the modified spectrum using DFT.

It is known that a change in amplitude at one frequency in the spectrum causes changes at each point in the numerical sequence. Thus, a new numerical series is obtained where each point is different from those in the original series. Detecting the amino acids corresponding to each element of this new numerical sequence can then be achieved using tabulated values of the EIIP or other appropriate amino acid parameters. The amino acids in the new sequence that differ from the original ones reside at the points most contributing to the frequency. These hot spots are related to this frequency and to the corresponding biological function. The procedure described was used in a number of examples: IL-2 [8]; hemoglobins, myoglobins, and lysozymes [7]; chymotrypsins [11]; glucagons and tumor necrosis factors [9]; EGFs, fibroblast growth factors (FGFs), etc. [2]. These examples have shown that such predicted amino acids denote residues crucial for protein functions. Consequently, these hot spot amino acids are found spatially clustered in the protein's 3-D structure in and around the protein active site. As these specific amino acids strongly influence the characteristic frequency, their cluster represents a site in the protein where the signal of characteristic frequency for the specific protein property is dominant. Since this cluster of amino acids has been found positioned in and around the active site (see Fig. 2), it is proposed that these specific amino acids play a crucial role in determining the structure of the active site, and possibly the active structure of the whole molecule [7], [10].

C. Bioactive Peptide Design

Following the determination of the RRM characteristic frequencies and corresponding phases for particular biological functions, it is possible to design amino acid sequences having those spectral characteristics only. It is expected the designed peptide will exhibit the desired biological activity. The strategy for the design of such defined peptides is as follows.

1) Within the multiple cross-spectral analysis of the group of protein sequences sharing the corresponding biological function, determine the unique RRM frequency characterizing this specific biological function/interaction.

Fig. 2. Three-dimensional structure of the Ha-*ras* oncogene product p21 shown using the ribbon with CPK surfaces. The active site is denoted by the guanine triphosphate (GTP) molecule (in stick presentation) bound to the p21 oncogene product. It can be observed that all predicted hot spot amino acids are clustered around the active site.

2) Define the characteristic phases at the characteristic frequencies for the particular protein that is chosen as the parent for agonist/antagonist peptide design.
3) Derive a numerical sequence from the known characteristic frequencies and phases. This can be done by summing sinusoids of the particular frequencies, amplitudes, and phases. The length of the numerical sequence is defined by the appropriate frequency resolution and the required peptide's length.
4) Determine the amino acids that correspond to each element of the new numerical sequence. It can be achieved by the tabulated EIIP or other appropriate amino acid parameters [1], [2], [9].

D. The Continuous Wavelet Transform Model

Using IFT, we can identify only a number of single amino acids mostly contributed to the particular frequency. However, the protein active site is usually composed of domain(s) within the protein molecule. Applying the WT, we observe a whole frequency/spatial distribution and thus are able to identify the domain(s) of high energy of a particular frequency along the sequence. The CWT is a relatively new signal-processing tool effective for multiresolution analysis and local feature extraction of nonstationary signals [4]. The WT can be viewed as an inner product operation that measures the similarity or cross correlation between the signal and the wavelets. The continuous version of the WT of the signal (t) is defined as

$$cwt(a, b) = \int s(t) \frac{1}{(a)^{1/2}} \Psi \frac{(t-b)}{a} dt \qquad (5)$$

where b is the shift factor (the translation factor of the wavelet function along the time axis) and a is the scale factor (it scales a function by compressing or stretching it). CWT is one of the time- or space-frequency representations. A time

(space)-frequency representation of a signal provides information about how the spectral content of the signal evolves with time (space), thus providing an ideal tool to dissect, analyze, and interpret signals with transients or localized events. This is performed by mapping a one-dimensional signal in the time (space) domain into two-dimensional time (space)-frequency representation of the signal. Because CWT provides the same time/space resolution for each scale, CWT can be chosen to localize individual events, such as the active site identification. The particular wavelet chosen here for critical amino acid identification is the Morlet, which is a locally periodic wave-train

$$\omega(t) = C \exp\left(\frac{-t^2}{2} + j\omega_0 t\right) \tag{6}$$

where $\omega_0 = 5.33$ and C is the constant used for normalization.

From (6), it can be seen that the Morlet wavelet is a complex sine wave modulated by a Gaussian function. The time-frequency version of CWT can be achieved by making the substitution $a = f_0/f$

$$cwt(t, f) = \int s(\tau)(f/f_0)^{1/2}\psi(f/f_0(\tau - t))\, d\tau \tag{7}$$

in which the analyzing wavelet becomes essentially a prototype band-pass filter with center time $t = 0$ and center frequency f_0. The center frequency and frequency bandwidth of the CWT vary with scale. However, their ratio remains fixed. It is the constant property of the wavelet. The underlying property of wavelets is that they are pretty well localized in both time and frequency [4]. A product of the uncertainties of both time and the frequency is bound by the Heisenberg uncertainty principle; no filter can have a width product smaller than $1/\pi$. The Gaussian filters (Morlet wavelet) attain this theoretical limit. Strictly speaking, the CWT provides a time-scale representation rather than a time-frequency representation. However, the scale factor of CWT is closely related to the frequency, and this makes the mapping from time-scale representation to time-frequency representation possible. The active sites along the protein sequence are determined through studying the set of local extrema of the moduli in the WT domain. Those energy-concentrated local extrema are the locations of sharp variations points of the EIIP and are proposed as the most critical locations for protein's biological functions. The wavelet approach incorporated to the RRM has been tested on a number of different protein groups [3], , [5], [12]–[15].

III. Results

In this paper, the RRM approach has been applied to the analysis of the oncogene protein family for the understanding of the structural and functional relationship within this protein group. All sequences have been taken from the Protein Data Bank, Brookhaven National Laboratory. Twenty-eight viral and 18 cellular proteins, which are the products of *myc*, *myb*, *mos*, *fes*, *fps*, *fgr*, *fms*, *erb*, *ras*, *src*, *abl*, *yes*, *syn*, and *int*, have been analyzed within the RRM. Here we have determined the RRM characteristic frequencies of analyzed oncogenes as a whole functional group (46 sequences), as well as the specific characteristics of different subgroups: viral (28 sequences) and cellular (18 sequences) oncogene proteins. Also we have identified the hot spots, or domains that contribute mostly to the observed protein's biological function of the selected protein sequences. The model protein studied here is the p21 *ras* oncogene product (Harvey Murine sarcoma virus) shared with a number of other oncogenic proteins the ability to transform cells. In this paper, we have predicted the active sites for the chosen protein example using the CWT incorporated into the RRM. In addition, the computational design of peptide analogous, based on the frequency and phase predicted by the RRM, has been performed.

A. The RRM Characteristic Frequency of Oncogenes

The RRM approach has been applied to a group of 46 oncogenes, with the aim of ascertaining their RRM frequency characteristics. As a result, there is one prominent frequency component at $f = 0.0322 \pm 0.004$, $S/N = 467.98$ in the cross-spectral function (see Fig. 1), common to the analyzed protein sequences related to the Ha-*ras* family. According to the RRM axioms, the result suggests that this common frequency characterizes a common biological activity of this group of oncogene products, i.e., their ability to transform cells. Then, the whole oncogene functional group consisting of 46 sequences was divided into two subgroups according to their originality, and the RRM analysis was performed for the group of 28 viral oncogenes and 18 cellular proteins respectively. The obtained RRM characteristic frequencies are as follows: $f_1 = 0.0322$, $S/N = 297.29$, and $f_2 = 0.0537 \pm 0.004$, $S/N = 199.74$. As is evident from Fig. 1, both identified frequencies with significantly different amplitude ratios are observed in the cross-spectral function of all oncogene proteins. The following fundamental conclusion was drawn from our previous studies [1], [2]: each specific biological function of the protein is characterized by a single frequency. Thus, two peak frequencies detected by the RRM correspond to two different protein functions identified for these groups of protein sequences.

B. Hot Spots Prediction and the Peptide's Design

Ras-p21 proteins, the products of the *ras* oncogenes and proto-oncogenes, are guanine nucleotide binding proteins functioning as molecular switches in the signal transduction processes in the cell, regulating cell proliferation and differentiation. These proteins exist in an active or inactive conformational state. This state depends on the attachment of growth factors to the extracellular receptors, and a large number of effector molecules to the protein. Certain mutations of the protein determined in 30% of human tumors have been found to negatively regulate the intrinsic as well as effector-activated GTPase activity of the protein (GAP). The mutations are usually found in only two residues: Gly-12 or Gln-61 of the p21 protein. Gln61 was found to be particularly important, as mutations of this residue eliminate the sensitivity to GAPs. In contrast to cellular p21, oncogenic p21 mutants are not able to function

as signal switch molecules and are constantly producing a growth-promoting signal [16], [17].

In this paper, Ha-*ras* p21 sequence (Harvey Murine sarcoma virus) was used as a protein model for further analysis within the RRM. Once the common characteristic frequency of the analyzed oncogene proteins was identified, the hot spot analysis of the Ha-*ras* p21 sequence was performed using IFT. This analysis was intended to allocate the amino acids related to the frequency $f = 0.0322 \pm 0.004$ identified previously. In our results, the following amino acids were found as hot spots for the analyzed sequence: Gly 10, Gly 12, Gly 13, Gly 15, Phe 28, Gly 48, and Ser 65. Furthermore, the analysis indicated that Gly 10, Gly 12, Gly 13, Gly 15, and Phe 28 contribute to the characteristic frequency for p21 oncogenic and GTP-binding function to a greater extent than Gly 48 and Ser 65 [18], [19]. These predictions have been compared with other biological and crystallographic findings [16], [17] relevant to the functional and binding sites of the analyzed Ha-*ras* p21 protein sequence. To validate the above predictions, the hot spot amino acids were superimposed on the known 3-D crystal structure of p21. The results are shown in Fig. 2. In addition, we have identified the hot spot amino acids related to the "nonsignificant" frequency $f = 0.0537 \pm 0.004$. These amino acids have been superimposed previously and are shown in Fig. 3.

Following the preliminary gains made through our study, the RRM has been applied to Ha-*ras* p21 protein to design the peptide that exhibits *ras*-like activity, i.e., the ability to transform cells. The design is based on the characteristic frequency and phase determined within the RRM for the oncogene proteins ($f = 0.0322$, $\varphi = 2.027$). We have designed an amino acid sequence that has only this spectral characteristic. This bioactive sequence designed *de novo* has only the desired biological function related to the chosen characteristic frequency and purported to have the corresponding biological activity. The designed sequence is as follows:

— One-letter abbreviation:
DDRTQWYKHPENLINEPHA
— Three-letter abbreviation:
Asp-Asp-Arg-Trn-Gln-Trp-Tyr-Lys-His-Pro-Glu-Asn-Leu-Ile-Asn-Glu-Pro-His-Ala

C. CWT in the Protein Sequence Analysis

It should be mentioned that by using IFT, it is possible to identify only a number of single amino acids that contribute to the particular frequency. However, the protein active site is usually built up of domain(s) within the protein sequence. Applying WT leads to the possibility of observing a whole frequency/spatial distribution along the sequence and thus identifying domains of high energy of particular frequency for this protein molecule. The results obtained within the study are considered to be useful inputs toward the finding of the appropriate wavelet function(s) for the analysis of different protein sequences. Here we have compared the performance of different wavelet functions, including Morlet, Meyer, Daubechies, Simlets, Coiflets, and Mexican Hat, to improve the detection of the active sites of the oncogene protein with previously determined characteristics. The continuous scalograms of Ha-*ras* p21 oncogene product using different wavelets functions are shown in Figs. 4–9.

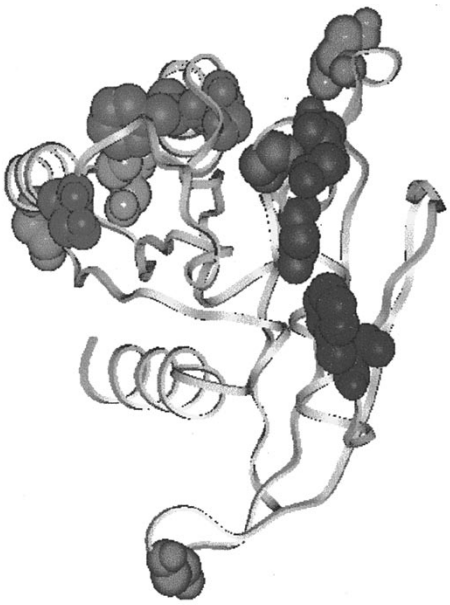


Fig. 3. Three-dimensional structure of Ha-*ras* oncogene product p21 shown in ribon presentation. Two separate clusters of the "hot spot" amino acids are highlighted with CPK surfaces and related to the frequency $f = 0.0322 \pm 0.004$ and to the $f = 0.0537 \pm 0.004$ respectively.

IV. Discussion

The results of hot spot analysis within the RRM and the further comparison of predicted and experimentally identified active sites (see Fig. 2) have led us to the following significant conclusions. As mentioned, the mutations in the p21 protein at amino acid positions 12, 13, and 61 have been found in a high percentage of human tumors. The mutations at these residue positions have a significant effect on the nucleotide dissociation rate constant of p21 and have been implied in oncogenic activation [16], [17]. Three out of five predicted hot spot amino acids are found to be among the active site residue (Gly 12, Gly 13, and Phe 28), while the remaining residues Gly 10 and Gly 15, together with Gly12, Gly 13, and Phe 28, represent the part of the continuous topological surface around the guanine-binding site. Thus, the predicted hot spot positions represent the site where the signal of the characteristic frequency ($f = 0.0322 \pm 0.004$) is dominant and consequently can act as a resonator for this characteristic frequency. Therefore, this characteristic frequency may dictate the specificity of the protein interactions and the selectivity of the subsequent energy transfer associated with the functional consequences of the biomolecular interactions. The results with the Ha-*ras* p21 oncogene product validate the RRM concepts and indicate a new strategy to characterize and interpret the informational content of oncogene proteins relevant to the cellular transformation.

Results of the use of CWT to the structural and functional analysis of the Ha-*ras* p21 protein sequence (see Figs. 4–9) reveal that the best prediction of functional active sites of this protein molecule can be gained by applying the Morlet

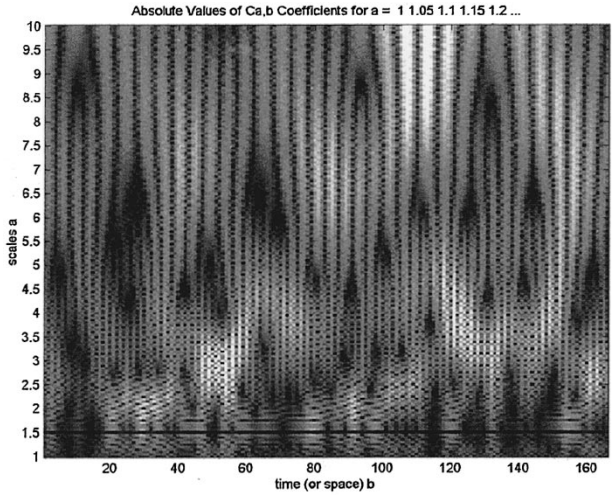


Fig. 4. CWT of Ha-*ras* p21 oncogene using Morlet wavelets function.

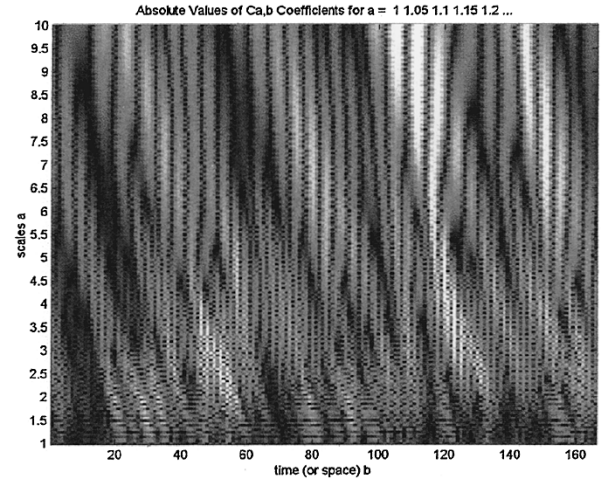


Fig. 5. CWT of Ha-*ras* p21 oncogene using Daubechis wavelets function.

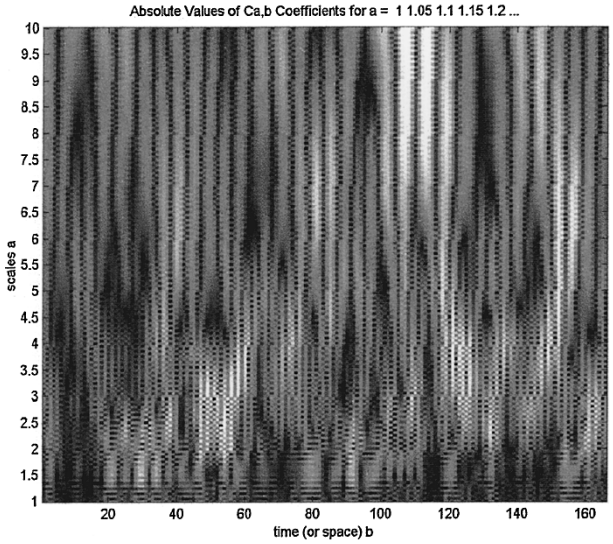


Fig. 6. CWT of Ha-*ras* p21 oncogene using Myer wavelets function.

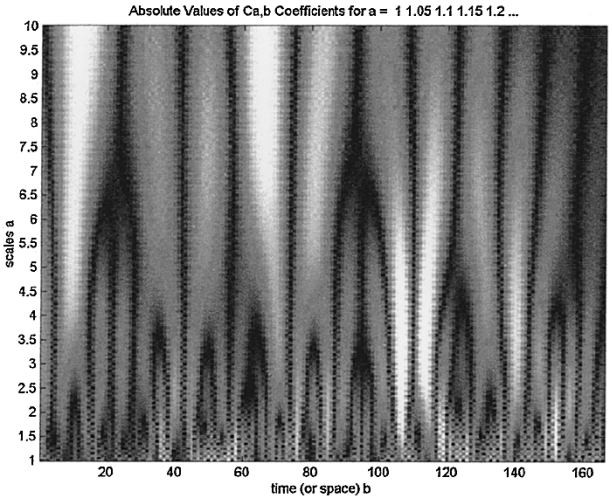


Fig. 7. CWT of Ha-*ras* p21 oncogene using Mexican hat wavelet function.

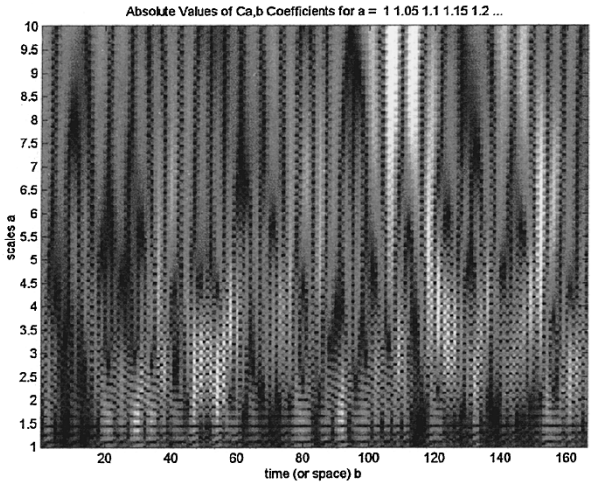


Fig. 8. CWT of Ha-*ras* p21 oncogene using Coiflets wavelets function.

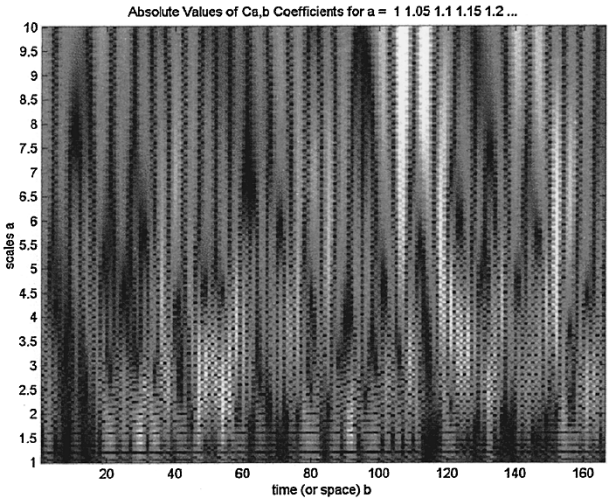


Fig. 9. CWT of Ha-*ras* p21 oncogene using Simlets wavelets function.

wavelets. This corresponds to our previous work, [14], [15], where we have shown that Morlet wavelets were the most suitable for the identification of active sites of EGF, FGF, and other protein sequences. The continuous scalogram of the

Ha-*ras* p21 protein (Harvey Murine Sarcoma virus) obtained by using CWT (Morlet wavelets) is presented in Fig. 4. As mentioned, the common frequency component of oncogene proteins corresponding to the common biological function (the ability to transform cells) was found at $f = 0.0322 \pm 0.004$. It can be clearly observed that at lower frequencies (the upper part of the scalogram), there is a definite area of high energy between the 100th and the 120th amino acid. This domain corresponds to the last part of the domain experimentally predicted to be the Ha-*ras* guamine-binding domain [16], [17]. The real form of the Morlet wavelet function is $w(t) = \exp(-t^2/2) * \cos(5t)$. There are two constants in this function, two and five. Because two determines the waveform amplitude modulation degree and five determines the center frequency, they are named here as the amplitude factor and the frequency factor. To find out if we can optimize the function for the use with proteins, we have modified both these factors to produce wavelets having similar shape but different center frequencies and modulation degrees. All the scalograms generated here have a maximum scale of ten.

V. Conclusion

In this paper, we have tested the RRM concepts applied to the analysis of oncogene proteins to gain knowledge about their structural and functional relationships. Although the analyzed oncogenes are discerned by the diversity of their functions, and no homologous sequences have been selected for the analysis, the RRM approach was revealed to be efficient to identify the common characteristic frequency and thus the common biological activity within the group, as well as to predict the oncogene functional active site. Particularly, the results obtained clearly indicate that while the Fourier approach for active site identification finds the specific residues that affect the RRM characteristic frequency, the wavelet approach identifies the amino acid domains indicated the sharpest variation locations of energy states and hence possibly form the so-called resonant boxes. This paper presents the analysis of the application of different wavelets' functions for their possible use in the identification of active sites of the Ha-*ras* p21 oncogene protein. It has been shown that the results depend on the particular function used, and it was suggested that Morlet wavelets are the most appropriate function to use in the active site prediction analysis. Thus, the domain identified at the 100th and 120th amino acid within the CWT is proposed as Ha-*ras* p21 protein functional and structural active site. This site predicted by the RRM corresponds to the site(s) determined experimentally by other authors. Importantly, with the incorporation of the CWT into the RRM, the prediction of the protein's active sites has been improved. This is largely due to the advantageous properties of the space-frequency analysis pertinent to the CWT. We believe that this paper, based on the protein analysis within the RRM, presents an additional insight toward the understanding of the structural and functional relationships of oncogene proteins. Finally, the next step in our research would be to generalize this analysis for a number of other protein groups.

References

[1] I. Cosic, "Macromolecular bioactivity: Is it resonant interaction between molecules?—Theory and applications," *IEEE Trans. Biomed. Eng.*, vol. 41, pp. 1101–1114, Dec. 1994.

[2] ——, *The Resonant Recognition Model of Macromolecular Bioactivity*. Basel, Switzerland: Birkhauser Verlag, 1997.

[3] I. Cosic, C. H. deTrad, Q. Fang, and M. Akay, "Protein sequences analysis using the RRM model and wavelet transform methods: A comparative study analysis," in *Proc. IEEE-EMBS Asia-Pacific Conf. Biomed. Eng.*, 2000, pp. 405–406.

[4] M. Akay, *Time Frequency and Wavelets in Biological Signal Processing*. Piscataway, NJ: IEEE Press, 1998.

[5] C. H. de Trad, Q. Fang, and I. Cosic, "The resonant recognition model (RRM) predicts amino acid residues in highly conservative regions of the hormone prolactin (PRL)," *Biophys. Chem.*, vol. 84, no. 2, pp. 149–157, 2000.

[6] I. Veljkovic and M. Slavic, "General model of pseudopotentials," *Phys. Rev. Lett.*, vol. 29, pp. 105–108, 1972.

[7] I. Cosic, A. N. Hodder, M. I. Aguilar, and M. T. W. Hearn, "Resonant recognition model and protein topography: Model studies with myoglobin, hemoglobin and lysozyme," *Eur. J. Biochem.*, vol. 198, pp. 113–119, 1991.

[8] I. Cosic, M. Pavlovic, and V. Vojisavljevic, "Prediction of 'hot spots' in Il-2 cased on information spectrum characterictics of growth regaling factors," *Biochemie*, vol. 71, pp. 333–342, 1989.

[9] I. Cosic, "Virtual spectroscopy for fun and profit," *Bio/Technology*, vol. 13, pp. 236–238, 1995.

[10] L. Baranyi, I. Cosic, W. Campbell, E. Deretey, V. Deretey, N. Okada, and H. Okada, "Antisense homology boxes coincide with the 'hot spot' regions predicted by resonant recognition theory," in *Proc. 2nd Int. Conf. Bioelectromagnetism*, 1998, pp. 67–68.

[11] I. Cosic, "Correlation between predicted and measured characteristic freqiency of chymotrypsin activation," in *Proc. 15th Ann. Conf. IEEE EMBS*, vol. 15, 1994, pp. 265–266.

[12] I. Cosic, Q. Fang, and E. Pirogova, "Modification of the RRM model using wavelets transform and ionization constant to predict protein active sites," *Proc. 21th Ann. Conf. IEEE EMBS* , vol. 21, no. 2, p. 1215, 1999.

[13] I. Cosic and Q. Fang, "Prediction of proteins active sites using digital signal processing methods," in *Proc. 2nd Int. Conf. Bioelectromagnetism*, 1998, pp. 69–70.

[14] Q. Fang and I. Cosic, "Prediction of active sites of fibroblast growth factors using continuous wavelets transform and the resonant recognition model," in *Proc. Inaugural Conf. Victorian Chapter IEEE EMBS*, 1999, pp. 211–214.

[15] I. Cosic and Q. Fang, "Evaluation of different wavelet constructions (designs) for analysis of protein sequences ," in Proc. of the 14th Int. Conference on DSP, 2002, to be published.

[16] E. F. Pai, W. Kabsch, U. Krengel, K. C. Holmes, J. John, and A. Wittinhofer, "Structure of the guanine–nucleotide binding domain of the Ha-ras oncogene product p21 in the triphosphate conformation," *Nature*, vol. 31, pp. 209–214, 1989.

[17] M. Barbacid, "Ras genes," *Ann. Rev. Biochem.*, vol. 56, pp. 779–827, 1987.

[18] I. Cosic and M. T. W. Hearn, "'Hot spot' amino acid distribution in Ha-ras oncogene product p21: Relationship to guanine binding site," *J. Mol. Recog.*, vol. 4, pp. 57–62, 1991.

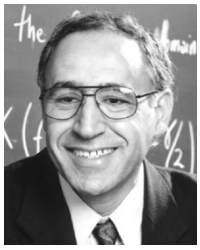
[19] ——, "Studies on protein–DNA interactions using the resonant recognition model. Application to repressors and transforming proteins," *Eur. J. Biochem.*, vol. 205, pp. 613–619, 1992.

Elena Pirogova received the B.Eng. degree in chemical engineering in 1991 from National Technical University of Ukraine, Kiev, Ukraine, and the Ph.D. degree in biomedical engineering from Monash University, Melbourne, Australia.

She is currently with the School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia. Her particular research interest is in the area of structure-function analysis of proteins and the design of peptide analogous. She is investigating different physicochemical properties of amino acids and their applicability for protein signal analysis within the resonant recognition model.

Qiang Fang (Member, IEEE) received the B.Sc. degree in physics from Tsinghua University, Beijing, China, in 1991 and the Ph.D. degree in biomedical engineering from Monash University, Melbourne, Australia, in 2001.

He is currently a Lecturer at the School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia. His main research interests include protein structure-function analysis, biological data retrieval and mining, and digital signal processing.



Metin Akay (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the Bogazici University, Istanbul, Turkey, in 1981 and 1984, respectively, and the Ph.D. degree from Rutgers University, New Brunswick, NJ, in 1990.

He is currently Associate Professor of Engineering, Psychology and Brain Sciences, and Computer Science at Dartmouth University, Hanover, NJ. He has played a key role in promoting the biomedical education in the world by writing several prestigious books and editing the *IEEE Biomedical Engineering Book Series* (New York: Wiley/IEEE Press), sponsored by the IEEE Engineering in Medicine and Biology Society (EMBS). He is the author or coauthor of 12 books, including *Theory and Design of Biomedical Instruments* (New York: Academic, 1991), *Biomedical Signal Processing* (New York: Academic, 1994), *Detection Estimation of Biomedical Signals* (New York: Academic, 1996), *Time-Frequency and Wavelets in BME* (New York: Wiley/IEEE Press, 1997), *Nonlinear Biomedical Signal Processing* (New York: Wiley/IEEE Press, 2000), *Information Technology in Medicine* (New York: Wiley, 2000). He gave 30 keynote and plenary talks and several invited talks at the international meetings including the ICAP'94, IFSA'95, the DSP applications and Exhibition Conference '96, the Satellite Symposium of the IEEE EMBS'98 in China, the 12th Annual Conference of Japanese Society of Medical Electronics and Biomedical Engineering, and the first and second Latin-American Conference on Biomedical Engineering in 1998 and 2001. His Neural Engineering and Informatics Lab is interested in investigating the respiratory somatosensory (RSS) responses of patients with obstructive sleep apnea syndrome (OSAS) and the effect of developmental abnormalities and maturation on the dynamics of respiration.

Prof. Akay is a recipient of the IEEE EMBS Career Service "for his outstanding contributions to the advancement of the scientific stature and visibility of IEEE-EMBS and extraordinary dedication to the promotion of biomedical engineering education in the world." He is also a recipient of the IEEE Third Millennium Medal for "his contributions to biomedical engineering research and education." He served as the invited Guest Editor for 12 special issues of *IEEE Engineering in Medicine and Biology* magazine, *Annals of BME*, and *Journal of BME* in the areas of cardiovascular engineering, virtual reality in medicine, advances in biomedical signal processing, and fuzzy logic in medicine. He is also the invited Guest Editor for PROCEEDINGS OF THE IEEE (Special Issue on Neural Engineering). He is also the Guest Editor for the two special issues of PROCEEDINGS OF THE IEEE on Functional Genomics, which will be published in 2002. He was the Chair of the IEEE EMBS Summer School in 1997, 2001, and 2002. He was also the Program Chair of the Annual IEEE EMBS Conference in 2001. These activities were sponsored by the National Science Foundation and largely attended by women and minorities. He is a strong supporter of women and minorities in engineering, medicine, and science in the world. He is also the IEEE Distinguished Lecturer in Bioengineering. He received the IEEE Engineering in Medicine and Biology Society Early Career Achievement Award in 1997 "for outstanding contributions in the detection of coronary artery disease, in understanding of early human development, and leadership and contributions in biomedical engineering education." He received the Young Investigator Award of the Sigma Xi Society, Northeast Region in 1998 and 2000 for "his outstanding research activity and the ability to communicate the importance of his research to the general public." He is a Member of Eta Kappa, Sigma Xi, Tau Beta Pi, The American Heart Association, and The New York Academy of Science. He also serves on the advisory board of several international journals including the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, IEEE TRANSACATION ON INFORMATION TECHNOLOGY IN BIOMEDICINE, and the National Institutes of Health Bioengineering partnership study session and several NSF review panels.

Irena Cosic received the B.Eng. degree in electrical engineering in 1976 and the M.Eng. and Ph.D. degrees in biomedical engineering from the University of Belgrade, Belgrade, Yugoslavia, in 1982 and 1985, respectively.

She is currently a Head of the School of Electrical and Computer Systems Engineering at RMIT University, Melbourne, Australia. She has published more than 100 papers, including journal papers and refereed conference papers, book chapters, and a research book. Her major research interests are in the area of biomolecular electronics, the influence of electromagnetic radiation on the human body and tissues, and complementary medicine. The main breakthrough in her research is the invention of the resonant recognition model, an innovative approach to the analysis of proteins and DNA using digital signal-processing methods and physicochemical characteristics of biomolecules. She holds one international patent.