Available online at www.sciencedirect.com
ScienceDirect
Signal Processing 88 (2008) 877–890
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

Macromolecular sequence analysis using multiwindow Gabor representations

Nagesh K. Subbanna*, Yehoshua Y. Zeevi

Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, Israel

Received 19 August 2006; received in revised form 6 August 2007; accepted 5 October 2007
Available online 22 October 2007

Abstract

Multiwindow Gabor representations highlight fingerprints suitable for indexing of macromolecules, based on their local periodic structures. This paper presents a technique for analysis, and comparison of DNA and protein sequences. We use local periodicities to compare sequences and develop techniques that can compare the similarity between sequences in the combined space. We further show that using correlation, and absolute error minimization between the query sequence and sequences in the database, one can search for sequences very efficiently. Thus, from the viewpoint of indexing (and otherwise), macromolecules are much simpler to deal with than images, in that a much more limited, and well-defined dictionary is sufficient for labeling the molecules.

Keywords: Macromolecule sequence analysis; Multiwindow Gabor representation; Correlation; Minimisation of differences

1. Introduction

Databases of proteins and DNA are expanding very rapidly and already contain millions of sequences, often with similar number of units. One such widely used database is the one available at the NIH. Managing such databanks requires careful indexing of macromolecules and a fast search for their efficient retrieval. The most widely used technique currently available is based on BLAST,[1] which relies extensively on searching a database and deciding whether a match exists, based on the number of correct matches between the query sequence and the sequence stored in the database. In this paper, we present an alternative technique based on local periodic fingerprints, highlighted by multiwindow Gabor signatures used in compact storage, search and retrieval of sequences from databases.

DNA sequences are comprised of four bases, adenine, guanine, thymine, and cytosine, identified by the symbols A, G, T, and C. DNA molecules have a double helical structure, with the two individual strands linked by complementary bases (Fig. 1). It is well known that A and T are complementary and G and C are likewise complementary. A thorough review regarding the structure of DNA sequences can be found in [1]. However, in the context of our study, it is sufficient to point out

*Corresponding author. Tel.: +972 4 8294726.
E-mail addresses: nagesh@tx.technion.ac.il (N.K. Subbanna), zeevi@ee.technion.ac.il (Y.Y. Zeevi).

[1]Basic Local Alignment Search Tool.

0165-1684/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2007.10.011

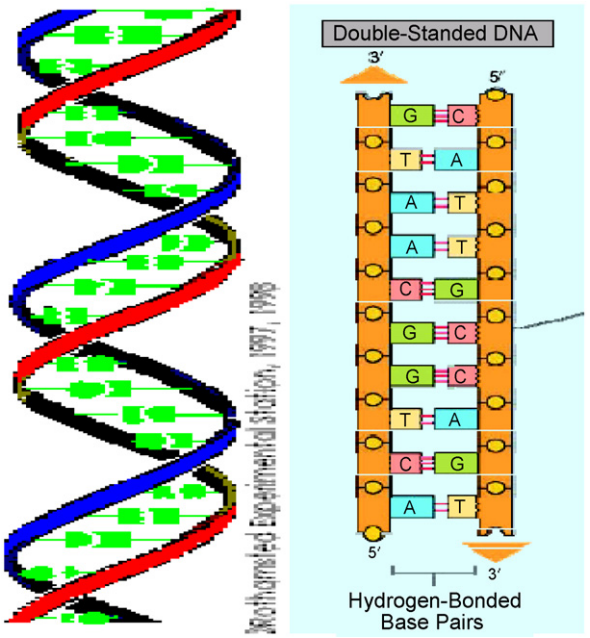

Fig. 1. (left) The double helical structure of a DNA molecule and (right) the structure of the molecule inside the double helical structure [30].

here that all DNA sequences are identified by the four base character strings. A typical, double helical DNA structure is accordingly represented by

ATTAGCGATTGCAACGCTGCATGCA

TAATCGCTAACGTTGCGACGTACGT.

Thus far, mostly character string classification and search techniques have been employed for the purpose of classifying such sequences. However, the application of signal processing techniques to the classification and analysis of macromolecular sequences, require the conversion of character strings to numerical sequences. Several techniques have been used [1]. Of these, we will utilize and compare two techniques employed most commonly. In [2], the author introduced a technique that explicitly preserves the complementary properties of the bases, and also maps the character string into a numerical sequence. We adopt this technique of mapping A, G, T, and C into $1+j, -1+j, 1-j$, and $-1-j$, respectively. The conversion of a DNA sequence into a complex numerical sequence preserving the properties of the DNA sequence allows us to apply signal processing techniques without affecting the properties of the DNA sequence. Another technique, introduced in [3], sets the values of the four nucleotides to the four vertices of a regular tetrahedron. The advantages of the technique, along with some improvements and implementational techniques, have been more thoroughly investigated in [4]. For the purpose of compactness and coherent presentation, we refer to this set of techniques as 'tetrahedron based discretization'.

Protein sequences, comprised of amino acids, present a more complicated problem for analysis. There have been many attempts to analyze protein sequences. In [5], for example, the authors show that protein sequences of certain species are incompressible with lossless compression using prediction techniques. We, however, do not aim to compress sequences of a single, specific species, but across the board of several species, where we can show that there exist many similarities.

For the case of protein sequences, we focus our attention primarily on the transmembrane sequences as they stand out in protein research and are, consequently, better understood insofar as both their structure and functions are concerned. We convert the protein sequences to numerical sequences using the technique of scale transmembrane helical propensities, proposed in [6]. This technique generates the helical propensities using the transmembrane database of globular protein sequences. Once we convert the character strings to numerical sequences, we are in a position to apply the multiwindow Gabor transforms.

Signal processing techniques are well suited to extracting local periodicities, and in particular, it is the position-frequency techniques that have been used in both images and signals to extract local periodicities (audio patterns and textures). Among the position-frequency techniques, it is the linear representation of signals (especially Gabor and wavelet representations) that are popular. In [4], the authors used a type of short time Fourier transform, cutting the signal into pieces and performing discrete Fourier transform on each of the pieces. This succeeds in combining time (position) and frequency together in plane of reference, but suffers from Gibbs ringing and effects of discontinuity. The technique has been revised and allows different windows to be used, and the code for the technique is available at [7]. The refinement is more flexible than the original and allows better results to be achieved, but suffers from a couple of drawbacks. Firstly, it is confined to one window (whatever the window function) and more importantly, the shifts are fixed along both the time and frequency windows.

It is well known that single window Gabor representation is insufficient for capturing the local changes in signals accurately [8]. Wavelets have also been used for extracting local periodicities. But, dyadic wavelets which are tailored for low frequency domain resolution at high frequencies may not be the best choice for handling DNA and protein sequences. However, the multiwindow Gabor representation is more robust (less sensitive to the shifts in the sampling intervals along the time and frequency axes), goes through all the frequencies at all resolutions and is more accurate (less sensitive to noise since several windows process the signal simultaneously at all frequencies) than both single window Gabor representations and dyadic wavelets (which have been employed in DNA and protein sequence analysis with some success [6,9,10]). Multiwindow Gabor representations combine, in a way, the advantages afforded by both single window Gabor (in the form of uniform resolution throughout the position-frequency plane) and wavelet schemes (in the form of using windows of different scales). Multiwindow Gabor schemes also eliminate the assumption of the geometric decrease of frequency resolution inherent to dyadic wavelet schemes.

Classification of macromolecules is different from segmentation and classification of images [11], video [12], and audio signals [13], in that macromolecules have a well defined alphabet and the set of alphabet is limited (just four in the case of DNA and 20 in the case of proteins). This turns the process of feature selection to be much easier. However, the issue of indexing macromolecules, i.e., the definition of macromolecular indexing vocabulary, is more complicated since only partial knowledge of the nature of macromolecules exists at the present time. This is, in fact, the subject of concerted effort in ongoing research.

As far as the relevant tools for analysis are concerned, Zibulski and Zeevi [14] established that the multiwindow Gabor frames map a sequence unitarily into the combined time (or position, in the case of spatial variables or sequences)-frequency domain. In [15], the technique was extended to discrete signals, i.e., sequences in the present context. We utilize the discrete multiwindow Gabor technique to map the sequences into the combined position-frequency domain.

However, as was proved in [16,17], a few disadvantages are associated with the canonical multiwindow Gabor representations. To overcome these disadvantages, we utilize non-canonical representations. Especially, in cases where the expansion frame is already fixed and we need to generate 'good' coefficients, it is useful to utilize non-canonical Gabor expansions.

The paper is organized as follows: In Section 2, we discuss briefly, non-canonical multiwindow Gabor representations of signals. In Section 3, we develop an algorithm for indexing, and storage of DNA sequences. In Section 4, we evolve efficient search techniques for retrieval. In Section 5, we extend the idea of classification and search for molecules to transmembrane protein sequences and, finally, compare our results with those obtained by others and discuss the implications.

2. Non-canonical multiwindow Gabor representations

We, briefly, discuss non-canonical multiwindow Gabor representations of signals; a full treatment of multiwindow Gabor functions can be found in [14,15,18]. Non-canonical Gabor representations have been considered in [16,17]. The multiwindow Gabor scheme is a combined time–frequency representation of signals, that captures local periodicities with considerable accuracy.

Throughout the paper, we consider L-periodic signals, i.e., signals that satisfy the condition $f[k]=f[k+L], k\in\mathscr{Z}$, where $\mathscr{Z}$ is the set of integers. In the context of our current study, any macromolecular finite sequence of length L can be casted as an L periodic signal.

The signal $f[k]$, can be reconstructed from the corresponding set of Gabor coefficients. The reconstruction of the signal $f[k]$ is given by [15]:

$$f[k]=\sum_{r=0}^{R-1}\sum_{m=0}^{\overline{b}-1}\sum_{n=0}^{\overline{a}-1}c_{r,m,n}\gamma_r[k-na]\mathrm{e}^{\mathrm{j}2\pi mbk/L}, \tag{1}$$

where $\gamma_r[k]$ are the dual windows, $\overline{a}=L/a\in\mathscr{N}$ and $\overline{b}=L/b\in\mathscr{N}$ are the number of sampling intervals along the time and frequency axes, respectively. We assume that both a and b are both divisors of L [19].

The coefficients of the multiwindow Gabor expansion are given by the projection of the finite signal $\mathbf{f}\in\mathscr{C}^L$ onto the combined space

$$c_{r,m,n}=\sum_{k=0}^{L-1}f[k]g_r[k-na]\mathrm{e}^{-\mathrm{j}2\pi mbk/L}, \tag{2}$$

where $g_r[k], r \in 0,\ldots,R-1$ are the window functions, a and b are the combined space sampling intervals along the sequence position and frequency axes, respectively.

It was established in [15] that a necessary condition for complete reconstruction in the case of multiwindow Gabor expansions is given by $R\overline{ab} \geqslant L$. In the case of critical sampling and a single window, the reconstruction is unstable according to the Balian–Low theorem [20]. This theorem extends to well-behaved multiwindows [14]. We, therefore, consider only the oversampling case where $R\overline{ab} > L$, which implies that the functions $g_{r,m,n}[k]$ are linearly dependent and the representation is overcomplete.

In vector form, (2) can be written as

$$\mathbf{c} = \mathbf{G}^*\mathbf{f}, \tag{3}$$

where $\mathbf{c}$ is the vector of coefficients, and $\mathbf{G}$ is the Gabor matrix

$$\mathbf{G} = \begin{bmatrix} g_{0,0,0}[0] & \cdots & g_{R-1,\overline{a}-1,\overline{b}-1}[0] \\ g_{0,0,0}[1] & \cdots & g_{R-1,\overline{a}-1,\overline{b}-1}[1] \\ \vdots & \ddots & \vdots \\ g_{0,0,0}[L-1] & \cdots & g_{R-1,\overline{a}-1,\overline{b}-1}[L-1] \end{bmatrix}, \tag{4}$$

with $g_{r,m,n}[k] = g_r[k-na]\mathrm{e}^{\mathrm{j}2\pi mbk/L}$. The reconstruction, inverse of (2), can be written in the following vector form version of (1):

$$\mathbf{f} = \mathbf{\Gamma}\mathbf{c}, \tag{5}$$

where $\mathbf{\Gamma}$ is the dual of the Gabor matrix.

Since the representation is overcomplete, there exist an infinite number of possible duals $\gamma_r[k]$. The canonical solution yields the minimum norm dual of the set of generalized Gabor elementary functions $g_{r,m,n}[k]$ by [19],

$$\tilde{\gamma}_r[k] = (\mathbf{G}\mathbf{G}^*)^{-1} g_r[k]. \tag{6}$$

However, it is often better to choose a different dual from a wider set of duals. Here, we extend the approach to non-canonical duals, introduced in [21]. The non-canonical dual is given by

$$d_{r,m,n} = d_r[k-na]\mathrm{e}^{\mathrm{j}2\pi mbk/L} = (\mathbf{H}\mathbf{G}^*)^{-1} h_r[k], \tag{7}$$

where $\mathbf{H}$ is a Gabor matrix of the form (4) with the vectors $h[\cdot]$ forming the columns of the matrix $\mathbf{H}$. It is of importance to mention that the set of vectors $h_{r,m,n}[\cdot]$ also constitutes a frame for $\mathscr{C}^L$. The only requirement for the existence of a frame of this form is that the matrix $\mathbf{H}\mathbf{G}^*$ be invertible [16]. The problem, therefore, is to ensure that the matrix $\mathbf{H}\mathbf{G}^*$ is invertible.

The problem of invertibility of the matrix $\mathbf{H}\mathbf{G}^*$ has been dealt at length in [17]. Conditions for the invertibility of $\mathbf{H}\mathbf{G}^*$ have been derived for both the rational and integer oversampling cases. In this paper, we have chosen integer oversampling of the sequences. Two properties of the matrix $\mathbf{H}\mathbf{G}^*$ that are of great use in ascertaining the invertibility of the matrix are mentioned below:

1. The matrix $\mathbf{H}\mathbf{G}^*$ has non-zero elements only on the principal diagonal and its $\overline{b}$th sub-diagonals, as shown in Fig. 2.
2. The matrix $\mathbf{H}\mathbf{G}^*$ is block circulant in terms of $a \times a$ blocks as shown in Fig. 2.

Utilizing these two properties of the matrix $\mathbf{H}\mathbf{G}^*$ and the conditions for invertibility for block

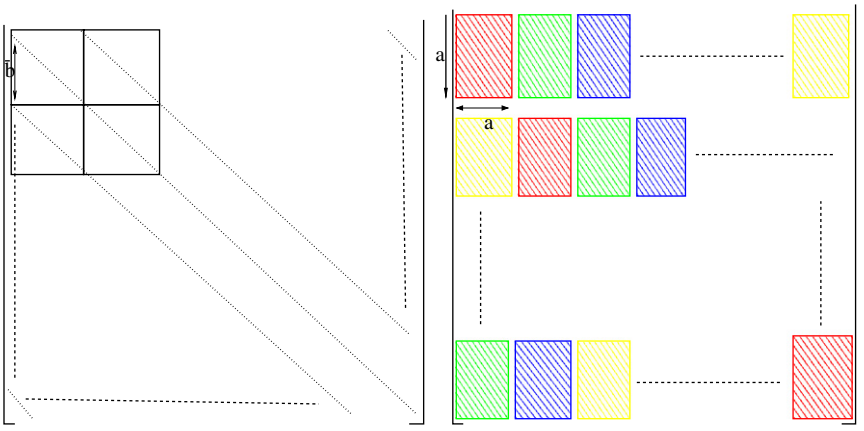

Fig. 2. (left) The banded structure of the matrix wherein the only non-zero elements appear on the principal and $\overline{b}$th subdiagonals. (right) The block circulant structure of the matrix with the $a \times a$ blocks rotating right and wrapping around at the end of the row.

circulant matrices [22], we can derive the condition for invertibility of the matrix $\mathbf{HG}^*$ under conditions for integer oversampling. The result can be stated as follows:

Theorem 2.1. *Under the conditions of integer oversampling*,[2] *if* $g_{r,m,n}[\cdot]$ *is a positive definite sequence*,[3] *and* $h_{r,m,n}[\cdot]$ *is positive at all points, then the matrix* $\mathbf{HG}^*$ *is always invertible.*

Proof. see [17]. □

Theorem 2.1 permits using all positive definite window functions (including Gaussians) in non-canonical multiwindow Gabor expansions. It is observed that there is no real restriction on the possible choices of the **H** and **G**, as long as one of them is positive definite and the other is completely positive (or negative). Further, we are no longer restricted to the least square solution as we were in the case of the canonical solution. We can even optimize over all the possible solutions using the non-canonical duals for a different norm (like the L_1 norm). The flexibility of the solution is the most important advantage of the non-canonical solution.

3. Storage and indexing of DNA sequences

We apply the multiwindow Gabor transform (3) to sequences of length L and obtain the set of coefficients **c**. One of the most curious features of multiwindow Gabor techniques is that with a proper choice of window and lattice parameters, one can represent a DNA sequence to a remarkable degree of accuracy with a very small number of coefficients. However, that this is possible mainly in cases where the segments of sequences under investigation incorporate coding regions. It has been observed that such segments exhibit many local periodicities [2,23]. Further, according to the uncertainty principle, and to Gabor's theory of signal representation in combined spaces, the wider the window the better is the frequency resolution (i.e., the definition of the index) and *vice versa* [14].

Separating out the coding regions of a DNA from the non-coding regions is an important task for biologists. It is well known from the theory of multiwindow Gabor representations that narrow windows detect sharp changes accurately, whereas wide windows can detect changes over a greater length [15]. Thus, it is possible to employ longer windows to detect the coding regions and perform an analysis on them using the narrow windows in a hierarchical way. However, we have used both narrow and wide windows on the entire sequence, in order to avoid missing short coding regions (which may be lost if broad windows alone are used to find the coding regions).

3.1. Compression of sequences

To illustrate our point, consider the example of the multiwindow Gabor representation of the sequence AF099922 depicted in Fig. 3.

It is evident from Fig. 3 that the blue colored Gabor pixels (codons) have little significance and can be easily ignored. In our experiment, we oversampled the sequence of length 960 by 15 times and tried to reconstruct the sequence from the coefficients. With as few as 397 highest coefficients, we are able to reconstruct the sequence with 93% accuracy. Actually, this is not surprising in view of the fact that local periodic components dominate in the process of generating the high coefficient values. A cross section of the local spectrum is shown in Fig. 4. This emphasizes the importance of local periodicities in generating the signature of the macromolecule. In the examples shown in Figs. 3 and 4, the coefficients corresponding to frequencies (21–23)/40 and (12–14)/40 are more than three and two times, respectively, larger than the next largest coefficients. The peaks of the transformed sequence identify the principal value of the frequency components at the positions $n = 3a$ and $n = 85a$. The technique of using non-canonical multiwindow Gabor window functions can be considered as a method of lossy compression. In fact, it is possible to recover even the 'lost' portions of the signal using approximation techniques, since there are only four possible values in case of DNA (or RNA) signals. An effective, albeit lossy compression can be achieved using non-canonical multiwindow Gabor coefficients.

There is another point to be considered here. In many cases, it is sufficient to reconstruct the areas of interest in the sequences. In such cases, the periodicity of the coding regions usually ensures that they are reconstructed faithfully and correctly. It is the regions that are completely unstructured that are more difficult to reconstruct with fewer coefficients. In such cases, any prior knowledge of the sequence can be exploited in the reconstruction.

[2]Under conditions of integer oversampling, $\overline{b}$ is divisible by a.

[3]Positive definite functions are those functions whose DFT is real and positive at all points.

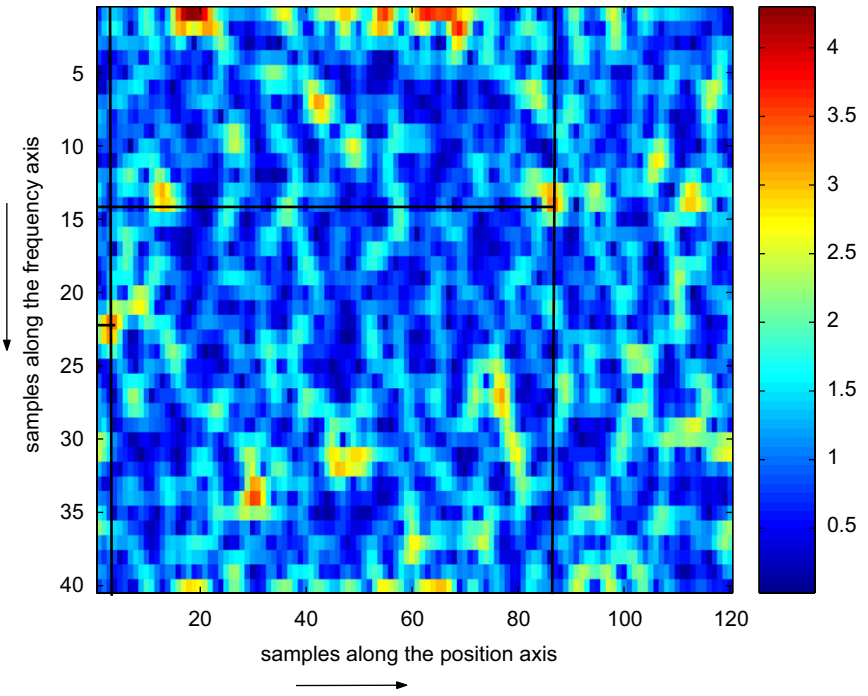

Fig. 3. Gabor coefficients of the sequence AF099922 ($L = 960$) using a Gabor window of effective width $\sigma = 64$, and the combined space sampling parameters $a = 8, b = 24$. The values of the peaks are indicated by 'T' joints at (3,22) and (85,14) and the dominant frequencies are shown by lines parallel to the frequency axis.

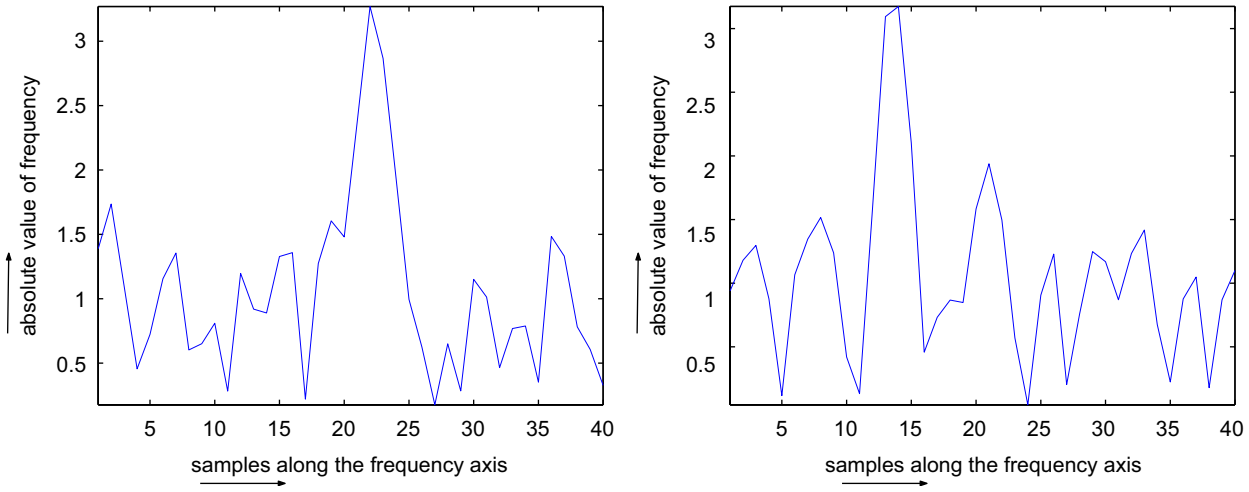

Fig. 4. Representation of the local frequency spectrum using a cross section of the coefficients of sequence AF099922 at position shifts $n = 3a$ and $n = 85a$.

4. Searching for DNA sequences

Given the coefficients representing sequences, to facilitate efficient search for a set of sequences, we develop a method of indexing molecules based on the local periodicities, where the energy of the coefficients is given by $\sum_{r,m,n}|c_{r,m,n}|^2$. Let us denote the total energy of the coefficients by E. In our case, it is given by

$$E = \sum_{r=0}^{R-1}\sum_{m=0}^{\bar{b}-1}\sum_{n=0}^{\bar{a}-1}|c_{r,m,n}|^2. \tag{8}$$

We select the individual coefficients $c_{r,m,n}$ in a decreasing order of energy, and do so until we reach a certain percentage η of the total energy, where η is specified by the user, in accordance with the structure of the molecules under consideration. Formally, we have

$$\eta = \frac{\sum_{s=0}^{V-1} |c_s|^2}{E}, \tag{9}$$

where c_s are the coefficients arranged in descending order and V is the total number of coefficients that satisfy the specified criterion. Now we have only a small subset of coefficients and these can be represented in one dimension using the position of the coefficient. Although, in theory, the coefficients constitute a three-dimensional array, i.e., they are labeled by the three parameters r, n, and m, the sequence obtained from the set of coefficients is a one-dimensional array. It suffices to store the position in a one-dimensional array, since the conversion from the one-dimensional position to the 'three dimensions' is given by

$$c_p = c_{r,m,n} = c_{r\overline{a}\overline{b}+n\overline{b}+m}, \tag{10}$$

where $p \in 0, \ldots, r\overline{a}\overline{b}$. In other words, we just store the positions and use the above formula during search.

We proceed to develop a method to search for sequences in the database based on correlation of coefficients [24]. We modify the method given in [24], to require only the important coefficients of two sequences. We then show that our method is not only faster, but also alleviates false negative results in the search process (this problem is endemic to BLAST and its clones).

Let the coefficients of the first sequence $f_1[k],\ k \in 0, \ldots, L_1 - 1$ (sequence in which we search for the match) be given by (2)

$$c^{(1)}_{r,m,n} = \sum_{k=0}^{L_1-1} f_1[k] g_r[k - na] \mathrm{e}^{-\mathrm{j}2\pi m b_1 k / L_1}, \tag{11}$$

where $c^{(1)}_{r,m,n},\ m \in 0, \ldots, \overline{b}, n \in 0, \ldots, \overline{a_1} - 1, r \in 0, \ldots, R-1$, are the coefficients, L_1 is the length of the sequence, b_1 is the shift along the frequency axis, and $\overline{a_1}$ is the number of shifts along the sequence (position) axis.

Similarly, the coefficients of the query sequence $f_2[k],\ k \in 0, \ldots, L_2 - 1$ are given by

$$c^{(2)}_{r,m,n} = f_2[k] * g[k - na] \mathrm{e}^{-\mathrm{j}2\pi m b_2 k / L_2}, \tag{12}$$

where $c^{(2)}_{r,m,n},\ m \in 0, \ldots, \overline{b}, n \in 0, \ldots, \overline{a_2} - 1, r \in 0, \ldots, R-1$, are the coefficients, L_2 is the length, b_2 is the shift along the frequency axis, and $\overline{a_2}$ is the number of shifts along the position axis. It is important to note that in both the labeled-and-stored and the query sequences, the values of a and $\overline{b}$ should be kept a constant. Adaptive approaches involving modifications of these are possible, but would require a more complicated approach to calculating the matches between the sequences than is indicated in this paper.

We define $\kappa[p], p \in 0, \ldots, \overline{a_1} - \overline{a_2}$ as the value of correspondence between the coefficients of the sequences. Formally, we define $\kappa[p]$ as

$$\kappa[p] = \sum_{n=0}^{\overline{a_2}-1} \sum_{r=0}^{R-1} \sum_{m=0}^{\overline{b}-1} |c^{(1)}_{r,m,n+p} \| c^{(2)}_{r,m,n}|. \tag{13}$$

A high value of $\kappa[\cdot]$ indicates a match and a low value of $\kappa[\cdot]$ indicates lack of a match. Since the acceptance threshold (value of $\kappa[\cdot]$ below which we denote no match) of κ is a tunable parameter, the degree of approximation regarding acceptance of a match is controlled by the user.

In our case, a vast majority of the coefficients are unnecessary in determining the match. A refinement would be to reduce the number of actual coefficients, using thresholds and normalizing the sequences. This would, therefore, compare only corresponding high value coefficients in the two sequences and reduce the computational time for searching. Formally, this can be written as

$$\kappa[p] = \sum_{v} |c^{(2)}_{v} \| c^{(1)}_{p\overline{b}+v}|, \tag{14}$$

where $p \in 0, \ldots, \overline{a_1} - \overline{a_2}$, and v is the set of coefficients chosen in the query sequence. Since multiwindow Gabor transforms use local properties, this technique would rule out false negatives, since similar sequences would have, at least reasonably high values in the corresponding positions along the sequence and frequency. The only need to find similar sequences would be to set a threshold for the correlation value $\kappa[\cdot]$, above which all sequences would be acceptable. In DNA sequence analysis, it is imperative, usually, to find all the nearly-similar sequences, not merely precise matches. Further, since slight differences in DNA sequences do not produce major difference in the coefficients (which is the principal problem with global Fourier methods), the localization of

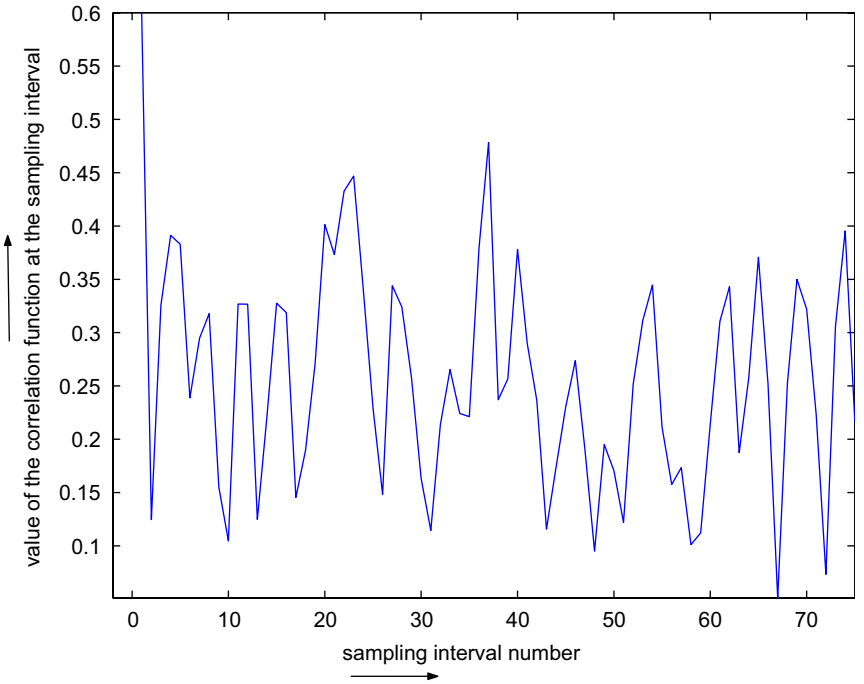

Fig. 5. The correlation function of a 'slightly-altered' subsequence of length 75 of the sequence AP000543.

variation in case of changes also helps in examining the periodicity properties in DNA sequences.

In the case of DNA sequences, it is well known that there are two types of repeating patterns. The first, which we have analyzed, is the feature of approximate matches. The second one is called the feature of reverse complements. It has been observed [25] that there are several sequences where the reverse complement of a subsequence occurs as a subsequence. As an example, a subsequence of *ATTGCA* would have a reverse complement of *TGCAAT*, and it is important to locate the reverse complement in the sequence. This problem is conveniently solved using the multiwindow Gabor representations.

If a subsequence $\mathbf{f_s}$ has a reverse complement $\hat{\mathbf{f}}_\mathbf{s}$, the reverse complementarity is characterized by equal values for the subsequences in the real part of zero frequency region for all the windows used. Consider the sequence $\mathbf{f_s}$ of length L_s and its reverse complement $\hat{\mathbf{f}}_\mathbf{s}$. The zero frequency multiwindow Gabor coefficients for the two sequences, are given by

$$c^{(s)}_{r,n,0} = \sum_{k=0}^{L_s-1} f_s[k] \mathrm{e}^{-(k-na)^2/(\sigma_r^2)}, \tag{15}$$

and

$$\hat{c^{(s)}}_{r,n,0} = \hat{f}_s[k] \mathrm{e}^{-(k-na)^2/(\sigma_r^2)}, \tag{16}$$

where $c^{(s)}$ and $\hat{c^{(s)}}$ are the Gabor coefficients of the sequences. Considering that $\hat{\mathbf{f}}_\mathbf{s}$ is nothing more than the conjugate of $\mathbf{f_s}$, we can easily see that the zero frequency values of the sequences, with a real window, are equal (they are just weighted averages) and their real parts being equal in the signal domain. These properties are preserved in the combined space–frequency domain, and as an advantage, we operate over a smaller set of shifts along the sequence $\overline{a}$ instead of the set of all elements in the sequence (a larger set L).

Since multiwindow Gabor handles changes on a local basis, the length of the subsequence in relation to the total does not play much of a role (this is especially true of the windows having a small effective width). In a large sequence, the above is easily achieved by comparing the zero frequency coefficients of a sequence $\mathbf{f}$ generated by Eq. (3) with the coefficients of the sequence $\hat{\mathbf{f}}$ and finding the equal values in the corresponding positions.

Shown in Fig. 5 are the results of the comparison of a subsequence within a sequence. We also see that it is very easy to capture the periodicities in coding

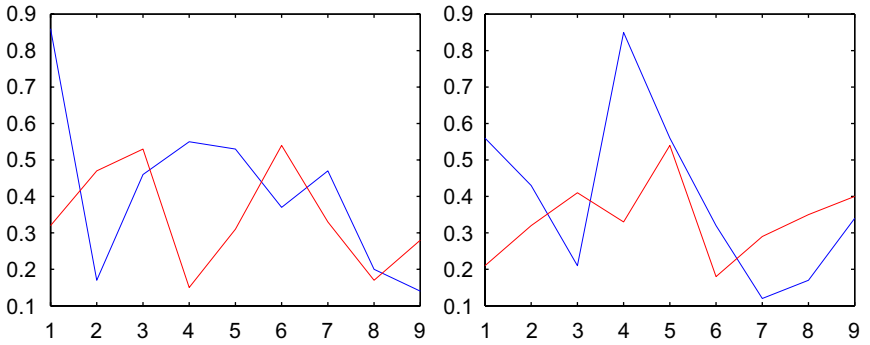

Fig. 6. The correlation function of two subsequences and their original sequences and each other. The correlation of the subsequence with the original sequence is given in blue, and the correlation of the sequence with the 'wrong' sequence is given in red. The sequences used are AP000543 and Y00821.

sequences, as is depicted in the example below. These periodicities can be used in the classification of similar coding sequences.

Finally, we can see that the correlation for the 'right match' is maximum due to the unitary nature of the Gabor coefficients, for a given set of analysis and synthesis frames. Shown in Fig. 6 are the Gabor coefficients of two subsequences of two different sequences matched against their own sequences and each other. We can see that the autocorrelation results in a much higher correlation factor than the cross correlation factor, even when the sequences are similar. Fig. 6 establishes that the 'right' matches have much higher correlation, and consequently, the chances of getting false negatives and positives is much smaller than in many other methods where local aberrations can result in large changes.

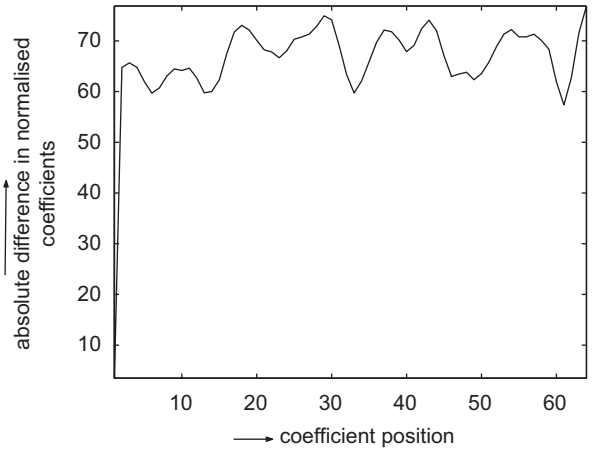

Fig. 7. Minimization of difference between the coefficients of the database and those of the query sequences. The sharp drop in differences indicates almost perfect match.

4.1. Minimization of differences

In the preceding section, we used the maximum correlation as a measure of similarity. Here, we illustrate that a similar, and in many cases, superior result, can be achieved by minimizing the differences between the coefficients of two sequences. As in the previous case, we threshold the coefficients and align the sequences over the shifts to minimize the error. It can be written as:

$$e[k] = \sum_{r=0}^{R-1} \sum_{n=0}^{\overline{a_2}-1} \sum_{m=0}^{\overline{b}-1} |c^{(1)}_{r,m,n+k} - c^{(2)}_{r,m,n})|, \tag{17}$$

where $e[\cdot]$ is the error (or the difference) between the coefficients, and $k \in 0, \ldots, \overline{a_1} - \overline{a_2}$. The minimum of $e[k]$ yields the best match. Fig. 7 shows that if the absolute differences between two normalized coefficients are used, the actual matching sequence is found just as effectively. Here the absolute difference between the 'most pronounced' coefficients of the sequence in the database and the query sequence are chosen. Clearly, the best match (which happens to be at the beginning of the sequence) is rather more pronounced than in the case of the correlation with the query sequence. The best match is nearly an order of magnitude smaller than the next best match!

5. Indexing of protein sequences

The method of multiwindow Gabor sequences is now extended to the case of protein sequences. In

this case, there are 20 amino acids that form the building blocks of these sequences. We use the method suggested in [6] for converting the character strings to protein sequences. Other possible methods are also available [26–28]. The method proposed by Lio et al. in [6] involves computing the propensity of the transmembrane helix, and scaling it, based on the amino acid normalized frequencies (number of occurrences) estimated from the transmembrane database (TMALN). The amino acid propensities after scaling are the same as suggested by Lio et al. [6].

For the case of protein sequences, we apply the same multiwindow Gabor transforms, and obtain the coefficients. In the case of transmembrane sequences, one of the most important requirements is to find the transmembrane helices. These subsequences show a certain periodicity and, therefore, concerted effort has been devoted to locating these periodicities using Fourier transforms [29], and dyadic wavelets [6,9].

To illustrate it, we consider the chemokine receptor (CKR5) sequence in humans, where there are five transmembrane helices [6]. We use the zero-mean sequence obtained by the discretization method proposed in [6]. It is easily observed that we obtain the transmembrane helix components by simply finding the peaks in the values of the coefficients as in the case of [6]. Since the sequence is characterized by its zero-mean, we capture both the positive and negative components of the frequency, as is apparent in the reflection of the coefficients across the central frequency (Fig. 9). Four of the five helices are identified by different spectral bands, indicating that their periodicities are of different lengths, while the TM helices 4 and 5 lie in the same spectral band, showing that they have roughly the same local periodicity length. (They are observed to have lengths of 27 and 25, respectively [6]. Thus one may conclude they are of roughly equal length.) It is also interesting to observe that there are other periodicities that have not been classified by other techniques. Since our method uses windows of different widths, it can capture 'hidden' periodicities and thus help biologists in their concerted efforts to understand these unknown structures better. Thus far, biologists have been constrained to have at least a rough idea of the periodicities of TM sequences, before they could attempt to find the exact location and periodicities. Our method alleviates this problem to a considerable extent, by removing the bar on the window spreads that were present in other methods. Further details will be provided in an upcoming paper.

By comparing our locations of the TM helices with the observed helices and the location proposed in [6], we can see that both the methods coincide in predicting the TM helices to the same locations and to have similar lengths. Another feature of interest is the periodicities at frequencies of (1/30) at position 135, which are not known to have any significance at present. It is interesting to investigate the validity and purpose of this particular periodicity.

6. Comparisons with existing techniques

There is a considerable amount of literature available on the utilization of position-frequency techniques to classify, and search sequences. In this paper, we choose three existing techniques to compare our method with. Apart from this, we will give a theoretical comparison with BLAST and its clones.

6.1. Comparison with DFT-based technique

This technique has been utilized by Anastassiou and his group [3,4]. The principal idea behind the technique is to cut the sequence into pieces and perform DFT on each part. Two separate representations of sequences have been used, and the authors have shown that they can find several characteristics in coding regions of sequences. The major problem of this technique is that the DFT is global, and further, the choice of using rectangular windows (cutting the sequence into pieces is essentially equivalent to utilizing the rectangular window) introduces Gibbs oscillations. The utilization of multiwindow Gabor windows can alleviate the problems to a large extent, since the Gaussian windows are less prone to the above mentioned problems. Further, it has been established that the Gaussian window achieves the best resolution in the position-frequency space.

6.2. Comparison with statistical-Fourier hybrid technique

We have implemented the multiwindow Gabor approach in the representation of a subsequence of c:20h12, in the ACES region of the chromosome 22 (Genbank accession number AP000543, [31] Dunham et al.). The sequence AP000543 is one of

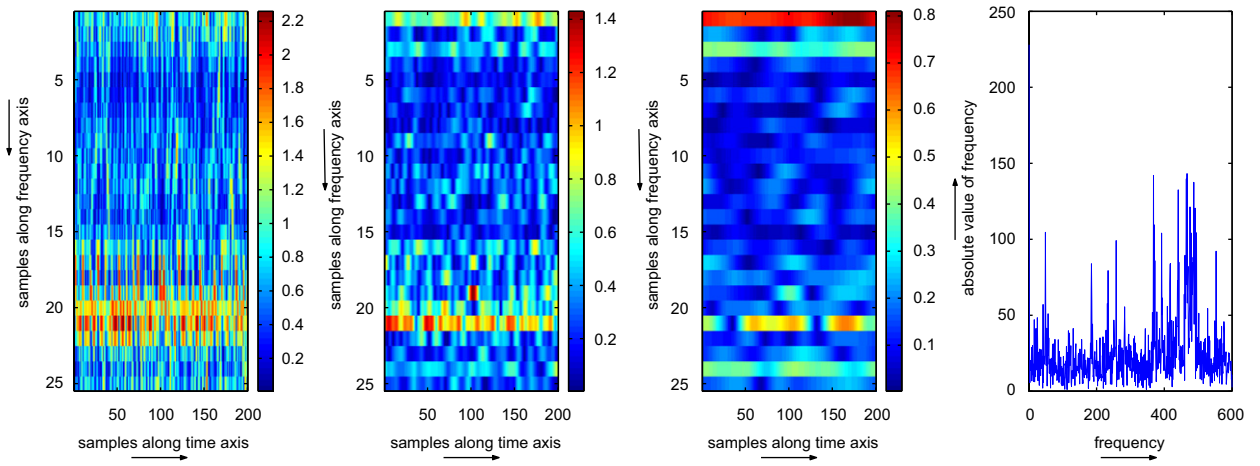

Fig. 8. 2D representation of the coefficients of sequence AP000543 of length $L = 600$ using three normalized Gaussian window functions with effective widths $\sigma_1 = 3$, $\sigma_2 = 12$, and $\sigma_3 = 48$, using combined space sampling parameters $a = 3, b = 24$, and representation of the absolute values of the DFT of AP000543 ($L = 600$).

the 27 sequences mapping the Cat's Eye syndrome genes in the centrometric region of the chromosome 22. This is the sequence utilized in [32]. The authors have used a mixed technique. Initially, a modified autocorrelation is utilized to detect similar bases in a DNA sequence, and then a DFT is performed on it. We have chosen a part of the sequence AP000543 and reimplemented the results of [32] in our formalism, and the results are displayed in Fig. 8. The repetition of the bases in the DNA sequence is highlighted very clearly at frequency 20/25–22/25, in the example shown in Fig. 8. The Fourier transform of the sequence also yields a peak in the region of (480/600), but there are other peaks in the region of (380–400)/600, which prevent clear detection of the accurate frequency of the periodicity. In fact, the peak at frequency (20–22)/25 is clearly the case in all three windows, emphasizing that the peak at this frequency is spread throughout the sequence. It also permits us to observe that the local variations—around 20/25–22/25 in the narrow window—gives way to a more stable structure in the larger windows (yielding the coefficients over a longer distance). This also seems to corroborate the conjecture that coding regions are very periodic, and that our technique allows detection of hidden periodicities. In [32], the authors try to find the long range correlations in the sequence with statistical techniques. The major disadvantage of this technique is that the modified autocorrelation function controls the effective area of search for similarities, whereas the multiple windows are more flexible in handling the various frequencies. The second problem is that almost periodic and minor aberrations are more difficult to detect using the modified autocorrelation technique. Our method is much more direct and efficient, since it allows us to see the patterns directly from the coefficients of the multiwindow Gabor transform.

Results such as those shown in Fig. 5 indicate that one can locate (slightly altered) a subsequence in a sequence where there are repetitive patterns. The subsequence consists of the first 75 bases of the sequence AP000543 (with some small alterations in the subsequence of 75 to ensure proper rendering within a certain, limited radius of the subsequences). The correlation at locations of coincidence is far greater than at other locations, (nearly 1.33 times the value at other places). This technique, thus easily affords finding the corresponding 'near matches' in other sequences.

6.3. Comparison with wavelet-based techniques

Wavelet-based techniques have been used with considerable success in several contexts such as locating the transmembrane helices in a transmembrane protein sequence. In [6], the authors utilized a thresholded Daubechies wavelet to detect the presence of transmembrane helices in the sequence. As seen in Fig. 9, we have reimplemented the method of [6] in our formalism. Other efforts such as [9,10], utilize similar techniques for the detection of hydrophobic cores in the sequences. The principal idea is to discretize the sequence based on the

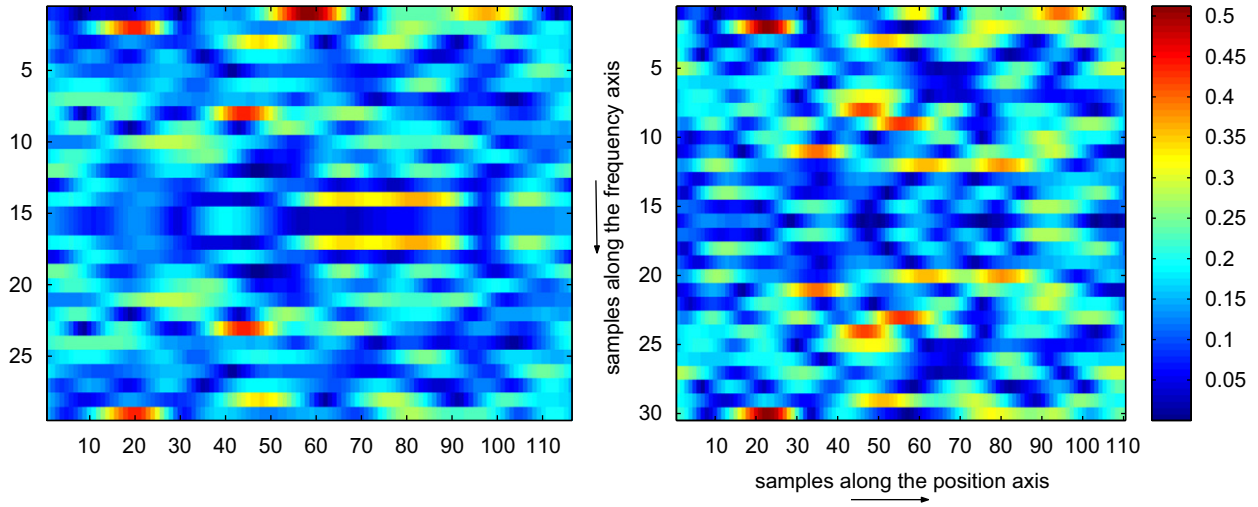

Fig. 9. A combined space representation of the protein CKR5 using a Gabor window of effective spread $\sigma = 27$. We can easily see all the five TM subsequences corresponding to the 'red' and 'orange' regions in the figure (left). The combined space representation of the same sequence using the dyadic Daubechies wavelets. All the regions are still easily recognizable, but the one problem is that there is a bit of scattering of frequencies due to the use of geometric progression for frequencies (right).

hydrophobicity of the residues and then use wavelet transforms to find the transmembrane helices.

As can be seen in Fig. 9, in the case of wavelet transforms, the dyadic wavelet causes the scattering of frequencies around position 40 (frequency (8–9/30)). This is due to the logarithmic progression of frequencies, as opposed to the arithmetic progression of frequencies in the case of Gabor functions. The disadvantage of using the logarithmic scale for frequencies is that proper resolution of frequency is impeded at high frequencies. Besides, dyadic wavelets are often tailored to human vision and are not necessarily the best for other signals/sequences. We see a rather better localization of frequencies in the case of Gabor transforms. A further problem of dyadic wavelets is harmonics which tend to generate artifacts—in Fig. 9, the high points at position 30 are seen at two frequencies 10/30 and 14/30.

In the case of multiwindow Gabor representations, since all the windows span the entire range of frequencies along with the positional information, we can capture both the transmembrane helices at their locations and other important periodicities. Given the entire protein sequence, the multiwindow Gabor coefficients help in locating and classifying the different transmembrane structures. One curious feature of the transmembrane helices is that when they are of different lengths, we need windows of different effective spreads to localize them. The utilization of the multiwindow Gabor representation permits us the degree of freedom of choosing the windows of convenience (something absent in Fourier and dyadic wavelets) and alleviates the problem inherent in the other techniques.

We have compared our method with the methods of [32,2] for DNA sequences, and with those presented in [6] for protein sequences. We show the advantages of using multiwindow Gabor representations for detecting local periodicities, storing and searching for sequences. In conclusion, our techniques have the potential of accelerating the access into the databases of macromolecules by properly indexing them according to the fingerprints of local periodicities. The need for such indexing is growing rapidly as more macromolecular sequence data becomes available.

6.4. Comparison with BLAST

A short comparison between the philosophies of traditional techniques such as BLAST and the position-frequency techniques is in order. BLAST and other techniques rely on aligning the sequences and finding matches (either between the DNA nucleotides, or between amino-acid residues). For each correct match, the sum (which is originally zero) is incremented by a predetermined amount (determined by a combination of context and heuristics), and for each mismatch, the sum is decremented by a predetermined amount. These techniques attempt to make use of the general

knowledge of sequences, as well as any specific knowledge about the sequence in question. The predetermined increments and decrements often cast away flexibility, and in case wrong heuristics are used, tend to produce false negatives sometimes. We, in contrast, rely on multiple windows (which essentially capture multiscale 'mean-tree paths') to identify similarities and local periodicities. This greatly enhances the robustness of the coefficients to local aberrations. The number of windows may be increased with no substantial increase in computational demands. As long as the choice of the lattice constants a and b is sane, multiwindow Gabor functions are less vulnerable to local variations and can find similarities across sequences without too much trouble). The sliding window correlation we use ensures we compare the sequences reasonably well. However, it must be admitted that our technique has been tested mainly on periodic sequences, especially coding regions, and has not been tested on other important DNA regions like binding sites. Consequently, the performance against BLAST with regard to the other important features like binding sites cannot be assessed.

7. Discussion

We have discussed the utility of multiwindow Gabor frames in searching, storing, retrieving, and classifying sequences. Another, perhaps just as important ability of Gabor coefficients is to help predict the secondary structure of protein sequences. To an extent, this ability has been observed in predicting the transmembrane helices in transmembrane sequences, as has been observed in the previous sections. This ability can be extended to other structures like alpha-helices, beta-strands, and observing defects in collagen fibers. These possibilities are being investigated and promising preliminary results indicate that it would be very useful to utilize this technique to predict secondary structures of protein sequences. A further advantage of the technique is that Gabor transforms are unitary, which means that it would be possible to use energy minimization techniques to find stable structures (something that is extensively used in molecular mechanics to find the structure of the protein). The application of position-frequency techniques to the detection of protein structures is likely to be of great use to biologists.

In conclusion, we can say that our technique measures up well against all the three techniques as we have shown above. Further, our technique can also be used to predict the secondary structure of the amino acid sequences. Given these advantages, it seems a fair supposition that utilising position-frequency structures in the field of macromolecule sequence analysis is an excellent idea.

Acknowledgments

The authors wish to express their gratitude to Professors Y.V. Venkatesh, N. Lotan, Y.C. Eldar, and M.R.N. Murthy for useful discussions on the subject. This research program has been supported in part by the Ollendorf-Minerva Research Center, by the HASSIP Research Network Program HPRN-CT-2002-00285, sponsored by the European Commission, and by the Fund for Promotion of Research at the Technion.

References

[1] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological Sequential Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, 1998.
[2] D. Anastassiou, Genomic signal processing, Signal Process. Mag. 18 (4) (September 2000) 8–20.
[3] D. Anastassiou, Frequency domain analysis of biomolecular sequences, Bioinformatics 16 (12) (December 2000) 1073–1081.
[4] D. Susillo, A. Kundaje, D. Anastassiou, Spectrogram analysis of genomes, EURASIP J. Signal Process. (Special Issue on Genomic Signal Processing) (1) (2004) 29–42.
[5] C.W. Neville-Manning, I.H. Witten, Protein is incompressible, in: Proceedings of the Conference on Data Compression 1999 (DCC99), pp. 257–265.
[6] P. Lio, M. Vanucci, Wavelet change-point prediction of transmembrane proteins, Bioinformatics 16 (2000) 376–382.
[7] D. Sussillo, Spectrofish, ⟨http://www.ee.columbia.edu/~sussillo/spectrofish/⟩.
[8] S.G. Mallat, A Wavelet Tour of Signal Processing, second ed., Academic Press, San Diego, CA, 1999.
[9] K.B. Murray, D. Gorse, J.M. Thornton, Wavelet transforms for the characterization and detection of repeating motifs, J. Mol. Biol. 316 (2002) 341–363.
[10] H. Hirakawa, S. Muta, S. Kuhara, The hydrophobic cores of proteins predicted by wavelet analysis, Bioinformatics 4 (1999) 141–148.
[11] A.K. Jain, F. Farrokhnia, Unsupervised texture segmentation using Gabor filters, Pattern Recognition 24 (12) (1991) 1167–1186.
[12] D. Zhong, S.F. Chang, Video object model and segmentation for content-based video indexing, in: International Symposium for Circuits and Systems (ISCAS97), pp. 1492–1495.
[13] E. Scheirer, M. Slaney, Construction and evaluation of a robust multifeatures speech/music discriminator, in:

Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP97), pp. 1331–1334.
[14] M. Zibulski, Y.Y. Zeevi, Analysis of multiwindow Gabor type schemes, Appl. Comput. Harmonic Anal. 4 (2) (1997) 188–221.
[15] M. Zibulski, Y.Y. Zeevi, Discrete multiwindow Gabor type transforms, IEEE Trans. Signal Process. 45 (6) (June 1997) 1428–1442.
[16] N.K. Subbanna, Y.C. Eldar, Efficient Gabor expansion using non-minimal dual Gabor windows, in: Proceedings of the International Conference on Electronics, Circuits, and Systems (ICECS), 2004.
[17] N.K. Subbanna, Y.C. Eldar, Y.Y. Zeevi, Oversampling of the generalized multiwindow Gabor scheme, in: International Workshop on Sampling Theory and Applications (SampTA), Samsun, Turkey, July 10–15, 2005.
[18] M. Dorfler, Gabor analysis for a class of functions called music, Ph.D. Dissertation, University of Vienna, 2003.
[19] S. Qiu, H. Feichtinger, Discrete Gabor structures and optimal representations, IEEE Trans. Signal Process. 43 (10) (October 1995) 2258–2268.
[20] I. Daubechies, The wavelet transform, time–frequency localization, and signal analysis, IEEE Trans. Inform. Theory 36 (5) (September 1990) 961–1004.
[21] Y.C. Eldar, O. Christensen, Characterization of oblique dual frame pairs, J. Appl. Signal Process. 2006 (ID 92674) (2006) 1–11.
[22] P.J. Davis, Circulant Matrices, Wiley, New York, 1979.
[23] P.P. Vaidyanathan, B.J. Yoon, The role of signal processing concepts in genomics and proteonomics, J. Franklin Inst. (Special Issue on Genomics) (2004) 1–27.
[24] E. Sejdic, J. Jiang, Comparative study of three time frequency representations with applications to a novel correlation method, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2004, pp. 633–636.
[25] T. Matsumoto, K. Sadakane, H. Imai, Biological Sequence Compression Algorithms, 1999.
[26] Y. Pilpel, N. Ben-Tal, D. Lancet, kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments, Application to membrane protein structure prediction, J. Mol. Biol. 294 (1999) 921–935.
[27] S.J. Hubbard, J.M. Thornton, Naccess Computer Program, Department of Biochemistry and Molecular Biology, University College London, 1993.
[28] T. Ertzold, A. Ulyanov, P. Argos, Information retrieval systems for molecular biology data banks, Methods Enzymol. 266 (1996) 114–128.
[29] Y. Pilpel, D. Lancet, The variable and conserved interfaces of modeled olfactory receptor proteins, Protein Sci. 8 (1999) 969–977.
[30] www.dwb.unl.edu/Teacher/NSF/C08/C08Links/www.iacr.bbsrc.ac.uk/notebook/courses/guide/dnast.htm.
[31] I. Dunham, N. Shimizu, B.A. Roe, S. Chissoe, The DNA sequence of human chromosome 22, Nature 402 (1999) 489–495.
[32] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier, L. Marcourt, Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences, J. Theoret. Biol. 206 (2000) 323–326.