Identification of Hot-Spot Locations in Proteins Using Digital Filters

Parameswaran Ramachandran, *Student Member, IEEE,* and Andreas Antoniou, *Fellow, IEEE*

***Abstract*—A technique for the identification of hot-spot locations in proteins using digital filters is described. In this technique, the characteristic frequency of the protein sequence of interest is first determined from the consensus spectrum of the corresponding functional group. The sequence is then filtered by using a specialized narrowband bandpass digital filter in order to select the characteristic frequency. The energy of the filtered output reveals the hot-spot locations. Zero-phase filtering is used to eliminate the need of computing the phase response of the digital filter. The technique has a unique advantage over existing computational hot-spot location techniques in that it identifies the hot-spot locations solely from the amino-acid sequence of a protein, which is usually the only information initially available for a newly discovered protein molecule. The paper deals also with a MATLAB implementation of the technique that incorporates a user-friendly graphical interface. The technique is illustrated using several protein examples and the results obtained are compared with corresponding results based on biological methodologies in order to demonstrate the usefulness, accuracy, and reliability of the technique.**

***Index Terms*—Proteins, hot spots, characteristic frequency, resonant recognition model (RRM), electron-ion interaction potential (EIIP), FFT, consensus spectrum, digital filters.**

I. Introduction

PROTEINS are the fundamental building blocks of all living organisms and include substances such as enzymes, hormones, and antibodies [1]–[4]. They are complex molecules known as *macromolecules*, and consist of linear chains of subunits known as *amino acids*. The individual amino acids in a protein molecule are linked by covalent linkages called *peptide bonds*. Hence, the chain of amino acids is called a *polypeptide* chain.

An amino acid consists of a carboxylic acid group, an amino group, and a variable side chain, all attached to a central carbon atom (also called the α-Carbon). The side chain is the only component that varies from one amino acid to another. The varying side chains are responsible for the chemical variety of the amino acids. Although numerous different amino acids are theoretically possible, only 20 of them are commonly found in proteins. These 20 amino acids make up the proteins found in all kinds of living organisms. The reason for the specific choice of this set of amino acids can only be attributed to millions of years of evolution.

Protein chains range in length from about 30 amino acids to more than 10,000 amino acids. However, the vast majority of proteins are between 50 and 2000 amino acids long. Although proteins can be conceptually thought to be linear chains of amino acids, in reality they fold in certain unique ways to form complex *three-dimensional (3-D)* structures. By virtue of these structures, proteins perform their biological functions through selective interactions with other molecules known as *targets*. Usually, the target molecules are also proteins, although they can sometimes be of other types such as parts of the DNA molecule.

The order of amino acids in a protein is known as its *primary structure*. It is known that the 3-D structure of a protein molecule is determined by its primary structure. However, the problem of how the 3-D structure and the biological function of a protein are coded into its primary structure is not fully resolved. Solving this problem is very crucial, and when this is achieved it will be possible to construct artificial proteins having desired or prescribed functions by carefully assembling the amino acids. Such custom-made proteins will lead to new cures for diseases such as paralysis and heart ailments. Consequently, solving the protein *structure-function* problem is a very active area of research in which biologists, chemists, computer scientists, and engineers are working in a collaborative effort.

One of the main steps in solving the protein structure-function problem is to understand the protein-target interactions. These interactions are very selective in nature. Specific regions in the protein and target molecules, known as *active sites*, bring about the protein-target interactions. Identifying the locations of the active sites is, therefore, crucial to the understanding of protein-target interactions. After locating the active sites in a protein, the next step is to identify the groups of amino acids that dominate the protein's function. These are known as *hot spots*. A popular experimental technique[1] for the identification of the locations of active sites and hot spots is *site-directed mutagenesis* [5]–[8]. In this technique, the amino acids at specific locations in a naturally-occurring protein are replaced by some other type of amino acid. Such replacements are known as *mutations*. As a result of these mutations, changes in the biological properties of the protein may occur. If a particular amino-acid location is very crucial to the biological function of the protein, then a mutation at that location will considerably hamper the protein's biological function. From this, we can conclude that the particular location belongs to a hot spot. The procedure must then be repeated for every suspected amino-acid location to determine all the hot spots. To perform the mutations, the amino acid *alanine* is usually chosen as the replacement [7]. The procedure is then

This work was funded by the Natural Sciences and Engineering Research Council of Canada.

The authors are with the Department of Electrical and Computer Engineering, University of Victoria, Canada (e-mail: rpara26@ieee.org, aantoniou@ieee.org).

[1] An *experimental technique* is carried out in a wet laboratory as opposed to a *computational technique* that can be carried out by using a digital computer.

called *alanine-scanning mutagenesis (ASM)*. The reason for the choice of alanine is because it has a methyl group (–CH_3) as the side chain making it one of the simplest amino acids with respect to molecular structure. The methyl group of alanine is nonreactive and thus it is almost never directly involved in protein function. Replacement of an amino acid by alanine eliminates the side chain yet it does not alter the main chain conformation nor does it impose extreme electrostatic or steric effects. Furthermore, alanine is abundant in proteins. A limitation of this technique is that it cannot be used if a suspected hot-spot amino acid happens to be alanine. However, this situation may be relatively rare due to the observation that three amino acids, namely, tryptophan, arginine, and tyrosine have a high probability of appearing in hot spots [9].

Although conceptually simple, the ASM procedure involves several delicate steps that need to be flawlessly executed at microscopic levels. Moreover, it is an expensive procedure as it requires the use of specialized chemicals and laboratory apparatus. Therefore, simpler and less expensive computational techniques that can yield estimates of the hot-spot locations will be of immense help to biologists in minimizing unnecessary mutations. By using the estimates obtained, biologists can selectively perform laboratory mutagenesis procedures to confirm the hot-spot locations thereby saving a considerable amount of laboratory resources.

In this paper, we propose a new computational technique for identifying hot-spot locations based on the use of digital filters. The main advantage of this technique over existing computational techniques is that the hot-spot locations are identified solely from the amino-acid sequence of a protein, which is usually the only information initially available for a newly discovered protein molecule. The paper deals also with a MATLAB implementation of the technique that incorporates an easy-to-use graphical user interface (GUI). The technique is illustrated in terms of several protein examples and the results obtained are compared with corresponding results based on biological methodologies.

The paper is organized as follows. Section II describes the fundamentals of hot spots in proteins. Section III describes the resonant recognition model which forms the basis for our technique. In Section IV, the proposed hot-spot location technique is described in detail. Section V describes the MATLAB implementation of the technique. A variety of examples are then presented in Section VI to demonstrate the technique and illustrate its capabilities relative to known biological methodologies. The computational complexity of the technique is discussed in Section VII.

II. Hot Spots in Proteins

The two interacting molecules in a protein-target interaction can either be both proteins or a protein and another type of molecule such as a DNA regulatory segment. When both the molecules are proteins, then the two terms 'protein' and 'target' have relative meanings. Protein 1 is the target for protein 2, and protein 2 is the target for protein 1. Hence, depending on the context, both molecules qualify to be called by either name. However, there is no ambiguity when only one of the molecules involved in an interaction is a protein.

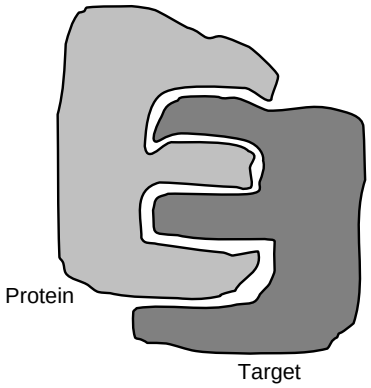


Fig. 1. An illustration of how a protein fits into its target.

The protein-target interactions are very specific in nature. This specificity is achieved by way of the unique 3-D structure of a protein molecule brought about by the folding of its amino-acid chain in a certain manner [10]. The protein-target interactions occur at predefined locations within the 3-D structure of a protein molecule. These locations are known as *active sites* and are typically the regions that make contact in an interaction [1], [11]. The active sites are formed as unique patterns in the arrangement of the amino acids in a protein. The shapes of the active sites are such that they can fit into the target molecules in a way analogous to a hand fitting into a glove, as shown in Figure 1.

A general property of protein-protein interfaces identified by various studies over the past two decades [12]–[16] is that most of the binding energy in an interaction is contributed by a small portion of the total number of amino acids comprising an active site. These few amino acids are termed *hot spots* and are responsible for the stability of the active sites as well as the protein-target complex as a whole [17]–[20]. Due to the crucial role played by hot spots, thorough knowledge about their locations is essential in understanding protein function. Therefore, reliable and efficient techniques for locating hot spots are required.

In concrete terms, a hot spot is defined using a thermodynamic quantity known as *Gibbs free energy*. This is the difference between the internal energy of a system and the product of its absolute temperature and entropy, and denotes a measure of the capacity of the system to do work.[2] It is measured in kilojoules or kilocalories per mole and is denoted as ΔG. The lower the free energy value, the easier it is for the system to do work. If the free energy is negative, the system will have a tendency to do work spontaneously, as in the case of an exothermic chemical reaction. In the context of a protein-target interaction, the work involved is the *binding* of the two molecules, and hence the term used is *binding free energy*. In order to determine whether a given amino acid is a hot spot, it is mutated to alanine and the binding free energy of the mutated protein-target complex is measured. The change in the binding free energy before and after the mutation is denoted as $\Delta\Delta G$. If the amino acid in question is a hot spot, then the

[2]Adopted from the American Heritage Science Dictionary.

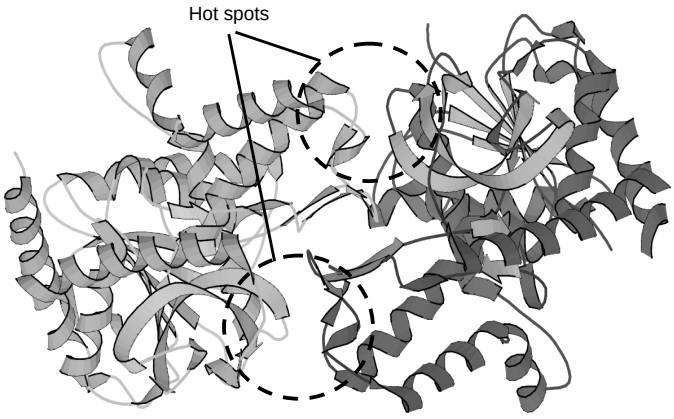


Fig. 2. Three-dimensional structure of a protein, with hot spots. The protein molecule shown is malate dehydrogenase and the interaction is between the chains A (left) and C (right) of the molecule. The circled regions contain the hot spots. Protein structures are generally represented in terms of structural motifs known as α-helix and β-sheet, the coils and sheets in the figure. This structure was obtained from the Protein Data Bank (PDB) at http://www.pdb.org using the protein ID '1guy'.

mutation reduces the binding affinity or, in other words, the binding becomes more difficult after the mutation. A reduction in the binding affinity corresponds to a positive $\Delta\Delta G$. A hot spot is defined as an amino acid whose mutation to alanine leads to a $\Delta\Delta G$ of at least 2.0 kcal/mol. This definition is commonly accepted in the biological community and has been widely used in the past (e.g., in [9], [19]). A schematic of a protein with hot spots is shown in Figure 2.

The hot-spot location technique proposed in this paper is based on a model of protein-target interactions known as the *resonant recognition model* (RRM), which is explained next.

III. The Resonant Recognition Model

Amino acids are represented by assigning distinct alphabets to them. Thus an entire protein sequence can be represented by a corresponding character sequence. In order to apply digital signal processing (DSP) for their analysis, the protein character sequences need to be mapped onto numerical sequences. This can be achieved by finding a set of numerals whose elements can be assigned to the individual amino acids.

Usually, the choice of the numerals is based on some physical property that is relevant to the biological function of the amino acids. A successful attempt to assign numerical values to the amino acids was made in [21] where each amino acid is assigned a numerical value called its *electron-ion interaction potential* (EIIP). The EIIP of an amino acid is a physical property denoting the average energy of the valence electrons in the amino acid, and is known to correlate well with a protein's biological properties [22]. Among over 200 different types of numerical mappings, it has been shown in [22] that EIIP values provide the most suitable mapping for a frequency-based analysis of protein sequences. The use of the EIIP value as a basis for characterizing the protein-target interactions assumes that the strength of the electromagnetic field surrounding a molecule can provide us with a preliminary indication of the molecule's capability to take part in biochemical processes [23].

TABLE I
EIIP Values for the 20 Amino Acids

Amino acid	EIIP	Amino acid	EIIP
Leucine (Leu)	0.0000	Tyrosine (Tyr)	0.0516
Isoleucine (Ile)	0.0000	Tryptophan (Trp)	0.0548
Asparagine (Asn)	0.0036	Glutamine (Gln)	0.0761
Glycine (Gly)	0.0050	Methionine (Met)	0.0823
Valine (Val)	0.0057	Serine (Ser)	0.0829
Glutamic acid (Glu)	0.0058	Cysteine (Cys)	0.0829
Proline (Pro)	0.0198	Threonine (Thr)	0.0941
Histidine (His)	0.0242	Phenylalanine (Phe)	0.0946
Lysine (Lys)	0.0371	Arginine (Arg)	0.0959
Alanine (Ala)	0.0373	Aspartic acid (Asp)	0.1263

An EIIP value exists for each of the 20 amino acids. It can be computed by using formulas based on the general-model pseudopotential as described in [24]. By representing amino acids by their EIIP values, a numerical sequence corresponding to the original character sequence can be obtained. In this manner, every protein character sequence can be transformed into a corresponding numerical sequence. DSP techniques can then be applied to the numerical sequences for detailed analysis. The EIIP values for the 20 amino acids are listed in Table I.

As a first step towards DSP-based analysis, Veljcović and co-workers have subjected the EIIP sequences of several proteins to Fourier spectral analysis [21]. They have observed that the discrete Fourier transforms (DFTs) of the EIIP sequences of the proteins belonging to a particular functional group share a unique spectral component. The frequency of this spectral component characterizes the protein function, and hence, it has been termed the *characteristic frequency* of the functional group. Thus, each protein function can be mapped onto a unique frequency in the frequency domain. Some proteins perform more than one function during their life cycle. For such proteins, each function will correspond to a different characteristic frequency.

For a successful protein-target interaction, both the protein and the target signals must share the same characteristic frequency but must have opposite phase. This matching resembles resonance, and hence, the characteristic frequency is said to provide *resonant recognition* between a protein and its target. This model of the protein-target recognition has been termed the *resonant recognition model* (RRM).

Based on the RRM, we can predict whether a particular protein will interact with an arbitrary target molecule by examining whether or not the protein and target share a common characteristic frequency. For more information on the RRM and its application to various protein sequences, the reader is referred to [25]–[29].

A. Determination of the Characteristic Frequency

The common characteristic frequency of a functional group of M proteins can be determined by computing the cross-spectral function

$$S(e^{j\omega}) = |X_1(e^{j\omega})\, X_2(e^{j\omega})\, \ldots\, X_M(e^{j\omega})| \qquad (1)$$

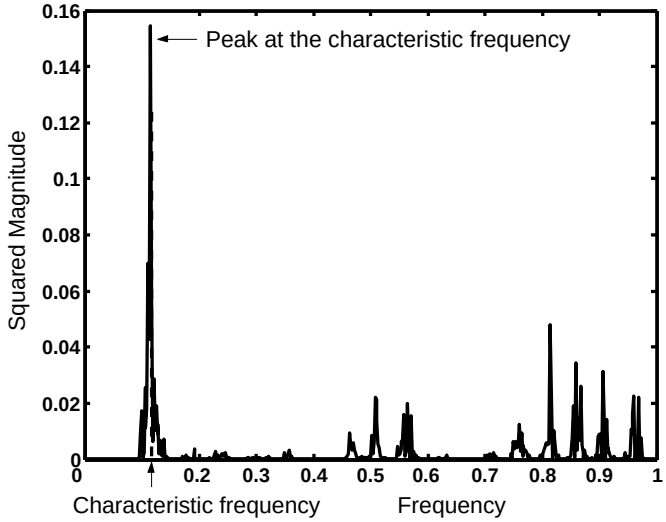


Fig. 3. Consensus spectrum of the epidermal growth factor functional group.

where $X_1, X_2, \ldots, X_M$ are the DFTs corresponding to the M proteins. In simpler terms, Eq. (1) corresponds to the product of the amplitude spectra of the protein sequences belonging to a functional group. Such a product is known as the *consensus spectrum* of the group. The consensus spectrum of a functional group has a distinct peak at the characteristic frequency. The consensus spectrum for a group of proteins called epidermal growth factor (EGF) is shown in Figure 3.

The number of protein sequences, M, required for a typical consensus spectrum varies from case to case. Typically, a sufficient number of protein sequences should be used to achieve a distinct peak at the characteristic frequency in the consensus spectrum. To start with, a set of two protein sequences may be tried. If there is ambiguity (i.e., if there are two or more peaks of approximately the same amplitude), then one more protein sequence from the functional group of interest is included in the computation. This procedure is repeated until the ambiguity is resolved, i.e., there is only one prominent peak with all other peaks well below it, thus clearly identifying the characteristic frequency.

If a protein performs more than one function, then, according to the RRM, each function will correspond to a unique characteristic frequency which can be identified by considering several consensus spectra, one for each function.

B. Hot Spots in Terms of the RRM

The hot-spot locations in a protein or a target molecule can be identified by determining the regions in the corresponding numerical sequence where the characteristic frequency is dominant. This corresponds to localization in the amino-acid domain of a protein numerical sequence, which can be easily achieved by using a variety of DSP techniques. A technique for the location of hot spots based on the use of the short-time discrete Fourier transform (STDFT) was proposed by the authors in [30]. In what follows, we explore the use of digital filters for the location of hot spots. As will be demonstrated, digital filtering is computationally more efficient than the STDFT technique.

IV. Location of Hot-Spots in Proteins Using Digital Filters

The consensus spectrum of the proteins belonging to a functional group reveals the characteristic frequency for the functional group. The hot-spot locations in a particular protein can be identified by first determining the corresponding characteristic frequency and then identifying the regions in the protein numerical sequence where the characteristic frequency is dominant. A simple strategy to identify such regions would be to alter the amplitude of the DFT coefficient corresponding to the characteristic frequency and determine the amino acids that are most affected by this alteration. This strategy is described in [25]. Its disadvantage is that a change in a single DFT coefficient affects all the elements of the original protein numerical sequence and, consequently, a hot-spot location technique based on this strategy is not reliable.

An alternative approach is the transform-based technique introduced in [30]. Although effective in locating hot spots, the technique is computationally expensive. A very efficient alternative based on the use of digital filters was introduced by the authors in [31]. Here we present the complete details of the technique along with several innovations and new results and comparisons with corresponding results based on biological methodologies. The innovations include the introduction of a control parameter that can be used to vary the resolution of the hot-spot location technique according to the user's preference. Varying this control parameter is analogous to varying the threshold value of 2.0 kcal/mol in the mutagenesis-based hot-spot location procedure and it would thus enable a typical user to group the identified hot spots based on their significance in the function of the corresponding protein.

A. The General Idea

The characteristic frequency of a protein numerical sequence obtained by using the RRM uniquely correlates with its biological function. We also know from Section III-B that the hot spots can be located by determining the regions in the protein numerical sequence where the characteristic frequency is dominant. If we can select the characteristic frequency from the large number of insignificant frequencies present in a protein numerical sequence, then it would become easier to locate the hot spots. This selection can be achieved with the aid of a *narrowband bandpass digital filter*. The output of this digital filter will be a sequence having energy only at the characteristic frequency. Therefore, a plot of the energy (squared magnitude) of the output sequence will reveal the locations of the hot spots in the form of distinct energy peaks.

B. Step-By-Step Procedure

A step-by-step procedure of the proposed hot-spot location technique is as follows:

1) Convert several protein sequences belonging to the functional group of interest into numerical sequences using EIIP values.
2) Compute the DFTs of the numerical sequences and their consensus spectrum to determine their characteristic frequency.

3) Design a narrowband bandpass digital filter that would select the characteristic frequency.
4) Filter the protein numerical sequence of interest by using the digital filter designed.
5) Compute the energy of the filter output to determine the regions in the numerical sequence where the characteristic frequency is dominant.
6) Locate the hot spots by locating the energy peaks in the filtered signal based on a suitable peak-to-average ratio.

To achieve a distinct peak at the characteristic frequency in the consensus spectrum, a sufficient number of protein sequences should be used. As discussed in Section III-A, the number of sequences required can vary from a minimum of two to as many as nine or more.

The lengths of protein sequences are usually less than 2^{16} and, consequently, the use of 2^{16}-point DFTs tends to give good results in practice. Evidently, the lengths of proteins need to be adjusted to 2^{16} points and this can be achieved by appending the appropriate number of trailing zeros in each sequence.

C. Choice Between IIR and FIR Filters

Several factors need to be taken into account while choosing the type of the digital filter that can be used for our application in order to assure accurate locations of the hot spots as well as minimal computational effort.

1) Linear Phase Response: In many applications, the choice between a *finite-duration impulse response* (FIR) and an *infinite-duration impulse response* (IIR) digital filter is dependent on whether or not linear phase response is required [32], [33]. This is because linear phase response is easily achieved in FIR filters while it would be somewhat involved to obtain even a near-linear phase response in an IIR filter. If the application at hand is real-time and a linear phase response is a requirement, then an FIR filter must be used. On the other hand, if the application is nonreal-time, then the filter delay can be eliminated altogether by using *zero-phase filtering*, which will be discussed in Section IV-E. In such situations, IIR filters are preferred since they offer several advantages over FIR filters.

2) Low Filter Order and High Selectivity: The two characteristics that are most critical for our application are a low order for the filter transfer function and a high selectivity. The rationales behind these requirements are as follows:

a) Low Filter Order: The higher the order of a digital filter, the longer would be its transient response. A long transient response would make the filtering of the protein sequence inefficient because a significant portion of the sequence would have already passed through the filter by the time steady state is attained. Thus, it is critical to have a low order for the transfer function of the filter.

b) High Selectivity: The frequency spectrum of a protein numerical sequence consists of a number of other frequency components in addition to the characteristic frequency component. Our aim is to select only the characteristic frequency component while attenuating all the other frequency components to an insignificant level. As it is possible to have unwanted frequencies very close to the characteristic frequency in the frequency spectrum, the digital filter must have a high selectivity (i.e., narrow transition bands).

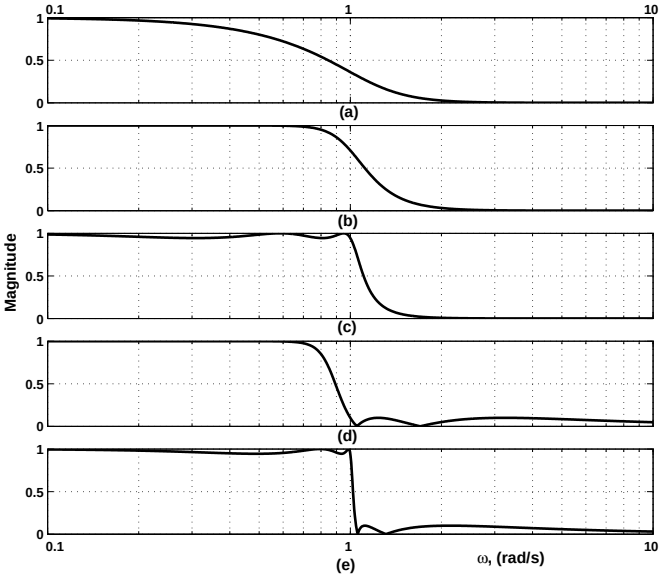


Fig. 4. Amplitude response of a fifth-order normalized lowpass (a) Bessel-Thomson filter, (b) Butterworth filter, (c) Chebyshev filter, (d) inverse-Chebyshev filter, and (e) elliptic filter.

Both of the above requirements can be simultaneously satisfied by using an IIR filter. This is because, in an IIR filter, the poles of the transfer function can be placed close to the unit circle. Hence a high selectivity can easily be achieved with a low-order transfer function. In an FIR filter, on the other hand, with the poles fixed at the origin, high selectivity can be achieved only by using a relatively high order for the transfer function.

Choosing an IIR digital filter for our application offers another advantage. Due to the much lower order of the IIR filter, the filtering would involve only a small amount of computation, which would lead to an efficient implementation of the hot-spot location system.

D. Choice Among Different Types of IIR Filters

IIR digital filters are classified on the basis of the analog-filter approximations from which they are derived. The most frequently used analog-filter approximations are the Bessel-Thomson, Butterworth, Chebyshev, inverse-Chebyshev, and the elliptic approximations. A digital filter derived from a particular approximation inherits the characteristics of the approximation. The narrowband bandpass digital filter required for our application can be designed in two steps. First a normalized lowpass transfer function is transformed into a denormalized bandpass transfer function by employing the standard lowpass-to-bandpass analog-filter transformation. Then the bilinear transformation is applied to obtain the transfer function of the digital filter (see Chap. 12 in [32]).

The Bessel-Thomson and Butterworth filters are not suitable for our application as they do not offer good selectivity when compared with the other three filter types. Among the other

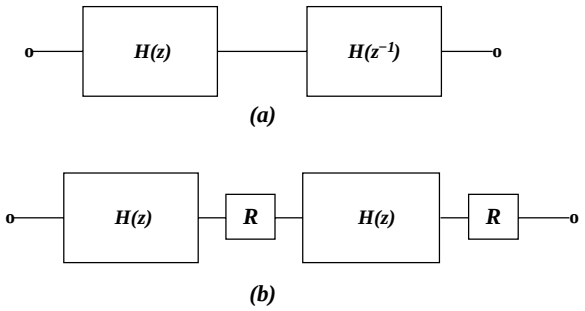


Fig. 5. (a) Zero-phase filter, (b) implementation.

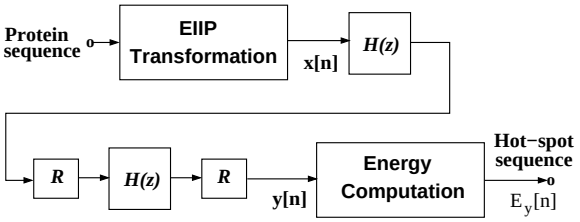


Fig. 6. Complete digital-filter based hot-spot location system.

three types, the Chebyshev and elliptic filters exhibit a ripple in their passband amplitude response, which would introduce increased amplitude distortion in the band of frequencies close to a characteristic frequency thereby resulting in errors in locating the hot spots. For this reason, Chebyshev and elliptic filters are unsuitable for this application. Inverse-Chebyshev filters have good selectivity and require similar filter orders as Chebyshev filters to satisfy a given set of specifications; in addition, they have a monotonic passband amplitude response which is highly desirable for the present application. For these reasons, the proposed hot-spot location system was implemented using inverse-Chebyshev filters.

E. Zero-Phase Filtering

At the filter output, the passband frequency component will be delayed by a certain period of time. Calculation of this delay is essential in order to collect the appropriate output samples. Unfortunately, this calculation is not very straightforward for an IIR filter. A simpler solution is to eliminate the filter delay by using *zero-phase filtering*.

In zero-phase filtering (see Sec. 12.5 in [32]), the signal is filtered through a cascade arrangement of two filters characterized by $H(z)$ and $H(z^{-1})$, as depicted in Figure 5a. The frequency response of the cascade arrangement can be expressed as

$$H_0(e^{j\omega T}) = H(e^{j\omega T})\, H(e^{-j\omega T}) = |H(e^{j\omega T})|^2 \qquad (2)$$

In other words, the frequency response of the arrangement is real and, as a result, has *zero* phase response. If the impulse response of the first filter is $h(n)$, then that of the second filter is $h(-n)$. The cascade of Figure 5a is implemented as depicted in Figure 5b where the protein numerical sequence is first fed to the filter characterized by $H(z)$ and the resulting output is then reversed and fed to the same filter again. The output of the second filtering operation is then reversed to obtain the final output. Devices R in Figure 5b, which are actually first-in last-out registers, are used to reverse the signals at the input and output of the second filtering operation. The delay introduced by the first filtering operation is canceled by the second filtering operation since the signal is fed backwards the second time. Thus, upon zero-phase filtering, the characteristic frequency component is not delayed at the output, and the need to compute the phase response of the IIR filter is eliminated. The locations of the peaks in the energy of the output signal identify the locations of the hot spots. If the output signal is denoted as $y[n]$, then its energy is given by

$$E_y[n] = |y[n]|^2 \qquad (3)$$

We refer to $E_y[n]$ as the *energy sequence* corresponding to a protein and to the peaks in $E_y[n]$ as the *energy peaks*.

The complete hot-spot location system is depicted in Figure 6.

F. Determination of the Filter-Design Specifications

From our discussion so far, we have concluded that an inverse-Chebyshev narrowband bandpass IIR digital filter is the best choice for locating the hot spots in proteins based on the RRM, and that zero-phase filtering must be used in order to eliminate the need to compute the phase response of the filter. We now discuss the factors determining the design specifications of the filter. The specifications that need to be determined are the selectivity, the maximum passband attenuation, the minimum stopband attenuation, and the filter order.

1) Selectivity: Selectivity depends on the locations of the passband and the stopband edges. The closer the locations of the stopband edges to the locations of the corresponding passband edges the higher would be the selectivity. The band edges of the digital filter can be determined from the frequency spectrum of the protein sequence of interest. For example, consider the frequency spectrum of the tuna cytochrome C protein sequence shown in Figure 7. The neighborhood of the characteristic frequency is shown as an inset. The value of the characteristic frequency is 0.944. As we need to pass the characteristic frequency through the filter, the maximum passband gain of 0 dB will correspond to the characteristic frequency. We can then determine the two passband edges ω_{p1} and ω_{p2} as the frequencies whose amplitudes are 3 dB below that of the characteristic frequency. Thus the difference, $\omega_{p2} - \omega_{p1}$, will constitute the 3-dB bandwidth of the filter. The stopband edges ω_{a1} and ω_{a2} should be such that the frequencies closest to the characteristic frequency on either side of the characteristic frequency in the frequency spectrum would be heavily attenuated. As long as this condition is satisfied, the stopband edges are allowed to be at the furthest possible location from the respective passband edges. This would be worthwhile, as it would then prevent the selectivity requirement from being oversatisfied, otherwise, the order of the filter would be unnecessarily increased thus deviating from an optimal design. The values of ω_{a1}, ω_{p1}, ω_{p2}, and ω_{a2} determined for the tuna cytochrome C protein are 0.928, 0.936, 0.952, and 0.960, respectively.

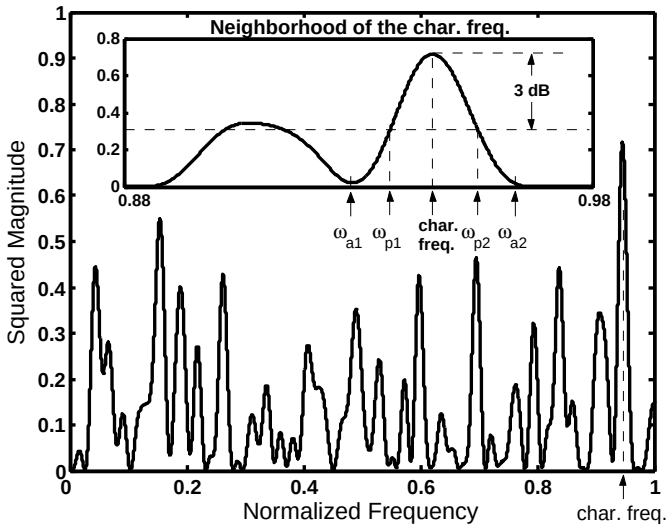


Fig. 7. Determination of the filter-design specifications from the frequency spectrum of a protein numerical sequence. (This illustration corresponds to the case of the tuna cytochrome C protein).

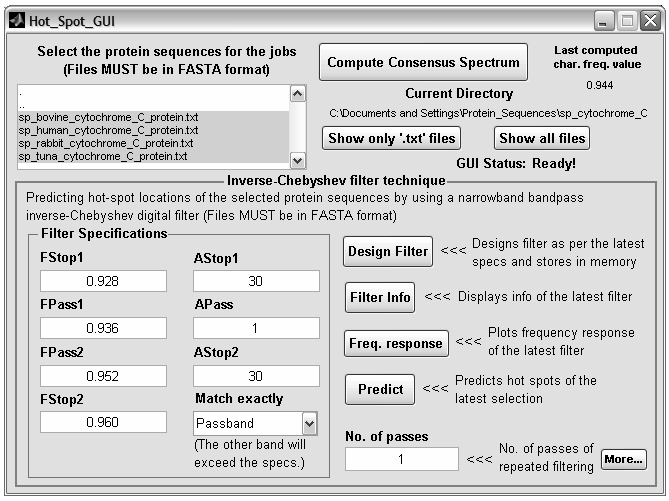


Fig. 8. A screen shot of the hot-spot GUI.

2) Passband and Stopband Attenuation: Since our filter is of the inverse-Chebyshev type, there is no passband ripple in the amplitude response. However, for the design process, we need to specify a value for the maximum passband attenuation, which occurs at the passband edges for an inverse-Chebyshev filter. A 10% reduction in the amplitudes of the frequencies at the passband edges is considered acceptable as this reduction will not affect the characteristic frequency. This corresponds to a maximum passband attenuation of 1 dB. The minimum stopband attenuation depends on the extent to which the stopband frequencies need to be attenuated. Here we consider a uniform reduction in the amplitudes of the stopband frequencies to a level below 5% of the amplitude of the characteristic frequency to be sufficient for efficiently locating the hot spots. This corresponds to a minimum stopband attenuation of approximately 30 dB.

3) Filter Order: The minimum filter order that would satisfy all the specifications can be determined by using a standard technique described in Chap. 12 of [32].

Mostly, the frequency spectra corresponding to the proteins belonging to a particular functional group are very similar, at least in the neighborhood of the characteristic frequency. Hence, in many cases, it is possible to use the same digital filter for locating the hot spots of different proteins as long as the proteins belong to the same functional group. However, if the functional group is different then the characteristic frequency is also different, which would alter the filter specifications.

Note that the design procedure is totally deterministic and it can, therefore, be fully automated whereby the IIR inverse-Chebyshev filter is designed by using the parameters of the frequency spectrum of a protein as input.

V. The Hot-Spot Software Package

In order to facilitate the application of the filter-based hot-spot location technique, we have implemented the technique using MATLAB and have also created a GUI to achieve a user-friendly implementation. A screen shot of the GUI is shown in Figure 8. The GUI is organized into two main sections. The top portion consists of the listbox that can be used to navigate to the desired folder and select the files of the protein sequences to be analyzed. The bottom portion consists of the fields where the various filter design parameters can be entered by the user. The basic operation of the software consists of the following steps. First, the user navigates to the folder containing the desired protein sequences belonging to the functional group of interest, selects them, and clicks the button labeled 'Compute Consensus Spectrum'. The software will compute the consensus spectrum and display it in a separate figure window. It will also automatically detect any distinct peak in the consensus spectrum and thereby determine the characteristic frequency. After this, the software will assign default values to all the design parameters. The user can keep these values or alter them as desired. Then the user clicks on the button labeled 'Design Filter'. Once the filter is designed and stored in memory, the user selects from the listbox the protein sequences whose hot-spot locations need to be computed and clicks on the button labeled 'Predict'. This will yield the hot-spot locations for all the selected protein sequences in separate figure windows. The software can also produce the amplitude-response plot of the designed filter or perform repeated filtering on the click of the appropriate buttons. The particular instance shown in Figure 8 corresponds to the stage where the consensus spectrum has just been computed for the cytochrome C functional group and the default values for the filter parameters have been assigned.

As can be seen, the software combined with the GUI is an easy-to-use and efficient implementation of the filter-based hot-spot location technique.[3] By using the software, biologists can efficiently perform numerous trials on their protein sequences in order to obtain estimates of the hot-spot locations before conducting wet laboratory experiments.

[3]The MATLAB source code can be obtained by contacting the authors.

VI. Illustrative Examples and Results

In order to illustrate the effectiveness of the proposed technique, we applied it to a set of protein sequences obtained from standard online databases. In this section, we present the results obtained and also compare them with corresponding results reported by the biological community in order to evaluate the usefulness, accuracy, and reliability of the technique.

A. Online Databases

Protein sequence data in the form of strings of alphabets with each alphabet representing an amino acid, are freely available at various Web databases. The most important databases of this class are the protein data bank (PDB) [34], [35] and Swiss-Prot [36], [37]. The PDB focuses on detailed 3-D structural information of proteins in the form of atomic coordinates, while the Swiss-Prot focuses on details about the amino acid sequences. Both these databases are considered to be very reliable by the biological community and are updated as and when new sequence information becomes available. All the protein sequences for our examples have been obtained from these databases.

Protein hot-spot location data obtained through alanine-scanning mutagenesis (ASM) have been compiled into an online database named the alanine scanning energetics database (ASEdb) [38], [39]. This is a standard repository for hot-spot location data used by the biological community and is updated as and when new data become available. We used this database as a benchmark to evaluate all the hot-spot locations identified by our technique.

B. Preliminaries

We applied our technique to a set of ten example protein sequences each chosen based on two factors. First, the proteins must be very different from each other in terms of biological functionality and, second, the hot-spot location results of the proteins obtained using biological methodologies must be available in ASEdb. This is to ensure that our results can be compared with corresponding results obtained using biological methodologies. Details pertaining to the protein examples are given in Table II. The protein sequences themselves can be downloaded from the databases by using their respective IDs. Proteins hGH and hGHbp bind to each other and hence have the same characteristic frequency as seen in Table II. This is consistent with the RRM which states that a protein and its target must share the same characteristic frequency (see Section III). Consequently, the same digital filter can be used to locate their hot-spots. The same applies to the proteins barnase and barstar. For each of the examples, a narrowband bandpass inverse-Chebyshev IIR digital filter was used with its passband centered at the corresponding characteristic frequency. For the sake of convenience, a sampling frequency of $\omega_s = 2$ was assumed which corresponds to a Nyquist frequency of $\omega_s/2 = 1$. The filter order used in each case was the lowest order for an inverse-Chebyshev filter that satisfied the set of specifications, and was computed by using the design formulas given in [32]. A maximum passband attenuation of 1 dB and a minimum stopband attenuation of approximately 30 dB were assumed for designing the filters. The passband and the stopband edges for the filters used are listed in Table III. A sample amplitude response plot corresponding to the filter used for the human basic fibroblast growth factor protein (bFGF) is shown in Figure 9.

C. Removing the DC Bias

Since the EIIP values assigned to the protein characters are all nonnegative (see Table I), the resulting protein numerical sequence is superimposed on a *DC bias*, i.e., it has a nonzero average value. Consequently, its DFT will show a tall peak at zero frequency, which corresponds to the DC bias and is not part of the actual signal. This peak can obscure other peaks including that of the characteristic frequency, and thus the DFT of the numerical sequence can be misleading. This problem can be avoided by subtracting the average of the numerical sequence from its samples before computing its DFT. This was performed for all the numerical sequences used to compute the consensus spectra of the examples.

D. Peak-To-Average Threshold Parameter

Our intention in identifying the hot-spot locations is to identify the regions in the protein numerical sequence where the characteristic frequency is dominant. The degree of dominance of the characteristic frequency in a particular region of the protein sequence can be estimated by computing the energy of the characteristic frequency component in that region and then comparing it with a reference energy level. The energy of the characteristic frequency component at various locations in the protein sequence can be obtained from the squared magnitude of the output sequence of the digital filter (see Eq. (3)). This is referred to as the *energy sequence* and its peaks as the *energy peaks*. The reference energy level is taken to be the average energy value for the particular protein sequence under consideration. Thus, in order to estimate the degree of dominance of an energy peak, we can first compute the average value of the energy sequence and then divide the magnitude of the energy peak by the average value. This ratio is referred to as the *peak-to-average ratio*.

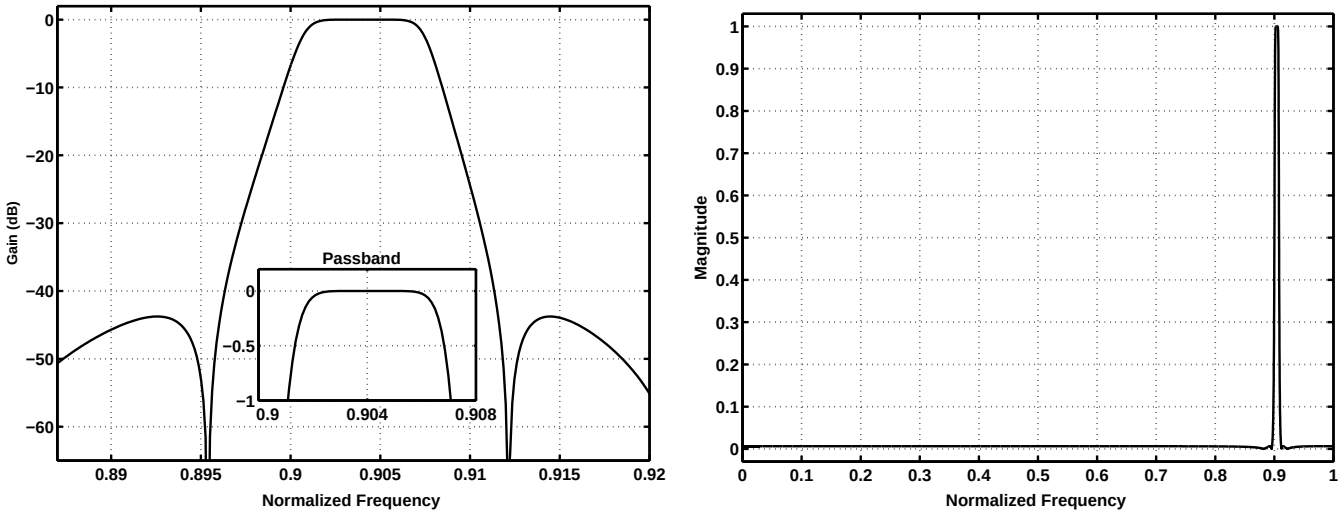
The peak-to-average ratio of an energy peak can be interpreted as an estimate of the significance of the corresponding hot spot in the functioning of the protein. This is one way of using the peak-to-average ratio. Alternatively, it is possible to define a threshold value for the peak-to-average ratio denoted as t_p, then compare the peak-to-average ratio of each energy peak with the value of t_p and designate only those peaks with peak-to-average ratios above the value of t_p as hot-spot locations. If t_p is assigned the value 1, then it means that the threshold is exactly at the average value, i.e., all peaks above the average value are designated as hot-spot locations. If t_p is assigned the value 3, then it means that the threshold is at 3 times the average value, i.e., all peaks above 3 times the average value are designated as hot-spot locations, and so forth.

Initially, a default value of 1 may be assumed for t_p, after which the value may be raised or lowered depending on the

TABLE II
PROTEIN EXAMPLES INVESTIGATED

Example number	Organism	Protein name	PDB ID	Swiss-Prot ID	Sequence length	Characteristic frequency	Filter order
1	human	basic fibroblast growth factor (bFGF)	4fgf	P09038	146	0.904	8
2	human	growth hormone (hGH)	3hhr	P01241	190	0.270	6
3	human	growth hormone binding protein (hGHbp)	3hhr	P10912	203	0.270	6
4	bacteria	barnase	1brs	P00648	110	0.321	6
5	bacteria	barstar	1brs	P11540	89	0.321	6
6	human	interleukin-4 (IL-4)	1rcb	P05112	129	0.587	6
7	*E. Coli*†	colicin-E9 immunity protein (IM9)	1bxi	P13479	86	0.190	6
8	human	neurotrophin-3 (NT3)	1nt3	P20783	119	0.069	8
9	bacteria	tryptophan RNA-binding attenuator protein (TRAP)	1wap	P19466	75	0.247	6
10	*C. fimi*	endoglucanase C	1ulo	P14090	152	0.093	8

† It is a common practice to italicize the scientific names of living organisms.



(a) (b)

Fig. 9. Amplitude response of the filter used to identify the hot-spot locations in the human basic fibroblast growth factor protein (example 1): (a) decibel (dB) scale, (b) linear scale.

TABLE III
PASSBAND AND STOPBAND EDGES FOR THE EXAMPLES

Example number	Stopband edges		Passband edges	
	ω_{a1}	ω_{a2}	ω_{p1}	ω_{p2}
1	0.896	0.912	0.901	0.907
2	0.262	0.278	0.267	0.273
3	0.262	0.278	0.267	0.273
4	0.313	0.329	0.318	0.324
5	0.313	0.329	0.318	0.324
6	0.579	0.595	0.584	0.590
7	0.182	0.198	0.187	0.193
8	0.061	0.077	0.066	0.072
9	0.239	0.255	0.244	0.250
10	0.085	0.101	0.090	0.096

specific case being investigated. A biologist using the proposed technique may use a t_p value larger than 1 to focus on the main hot spots or reduce the value below 1 to increase the resolution of the technique and thus identify the less significant hot spots. In the extreme case, if the user wishes to consider all the peaks revealed by our technique for further analysis, then t_p can be set to zero.

Since t_p serves as a threshold value for the peak-to-average ratio that can be used as a control parameter to control the resolution of the hot-spot location technique, we refer to t_p as the *peak-to-average threshold parameter*.

E. Advantages of Computational Hot-Spot Location Techniques Over Alanine-Scanning Mutagenesis

Although biological experimentation techniques such as alanine-scanning mutagenesis (ASM) will have to be ultimately employed to conclusively determine hot-spot locations, computational techniques such as the one proposed here can be very useful as a first line of attack thus aiding biologists to selectively focus on specific probable hot spots before they perform wet lab experiments. In analyzing newly discovered proteins, biologists can first apply computational techniques, consider the results obtained as a starting point, and then further analyze only those locations using the experimental techniques to confirm the hot-spot locations. In this way,

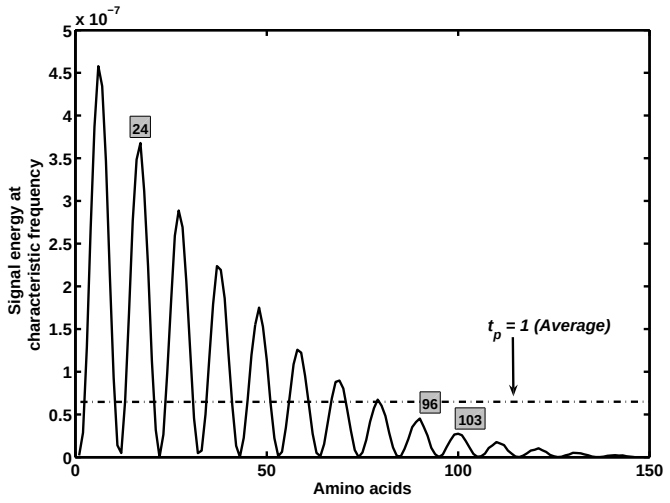


Fig. 10. Hot-spot locations of human basic fibroblast growth factor protein (example 1).

significant amounts of time, effort, and biological resources can be saved. On the basis of our examples, on the average, only a third of the amino acids in a protein sequence belong to hot-spot locations. Thus if computational techniques are made very reliable, then biologists would need to perform mutations for only a third of the total number of amino acids in a protein sequence.

Most of the existing computational techniques require detailed information about the 3-D structure of the protein being investigated [18], [19], [40]. However, for newly discovered proteins, only the amino-acid sequence information is initially available. The 3-D structural information becomes available only after researchers meticulously perform detailed structural analysis on the proteins such as *nuclear magnetic resonance* (NMR) *spectroscopy*. This is very laborious and takes a considerable amount of time, usually several years of further research from the time a protein has been discovered. Hence computational techniques capable of predicting hot-spot locations solely based on a protein's amino-acid sequence, such as the one proposed here, are very useful in analyzing newly discovered proteins.

F. Discussion of the Results

A sample plot illustrating the hot-spot locations corresponding to the human basic fibroblast growth factor (bFGF) protein is shown in Figure 10. The threshold level at the average value is marked. If we set $t_p = 1$, then only the peaks that are above the average threshold level will be designated as hot-spot locations. However, for the sake of completeness, we designate all the peaks as *probable* hot-spot locations, i.e., t_p is set to zero.

In Table IV, we list the probable hot-spot locations identified by our technique for all the ten example protein sequences. Also listed are the hot-spot locations reported in the biological community, identified by performing alanine-scanning mutagenesis. We obtained this data from ASEdb [38], [39]. The effectiveness of our technique can be evaluated by comparing the probable hot-spot locations obtained by using our technique with corresponding hot-spot locations reported in ASEdb. From Table IV, we can see that for the examples considered the filter-based technique correctly identified most of the hot-spot locations reported in ASEdb. Out of a total of 76 locations, our technique identified 61 locations, corresponding to a success rate of more than 80%. The filter-based technique also identified several amino-acid locations which could be new hot spots that have not yet been reported in the database. These are listed in Table V. Owing to the fact that the ASEdb is dynamic and that the hot-spot identification problem is far from being resolved, we believe that the new hot-spot locations predicted by our technique will match entries that may be added to the ASEdb by the biological community in the future. These new locations may provide new insights that may significantly improve the current understanding of the working of proteins.

In the immediate future, we plan to apply the technique to a more diverse range of protein samples, to improve the quality of the bandpass filter further by increasing its selectivity while reducing the duration of its transient response, and to compare the method with any other computational methods that may be proposed in the literature. We also hope to investigate the existence of a certain periodicity that was expected to manifest itself in our results by one of the reviewers but which did not reveal itself in our simulations.

VII. Computational Complexity

From an implementation perspective, the major advantage of the proposed filter-based technique over the transform-based technique introduced in [30] is its lower computational complexity. In order to establish this fact, we compared the computational complexity of the filter-based technique with that of the transform-based technique by recording the average CPU times over 1000 runs applied for each example sequence. The MATLAB commands `tic` and `toc` were used to compute the CPU time. In order to obtain a fair comparison, the same computer system was used throughout the procedure with its processor entirely dedicated to this task. The results are shown in Table VI. As can be seen from the table, the filter-based technique requires less than 15% of the computation required by the transform-based technique, i.e., the filter-based technique is significantly more efficient than the transform-based technique.

VIII. Conclusion

A new technique for the identification of hot-spot locations in proteins using digital filters was presented. The technique was applied to several example protein sequences and the results were compared with corresponding results obtained by the biological community. It was found that the technique correctly identifies more than 80% of the hot-spot locations surveyed. The technique also reveals several amino-acid locations that could be new hot spots.

The technique is particularly suitable for analyzing newly discovered proteins owing to its capability to predict hot-spot locations solely from the amino-acid sequence. However, incorporating the 3-D structural information into the technique

TABLE IV
HOT-SPOT LOCATIONS IDENTIFIED BY THE FILTER-BASED TECHNIQUE ALONG WITH THOSE FROM ASEDB

Example Number	Protein name	Hot-spot locations	
		Filter-based technique	ASEdb
1	human basic fibroblast growth factor (bFGF)	24, 96, 103	24, 96, 103, 140
2	human growth hormone (hGH)	14, 22, 26, 41, 45, 48, 56, 62, 63, 64, 68, 164, 167, 168, 171, 175, 178, 179, 183, 186	14, 21, 22, 26, 41, 45, 46, 48, 56, 61, 62, 63, 64, 68, 164, 167, 168, 171, 172, 175, 176, 178, 179, 183, 186
3	human growth hormone binding protein (hGHbp)	43, 44, 105, 106, 127, 164, 165, 169	43, 44, 103, 104, 105, 106, 126, 127, 164, 165, 169
4	bacteria barnase	27, 54, 59, 60, 73, 83, 87, 102, 103	27, 54, 59, 60, 73, 82, 83, 87, 102, 103
5	bacteria barstar	33, 35, 38, 42, 73, 76, 80	33, 35, 38, 39, 42, 73, 76, 80
6	human interleukin-4 (IL-4)	9, 88	9, 88
7	*E. Coli* colicin-E9 immunity protein (IM9)	34, 41, 50, 51, 55	33, 34, 41, 50, 51, 55
8	human neurotrophin-3 (NT3)	11, 68, 103	11, 68, 103
9	bacteria tryptophan RNA-binding attenuator protein (TRAP)	37, 40, 56	37, 40, 56, 58
10	*C. fimi* endoglucanase C	50	19, 50, 84

TABLE V
NEW HOT-SPOT LOCATIONS IDENTIFIED BY THE FILTER-BASED TECHNIQUE

Example number	Protein name	New locations
1	bFGF	6, 27, 37, 48, 58, 69, 79, 110, 120, 131
2	hGH	4, 7, 11, 19, 30, 33, 37, 52, 59, 70, 74, 78, 81, 85, 89, 93, 96, 100, 104, 107, 111, 115, 118, 122, 126, 130, 133, 137, 141, 144, 148, 152, 156, 159, 189
3	hGHbp	5, 9, 12, 16, 20, 23, 27, 31, 34, 38, 49, 53, 57, 60, 64, 68, 71, 75, 79, 82, 86, 90, 93, 97, 101, 112, 116, 119, 123, 130, 134, 138, 141, 145, 149, 153, 156, 160, 175, 178, 182, 186, 189, 193, 197
4	barnase	3, 6, 9, 12, 15, 18, 21, 25, 31, 34, 37, 40, 43, 46, 49, 56, 65, 68, 71, 77, 80, 90, 93, 96, 99, 105, 108
5	barstar	4, 7, 10, 13, 16, 19, 22, 25, 28, 44, 47, 50, 53, 56, 60, 63, 66, 69, 78, 84, 87
6	IL-4	3, 6, 11, 13, 15, 18, 20, 23, 25, 28, 30, 32, 35, 37, 40, 42, 45, 47, 49, 52, 54, 57, 59, 61, 64, 66, 69, 71, 74, 76, 78, 81, 83, 86, 91, 93, 95, 98, 100, 103, 105, 108, 110
7	IM9	4, 9, 14, 19, 25, 30, 46, 62, 67, 72, 77, 83
8	NT3	7, 9, 22, 35, 37, 50, 79, 93, 108
9	TRAP	3, 7, 11, 15, 19, 23, 27, 31, 35, 39, 44, 48, 52, 60, 64, 68, 72
10	endoglucanase C	4, 14, 26, 36, 57, 68, 79, 90, 101, 111

could significantly improve its accuracy and reliability. If this is achieved, then the technique can first be used to predict probable hot-spot locations of a newly discovered protein solely from its amino-acid sequence; subsequently, when 3-D structural information becomes available, the technique can be reapplied taking the available structural information into account to yield more accurate predictions. The problem of incorporating 3-D structural information into the technique opens up a significant scope for further research.

The paper dealt, in addition, with a MATLAB implementation of the technique that incorporates a user-friendly graphical interface.

REFERENCES

[1] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology*. New York: Garland Publishing, 1998.

[2] A. Datta and E. R. Dougherty, *Introduction to Genomic Signal Processing with Control*. Florida: CRC Press, 2007.

[3] D. Dressler and H. Potter, *Discovering Enzymes*. New York: Scientific American Library, 1991.

[4] C. Tanford and J. Reynolds, *Nature's Robots: A History of Proteins*. Oxford; New York: Oxford University Press, 2001.

[5] R. W. Old and S. B. Primrose, *Principles of Gene Manipulation: An Introduction to Genetic Engineering*, 4th ed. Oxford: Blackwell Scientific Publications, 1989.

[6] B. C. Cunningham, P. Jhurani, P. Ng, and J. A. Wells, "Receptor and antibody epitopes in human growth hormone identified by homolog-scanning mutagenesis," *Science*, vol. 243, no. 4896, pp. 1330–1336, Mar. 1989.

[7] B. C. Cunningham and J. A. Wells, "High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis," *Science*, vol. 244, no. 4908, pp. 1081–1085, Jun. 1989.

[8] J. A. Wells, "Systematic mutational analyses of protein-protein interfaces," *Methods in Enzymology*, vol. 202, pp. 390–411, 1991.

[9] A. A. Bogan and K. S. Thorn, "Anatomy of hot spots in protein interfaces," *Journal of Molecular Biology*, vol. 280, pp. 1–9, 1998.

[10] A. R. Dinner, A. Sali, L. J. Smith, C. M. Dobson, and M. Karplus,

TABLE VI
AVERAGE CPU TIMES

Protein name	Average CPU time (milliseconds)	
	Filter-based technique	Transform-based technique
bFGF	0.3937	3.7643
hGH	0.3110	4.6561
hGHbp	0.3145	4.8868
barnase	0.2749	3.0763
barstar	0.2693	2.6193
IL-4	0.2840	3.4171
IM9	0.2670	2.5612
NT3	0.2918	3.2255
TRAP	0.2630	2.3231
endoglucanase C	0.2956	3.9029

"Understanding protein folding via free-energy surfaces from theory and experiment," *Trends in Biochemical Sciences (TiBS)*, vol. 25, pp. 331–339, Jul. 2000.

[11] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 4th ed. New York: Garland Science, 2002.

[12] T. Clackson and J. A. Wells, "A hot spot of binding energy in a hormone-receptor interface," *Science*, vol. 267, no. 5196, pp. 383–386, Jan. 1995.

[13] T. Kortemme and D. Baker, "A simple physical model for binding energy hot spots in protein-protein complexes," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 99, no. 22, pp. 14 116–14 121, Oct. 2002.

[14] B. C. Cunningham and J. A. Wells, "Comparison of a structural and a functional epitope," *Journal of Molecular Biology*, vol. 234, pp. 554–563, 1993.

[15] I. Halperin, H. Wolfson, and R. Nussinov, "Protein-protein interactions: Coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking," *Structure*, vol. 12, no. 6, pp. 1027–1038, Jun. 2004.

[16] O. Keskin, B. Ma, and R. Nussinov, "Hot regions in protein-protein interactions: The organization and contribution of structurally conserved hot spot residues," *Journal of Molecular Biology*, vol. 345, no. 5, pp. 1281–1294, 2005.

[17] W. L. DeLano, "Unraveling hot spots in binding interfaces: Progress and challenges," *Current Opinion in Structural Biology*, vol. 12, no. 1, pp. 14–20, 2002.

[18] Y. Gao, R. Wang, and L. Lai, "Structure-based method for analyzing protein-protein interfaces," *Journal of Molecular Modeling*, vol. 10, no. 1, pp. 44–54, Feb. 2004.

[19] X. Li, O. Keskin, B. Ma, R. Nussinov, and J. Liang, "Protein-protein interactions: Hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: Implications for docking," *Journal of Molecular Biology*, vol. 344, no. 3, pp. 781–795, 2004.

[20] M. R. Arkin and J. A. Wells, "Small-molecule inhibitors of protein-protein interactions: Progressing towards the dream," *Nature Reviews Drug Discovery*, vol. 3, no. 4, pp. 301 –317, Apr. 2004.

[21] V. Veljković, I. Cosić, B. Dimitrijević, and D. Lalović, "Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?" *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 5, pp. 337–341, May 1985.

[22] J. Lazović, "Selection of amino acid parameters for Fourier transform-based analysis of proteins," *Computer Applications in the Biosciences (CABIOS)*, vol. 12, no. 6, pp. 553–562, 1996.

[23] V. Veljković, *A Theoretical Approach to the Preselection of Carcinogens and Chemical Carcinogenesis*. New York: Gordon and Breach, 1980.

[24] V. Veljković and I. Slavić, "Simple general-model pseudopotential," *Physical Review Letters*, vol. 29, no. 2, pp. 105–107, Jul. 1972.

[25] I. Cosic, "Macromolecular bioactivity: Is it resonant interaction between macromolecules?—Theory and applications," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 12, pp. 1101–1114, Dec. 1994.

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

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

[28] I. Cosic, E. Pirogova, and M. Akay, "Application of the resonant recognition model to analysis of interaction between viral and tumor suppressor proteins," in *Proc. 25^{th} Annual International Conference of the IEEE EMBS*, Cancun, Mexico, Sep.17-21, 2003, pp. 2398–2401.

[29] E. Pirogova, G. P. Simon, and I. Cosic, "Investigation of the applicability of dielectric relaxation properties of amino acid solutions within the resonant recognition model," *IEEE Transactions on Nanobioscience*, vol. 2, no. 2, pp. 63–69, Jun. 2003.

[30] P. Ramachandran, A. Antoniou, and P. P. Vaidyanathan, "Identification and location of hot spots in proteins using the short-time discrete Fourier transform," in *Thirty-Eighth Asilomar Conference on Signals, Systems, and Computers*, Pacific Groove, CA, Nov. 2004, pp. 1656–1660.

[31] P. Ramachandran and A. Antoniou, "Localization of hot spots in proteins using digital filters," in *IEEE International Symposium on Signal Processing and Information Technology*, Vancouver, Canada, Aug. 2006, pp. 926–931.

[32] A. Antoniou, *Digital Signal Processing: Signals, Systems, and Filters*. New York: McGraw-Hill, 2005.

[33] R. G. Lyons, *Understanding Digital Signal Processing*, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, 2004.

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

[35] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

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

[37] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Research*, vol. 31, no. 1, pp. 365–370, 2003.

[38] Alanine Scanning Energetics database (ASEdb). [Online]. Available: http://nic.ucsf.edu/asedb/index.php

[39] K. S. Thorn and A. A. Bogan, "ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions," *Bioinformatics*, vol. 17, no. 3, pp. 284–285, 2001.

[40] S. Huo, I. Massova, and P. A. Kollman, "Computational alanine scanning of the 1:1 human growth hormone-receptor complex," *Journal of Computational Chemistry*, vol. 23, no. 1, pp. 15–27, 2002.