AN EFFICIENT ANALYSIS TECHNIQUE FOR DNA SEQUENCES USING MULTIWINDOW GABOR REPRESENTATIONS

Nagesh K. Subbanna and Yehoshua Y. Zeevi

Department of Electrical Engineering,
Technion - Israel Institute of Technology,
Haifa - 32000, Israel
email: nagesh@tx.technion.ac.il, zeevi@ee.technion.ac.il

ABSTRACT

In this paper, we develop a new technique to store, search and compare DNA sequences. We utilize the concept of multiwindow Gabor representations and use these functions to handle the nucleotide sequences. We show specifically, that using multiwindow Gabor representations, it is possible to represent the sequences efficiently using very few terms. Further, we develop a search technique based on the correlation between the multiwindow coefficients of the query sequence and sequences in the database and show that our method has a smaller computational complexity. Most importantly, we show that using multiwindow Gabor representations, we can examine the periodicity properties of sequences very easily, without need to resorting to the string matching methods, the global Fourier techniques or the statistical correlation techniques.

1. INTRODUCTION

All too often, it has been noticed that the DNA sequences make huge databases. For instance, Genbank, a premier database, has a total of several million DNA sequences, not to mention its complement of amino acids, and proteins, among others. One of the necessities to be able to manage this database is to be able to store sequences efficiently, classify this huge database and be able to search a sequence of nucleotides, either as part of a larger sequence or as a stand alone sequence.

Databases store the DNA sequences as character strings of the nucleotide bases present in the sequence. Each nucleotide is one of the four distinct, possible types designated by the letter A, G, T, and C. A thorough note on the structure of the DNA molecule may be found in [7]. It is sufficient to mention here that it is a double helical structure with the two individual strands linked by complementary bases. A is complementary to T and C is complementary to G. The DNA sequence, in this paper, is thus a character string representing one of the two strands of the DNA. To apply signal processing methods, the character strings need to be mapped into the numerical sequences. To consider the DNA sequence as a signal, we first apply the method proposed by Anastassiou [7] that maps the character string into a numerical sequence. Initially, given a sequence of L nucleotides comprised of the alphabet of four letters, we represent the nucleotides A, T, G, and C by $1+j$, $1-j$, $-1+j$ and $-1-j$ respectively. These values have the advantage of correctly showing the complementary nature of the nucleotides (since A is complementary to T and G to C, the correct structure is preserved) in the double helical structure of the DNA molecule, being of the same absolute value and finally being exactly $\pi/2$ radians out of phase with each other. Using these values, we convert a sequence of nucleotides of length L into a discrete signal belonging to C^L.

Many techniques, among which the prominent ones include direct string comparison techniques like BLAST[1], Fourier analysis [7], pattern comparison using statistical tools like correlation [4], probabilistic techniques based on Hidden Markov models [5], among others, and finally dyadic wavelet techniques [3] have been applied in recent years for searching and extracting patterns. However, these techniques are all tailored to one particular use and do not combine the advantages afforded by the application of the multiwindow Gabor transform in the analysis of DNA sequences.

For the purpose of efficient representations of DNA sequences, we utilize discrete multiwindow Gabor transforms [2]. Proper detection of frequencies is directly related to the spread of the window function, i.e., correct recognition of low frequencies requires windows of large spread, while windows of narrow spread are required to recognize sharp changes (high frequencies) [6]. Keeping this in mind, we utilize multiwindow Gabor transforms that alleviate the problem of choice of window functions to a large extent and help cope with a greater margin of error in terms choice of lattice constants.

Since most of the researchers concerned with the structure of DNA and protein macromolecules are not familiar with the powerful signal processing techniques employed in our study, we first provide an overview of the multiwindow Gabor functions. We then apply the multiwindow Gabor frames to the sequences and develop a simple search technique in the combined space. Finally, we discuss the results obtained by our approach to DNA sequence analysis.

2. MULTIWINDOW GABOR FUNCTIONS

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 cast as an L periodic signal.

Given a discrete, finite signal $f[k]$, the multiwindow Gabor coefficients $c_{r,m,n}$, generated by the projection of the signal into the combined space, are given by

$$c_{r,m,n} = \sum_{k=0}^{L-1} f[k] {g_r}^*[k-na] e^{-j2\pi mbk/L}, \qquad (1)$$

[1]BLAST - Basic Local Alignment Search Tool - is a string based technique to compare DNA sequences, search and check patterns

where $g_r[\cdot], r \in 0,\ldots,R-1$ are the window functions chosen by the user, and form an over-complete basis for C^L, a and b are the fixed shifts of the tessellations along the time and the frequency axes respectively, and L is the periodicity of the signal. It is important to note that the multiwindow Gabor transform is unitary, and the set of coefficients uniquely determines a signal and a and b are chosen such that L is divisible by both a and b.

Given the coefficients $c_{r,m,n}$, the signal $f[k], k \in 0,\ldots,L-1$ is reconstructed by

$$f[k] = \sum_{r=0}^{R-1} \sum_{n=0}^{\overline{a}-1} \sum_{m=0}^{\overline{b}-1} c_{r,m,n} \gamma_r[k-na] e^{j2\pi mbk/L}, \qquad (2)$$

where $c_{r,m,n}$ are the coefficients, $\gamma_r[\cdot]$ is the r-th dual window function, R is the number of window functions, $\overline{a} = L/a \in \mathcal{N}$ and $\overline{b} = L/b \in \mathcal{N}$, and $\mathcal{N}$ is the set of natural numbers.

The necessary condition for the set of window functions $g_r[\cdot]$ to form a complete basis for C^L, and consequently for lossless reconstruction of the signal $f[k]$ from the set of coefficients $c_{r,m,n}$ is that $\overline{a}\overline{b}R \geq L$ [2]. If $\overline{a}\overline{b}R = L$, then the reconstruction exists, but is unstable and the localization properties are lost [1]. We shall, therefore, assume that $\overline{a}\overline{b}R > L$. Given the window functions $g_r[k]$, and the coefficients $c_{r,m,n}$, one has to compute the dual window function $\gamma[k]$.

In matrix form, equations (1) and (2) can be written as

$$\mathbf{c} = \mathbf{G}^* \mathbf{f}, \qquad (3)$$

and

$$\mathbf{f} = \mathbf{\Gamma} \mathbf{c}, \qquad (4)$$

respectively, where $\mathbf{f} = f[k], k \in 0,\ldots,L-1$ is a vector of length L, $\mathbf{c} = c_{r,m,n}$ is a vector of length $R\overline{a}\overline{b}$ and $\mathbf{G}$ is a matrix of dimensions $L \times R\overline{a}\overline{b}$ as shown below

$$\mathbf{G} = \begin{bmatrix} g_{0,0,0}[0] & \cdots & g_{R-1,\overline{a}-1,\overline{b}-1}[0] \\ g_{0,0,0}[1] & \cdots & g_{R-1,\overline{a}-1,\overline{b}-1}[1] \\ \vdots & \ddots & \vdots \\ g_{0,0,0}[L-1] & \cdots & g_{R-1,\overline{a}-1,\overline{b}-1}[L-1] \end{bmatrix}. \qquad (5)$$

Given the matrix $\mathbf{G}$, the inverse matrix $\mathbf{\Gamma}$, is computed by finding the pseudo-inverse of the matrix $\mathbf{G}$, [2]

$$\mathbf{\Gamma} = \mathbf{G}^\dagger = \mathbf{G}^*(\mathbf{G}\mathbf{G}^*)^{-1}. \qquad (6)$$

This solution corresponds to the least square solution to the problem. An efficient technique to compute $\mathbf{\Gamma}$ is given in [2]. Other duals have been proposed in [10] and [11], It is possible to find duals that optimize for different metrics rather than find the least square solution, without much greater computational effort. Consequently, we have a great degree of freedom with this choice of the analysis technique.

3. APPLICATION OF MULTIWINDOW GABOR TRANSFORMS TO DNA SEQUENCES

We principally project the sequence of length L into the two-dimensional Gaborian combined space and examine the coefficients. Fig.1 depicts the two-dimensional representation of the color coded value of the coefficients, with the x and y axes representing the number of sampling intervals shifted along the time and the frequency axes respectively. This two-dimensional representation of signals provides an easy way to assess, store and compare the similarity of signals, both visually and quantitatively. It can be observed in Fig.1, the representation of the coefficients of a signal of length 600, oversampled by a factor of 12.5, that there are very few coefficients that have a high value. This offers a new approach for compact storage of DNA/protein sequences. By choosing a proper set of lattice constants a, b and window functions $g_r[\cdot]$, it is possible to reduce the number of necessary coefficients very significantly. From the coefficients in Fig.1, the signal can be reconstructed to within 3% error with only 377 largest coefficients. This leads to a very efficient representation of signals. Since the alphabet is limited to only four possible values A, C, T, and G, we can approximate the sequence using very few coefficients. Indeed, we know from signal processing that the multiwindow Gabor scheme provides an efficient means for the sparsification and compression of non-stationary locally periodic sequences [1].

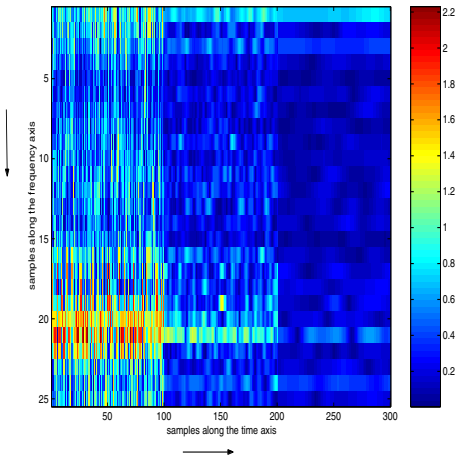

Figure 1: Coefficients of sequence AP000543 of length $L = 600$ using three normalized Gaussian window functions with effective widths $\sigma_1 = 4$, $\sigma_2 = 16$ and $\sigma_3 = 64$, with the combined space sampling parameters $a = 6, b = 12$

3.1 Algorithm for searching:

Utilizing the technique evolved in the last section for storing sequences compactly, we develop a technique to search and find patterns in sequences. It has been shown in [8] that correlations technique for recognizing whether a sequence matches a subsequence of another sequence can be effectively deployed. We utilize a variant of the technique to judge whether the query string is present in the string to which it is being compared.

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 (1)

$$c^{(1)}{}_{r,m,n} = \sum_{k=0}^{L_1-1} f_1[k] g_r[k-na] e^{-j2\pi mb_1 k/L_1}, \qquad (7)$$

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 time axis.

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

$$c^{(2)}{}_{r,m,n} = f_2[k] * g[k-na]e^{-j2\pi mb_2k/L_2}, \qquad (8)$$

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 time axis. It is important to note that in both the search and the query sequences, the values of a and $\overline{b}$ should be kept a constant. Changing these would be possible, but would require a more complicated approach to calculating the matches between the sequences than is indicated in this paper.

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

$$\kappa[p] = \sum_{n=p}^{p+\overline{a_2}-1} \sum_{r=0}^{R-1} \sum_{m=0}^{\overline{b}-1} c^{(1)}{}_{r,m,n} c^{(2)}{}_{r,m,n}. \qquad (9)$$

A high value of $\kappa[\cdot]$ indicates a match and a low value of $\kappa[\cdot]$ indicates a lack of matching. 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. A refinement would be to reduce the number of actual coefficients, using the thresholding suggested and normalizing the sequences. This would therefore, compare only corresponding high value coefficients in the two sequences and reduce the computational time for searching. Also, this technique would rule out false negatives, since similar sequences would have, at least reasonably high values in the corresponding positions. The only need to find nearly similar sequences would be to set a threshold for the correlation value $\kappa[\cdot]$, above which all sequences would be acceptable. Further, since slight differences in DNA sequences do not produce a lot of difference in the coefficients (which is the principal problem with global Fourier methods), the localization of variation in case of changes also helps in examining the periodicity properties in DNA sequences.

3.2 Pattern examination

There are two types of patterns of great importance in macromolecular sequence analysis viz, approximate repeats and reverse complements [12]. The multiwindow Gabor technique can easily detect both. The correlation technique above showed that it is very easy to detect approximate matches and periodicities. It is possible to reveal hidden patterns by looking directly at the coefficients in the combined space. The present technique of looking at the sequences in a two-dimensional representation allows us to appreciate patterns far more quickly and correctly. Since we have several windows used, and all the windows compute the coefficients at all frequencies, it is more effective than dyadic wavelets, that have been used in DNA sequence analysis [13]. The real advantage is in error tolerance in the choice of the lattice constants and the window functions.

Reverse complements can also be very easily detected using multiwindow Gabor techniques. To detect reverse complements, we simply reverse the sequence and apply the same

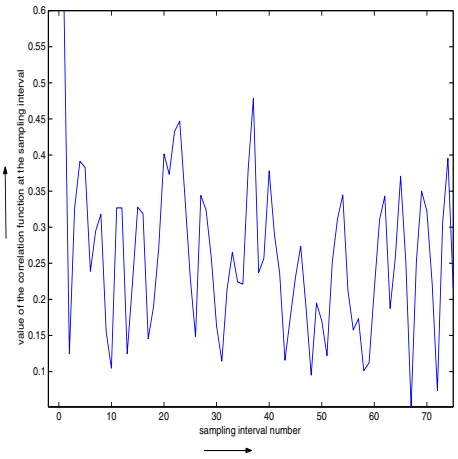

Figure 2: The correlation function of a 'slightly-altered' subsequence of length 75 of AP000543

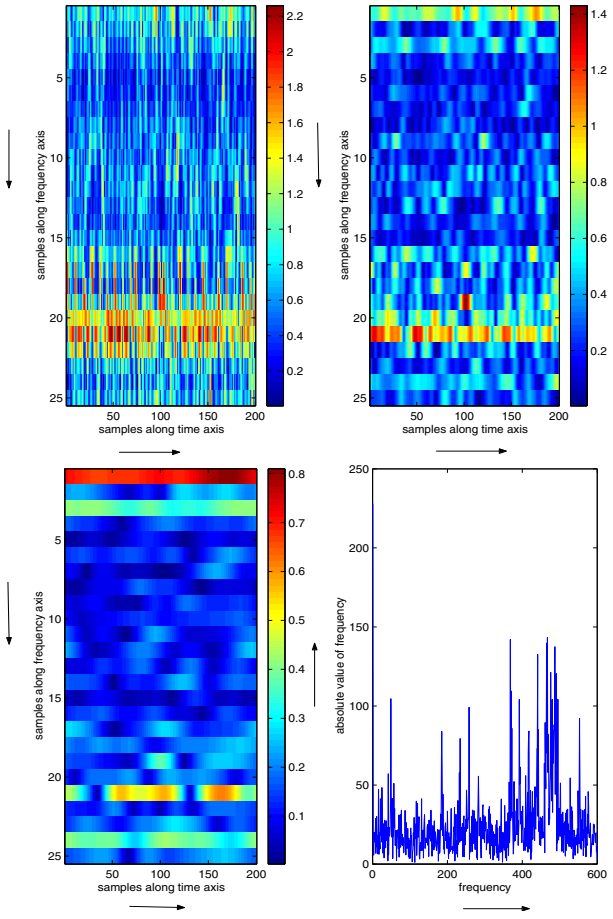

Figure 3: Coefficients of sequence AP000543 of length $L = 600$ using three normalized Gaussian window functions with effective widths $\sigma_1 = 3$, $\sigma_2 = 12$, and $\sigma_3 = 48$, with the combined space sampling parameters $a = 3, b = 24$, and the DFT of AP000543 of length 600.

multiwindow Gabor technique. The coefficients are given by (1). Setting the frequency to zero, we find that a sequence and its complement have the same real part when a symmetric Gaussian window is used. The proof is trivial.

4. RESULTS AND DISCUSSION

We have implemented the multiwindow Gabor representation of a subsequence of c:20h12 in the CES region of the chromosome 22 (Genbank accession number AP000543, [9] Dunham *et. al*). The sequence AP000543 is one of the 27 sequences mapping the Cat's Eye syndrome genes in the centrometric region of the chromosome 22.

In Fig.2, we see that we are able to locate a (slightly altered) subsequence even in a sequence where there are repetitive patterns. The subsequence comprised of the first 75 bases of the sequence AP000543 (with some small alterations in the subsequence of 75 to ensure that the technique was able to find slightly modified subsequences). The correlation with the place of coincidence is far greater than at other places, (nearly 1.33 times) the value at other places. It is also clear that the periodicity of the sequence seen in Fig.1 is reflected in the correlation values seen in Fig.2 since most of the other (apart from the exact match) correlations also have reasonably high values. This technique, thus, easily permits finding the corresponding 'near matches' in other sequences.

In Fig.3, we see that the repetition of the bases in the DNA sequence can be seen very clearly at frequency 20/25-22/25. The Fourier transform of the sequence also gives a peak in the region of (480/600), but there are other peaks in the region of (380-400)/460, which prevent us from recognizing the accurate frequency of the periodicity. In fact, it can be clearly seen in all the 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 (which studied the coefficients over a longer distance). This also seems to corroborate the conjecture that coding regions are very periodic, and our technique allows us to detect hidden periodicities. In [3], the authors try to find the long range correlations in the sequence with statistical techniques. Our method is much more direct, since it allows us to see it directly from the coefficients of the multiwindow Gabor transform.

It is important to mention that our preliminary investigations indicate that the multiwindow Gabor representations technique can also be applied to the analysis of amino acid and protein sequences.

5. ACKNOWLEDGMENT

The research was 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] M. Zibulski, and Y. Y. Zeevi, "Analysis of multiwindow Gabor-type schemes by frame methods", Applied and Computational Harmonic Analysis, Vol. 4(2), pp. 188-221, 1997.

[2] M. Zibulski, and Y. Y. Zeevi, "Discrete Multiwindow Gabor type transforms", IEEE Transactions on Signal Processing, Vol. 45(6), pp. 1428- 1442, June 1997.

[3] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier, and L. Marcourt, "Fourier and Wavelet Transform Analysis, a Tool for Visualizing Regular Patterns in DNA Sequences", Journal of Theoretical Biology, Vol 206, pp. 323-326, 2000.

[4] S. M. Ossadnik, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C. K. Peng, M. Simons, H. E. Stanley, "Correlation approach to identify coding regions in DNA sequences", Journal of Biophysics, Vol. 67(1), pp. 64-70, July 1994.

[5] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, "Biological sequential analysis: Probabilistic Models of proteins and nucleic acids", Cambridge University Press, 1998.

[6] S. G. Mallat, "A Wavelet tour of Signal Processing", Academic Press, 1999.

[7] D. Anastassiou, "Genomic Signal Processing", Signal Processing Magazine, Vol.18(4), pp. 8-20, September 2000.

[8] E. Sejdic, and J. Jiang, "Comparative study of three time frequency representations with applications to a novel correlation method", pp. 633-636, ICASSP-2004.

[9] I. Dunham, N. Shimizu, B. A. Roe, and S. Chissoe, "The DNA sequence of human chromosome 22", Nature, Vol. 402, pp. 489-495, 1999.

[10] I. Daubechies, H. Landau, and Z. Landau, "Gabor Time-Frequency lattices and the Wexler-Raz Identity", Journal of Fourier Analysis and Applications, Vol. 1(4), pp. 437-478, 1995.

[11] N. K. Subbanna, Y. C. Eldar, and Y. Y. Zeevi, "Oversampling of the Generalized Multiwindow Gabor Scheme", submitted to SampTA 05.

[12] T. Matsumoto, K. Sadakane, and H. Imai, "Biological sequence compression algorithms", 2001.

[13] K. B. Murray, D. Gorse, and J. M. Thornton, "Wavelet Transforms for the characterisation and detection of Repeating Motifs", Journal of Molecular Biology, Vol. 316, pp. 341-363, 2002.