An improved digital filtering technique using nucleotide frequency indicators for gene identification

Achuthsankar S. Nair, S.Sreenadhan

Abstract

The pivotal problem of Gene Identification in Eukaryotes is to locate exons. For that, a number of methods which utilize different types of coding measures and the signals at gene premises are in use. Combinations of these methods are more advisable than a single one as none of them has 100% selectivity and sensitivity. In this paper, two of the existing methods for finding exons from open reading frames using digital signal processing, namely the spectral analysis technique and the filtering technique, are reviewed. Then a method for improving the discriminating capability of digital filtering technique is presented.

We suggest that the existing filtering technique for locating exons can be improved if various coding statistics like single, dinucleotide and trinucleotide biases are incorporated. The model independent method then becomes model dependent and the selectivity and sensitivity are improved. Examples using Open Reading Frames (ORFs) of various organisms are given to illustrate this.

Index Terms

genomic signal processing, digital signal processing, filtering of gene sequences, gene finding, coding statistics.

I. Introduction

IN a genome, millions of nucleotides may be present but only less than 3% of them are part of genes, which are stretches of DNA that 'code' for proteins . And only very rarely the genes appear at a stretch from start to end. Usually genes are seen split and distributed in the genome. Coding regions in DNA are called exons and exons are separated by introns which are noncoding. Also in between the genes, lie intergenic regions, which are mostly not having any meaning. The area in a genomic sequence from the 'start' of a gene to the 'stop' of a gene which includes exons and introns is known as an Open Reading Frame (ORF). A large number of gene finding algorithms use probabilistic models like

Hidden Markov Models (HMM) [1]. They depend on gene models or on an a priori knowledge of coding statistics of the genome of the organism under consideration. They utilize coding measures like codon usage measures, oligonucleotide counts, aminoacid usage, codon preference, hexamer usage etc. Papers [2–4] contain reviews and assessments. On the contrary, digital processing techniques of gene finding are model independent. They utilize the general features of coding rather than the organism dependent measures. This paper focuses on the digital signal processing technique of gene finding and suggests a method by which coding statistics can be incorporated into it for improving the performance.

II. Genomic Signal Processing- A Review

The short range correlations, periodicities and patterns existing in genomic data naturally led to the use of digital signal processing (Fourier transform and wavelet analysis) and filtering techniques for identifying protein coding regions. In papers [5–20], various methods, reviews and applications can be found on the subject. Here we will have a short review of the important aspects of the technique which have direct relationship with the matters discussed in this paper.

A. A preliminary spectral measure for coding regions

For a DNA string x[n] of N characters (with an alphabet A,G,C and T) let us define four binary indicator sequences $u_A[n], u_G[n], u_C[n]$ and $u_T[n]$. Each indicator sequence has a 1 if the corresponding base exists at position n, otherwise a 0. For example, if x[n]=AATGCATCA, $u_A[n] = [110001001]$. Obviously, the sum of all binary indicators at any position n is 1 for all n.

$$u_A[n] + u_G[n] + u_C[n] + u_T[n] = 1 \quad for\ n = 0, 1, 2, \cdots N-1 \tag{1}$$

Let $U_A[k], U_G[k], U_C[k]$ and $U_T[k]$ be the Discrete Fourier Transforms (DFT) of the binary sequences, $u_A[n], u_G[n], u_C[n]$ and $u_T[n]$ respectively.

$$U_X[k] = \sum u_X[n] e^{(-j2\pi kn/N)},\ for\ X = A, G, C\ or\ T\ and\ k = 0, 1, 2, \cdots N-1 \tag{2}$$

The total spectral content at k is,

$$S[k] = \sum \mid U_X[k] \mid^2 \tag{3}$$

S[k] may be used as a preliminary indicator of a coding region as a plot of S[k] against n reveals a peak at k=N/3 for a coding region and shows no such peak for a noncoding region [5–7].

B. Digital Filtering Techniques for Gene Finding

The DFT method of finding genes as described above may be viewed essentially as a filtering technique [9–11]. As N corresponds to 2π, the period 3 components may be isolated by filtering the sequence through a bandpass filter H(z) with the pass band centred around $2\pi/3$. If we give $u_X[n], (X = A, G, C\, or\, T)$ as an input to H(z), the output $y_X[n], (X = A, G, C\, or\, T)$ will have peaks at coding regions as $u_X[n]$ has period 3 components and the filter has a passband around $2\pi/3$.

The total output is

$$y[n] = \sum \mid y_X[n] \mid^2 \tag{4}$$

This can be utilized for finding the exons in an ORF. H(z) can be implemented as an antinotch filter.

C. An antinotch filter for gene identification

Consider a second order all pass filter

$$A(z) = \frac{R^2 - 2Rcos\theta z^{-1} + z^{-2}}{1 - 2Rcos\theta z^{-1} + R^2 z^{-2}} \tag{5}$$

with poles at $Re^{\pm j\theta}$ and zeros at $1/Re^{\pm j\theta}$. Also consider a filter bank with G(z) and H(z) defined as

$$G(z) = 0.5[1 + A(z)] \tag{6}$$

$$H(z) = 0.5[1 - A(z)] \tag{7}$$

From 5 and 6

$$G(z) = \frac{(1 + R^2)(1 - 2cos\omega_0 z^{-1} + z^{-2})}{2(1 - 2Rcos\theta z^{-1} + R^2 z^{-2})} \tag{8}$$

where $cos\omega_0 = 2Rcos\theta/(1 + R^2)$

It is evident that G(z) is a notch filter with a zero at the frequency ω_0. For stability, R should be less than 1. It is clear that as the pole radius R gets close to unit circle, ω_0 gets close to θ. So, at any frequency sufficiently away from ω_0 the contribution of zero and pole are almost same and G(z) has unity gain. It can be easily verified that G(z) and H(z) are power complementary. So, H(z) is a good antinotch filter which can be used to identify exons.

D. Realisation of H(z)

Substituting the value of A(z) from 5 and noting that $cos\omega 0 = cos2\pi/3$, the transfer function of the antinotch filter is,

$$H(z) = \frac{(1-R^2)(1-z^{-2})}{2(1+\frac{1+R^2}{2}z^{-1}+R^2z^{-2})} \tag{9}$$

which may be realised in direct form II or in lattice structure form.

III. Incorporating nucleotide biases.

We propose a modification to the filtering technique reported in [9–11] through which its performance can be further improved. The basic filtering method is a model independent one as we do not make use of any kind of the coding/noncoding statistics derived from the genome. There We depend only on the period three component of the spectra of the coding regions which are higher compared to those of the noncoding regions. Now if we incorporate coding statistics in terms of nucleotide biases in exons and introns to the filtering technique, the discriminative power of the method can be enhanced. We will incorporate the statistics in three ways (numerous ways are possible) and the method becomes model dependent.

A. Using single nucleotide bias

Let us find the average value of the normalized frequencies of nucleotides, A, G, C & T in coding regions, and noncoding regions and take the ratios. That is, if $f_e(A), f_e(G), f_e(C)\&f_e(T)$ are the frequencies of A, G, C & T normalised to the length of exons and $f_i(A), f_i(G), f_i(C)\&f_i(T)$ are the corresponding ones for introns, of the same genome, we may define (nucleotide) relative frequency indicator,

$$rf(X) = f_e(X)/f_i(X), X = A, G, C \, or \, T. \tag{10}$$

This can replace the *ones* in the binary indicator sequences in the original technique of filtering. The sequences may now be renamed as *probability indicator sequences*. As an example, if the sequence is

$$x[n] = AATGCATCA \tag{11}$$

and the relative frequency indicators are rf(A)=0.19, rf(G)=0.2, rf(C)=0.27 and rf(T)=0.36, probability indicator sequences are,

$p_A[n] = [0.19\ 0.19\ 0\ 0\ 0\ 0.19\ 0\ 0\ 0.19]$

TABLE I

AN EXAMPLE OF DINUCLEOTIDE BIAS TABLE

	A	G	C	T
A	0.7	0.9	0.7	0.8
G	1.2	0.7	2.3	0.8
C	1.1	0.8	2.4	0.9
T	0.9	1.2	0.8	2.3

$p_G[n] = [0\ 0\ 0\ 0.2\ 0\ 0\ 0\ 0\ 0]$

$p_C[n] = [0\ 0\ 0\ 0\ 0.27\ 0\ 0\ 0.27\ 0]$

$p_T[n] = [0\ 0\ 0.36\ 0\ 0\ 0\ 0.36\ 0\ 0]$

Now let the outputs of the antinotch filter for the inputs $p_X[n]$, X=A,G,C or T be $y_X[n]$ then,

$$y[n] = \sum |y_X[n]|^2 \cdots \tag{12}$$

y[n] plotted against nucleotide positions of the given sequence is a better predictor.

B. Using dinucleotide bias

We may form two (4,4) matrices $f_e(4,4)$ for exon regions and $f_i(4,4)$ for intron regions where the element (i,j) of these matrices is the probability of the nucleotide i occurring after nucleotide j. A dinucleotide probability indicator matrix $f_d(4,4)$ may be formed by dividing $f_e(4,4)$ by $f_i(4,4)$ element by element. The elements of f_d may now be used to form the indicator sequences. For example let f_d is as given in Table(I) Then the new Probability Indicator Sequences for 11 are

$p_A[n] = [0.19\ 0.7\ 0\ 0\ 0\ 0.7\ 0\ 0\ 0.7]$

$p_G[n] = [0\ 0\ 0\ 0.8\ 0\ 0\ 0\ 0\ 0]$

$p_C[n] = [0\ 0\ 0\ 0\ 2.3\ 0\ 0\ 0.9\ 0]$

$p_T[n] = [0\ 0\ 0.9\ 0\ 0\ 0\ 0.9\ 0\ 0]$

The new $p_X[n]$ may be filtered as in the previous case for locating genes.

C. Using trinucleotide bias

Also we may form two (4,16) matrices $f_e(4,16)$ and $f_i(4,16)$ with 4 rows for A,G,C &T and 16 columns for 16 nucleotide pairs, for exon and intron regions respectively. An element $f(i,jk)$ of these

TABLE II

A FORMAT FOR TRINUCLEOTIDE BIAS TABLE

	AA	AG	AC	AT	GA	GG	GC	GT	CA	CG	CC	CT	TA	TG	TC	TT
A																
G																
C																
T																

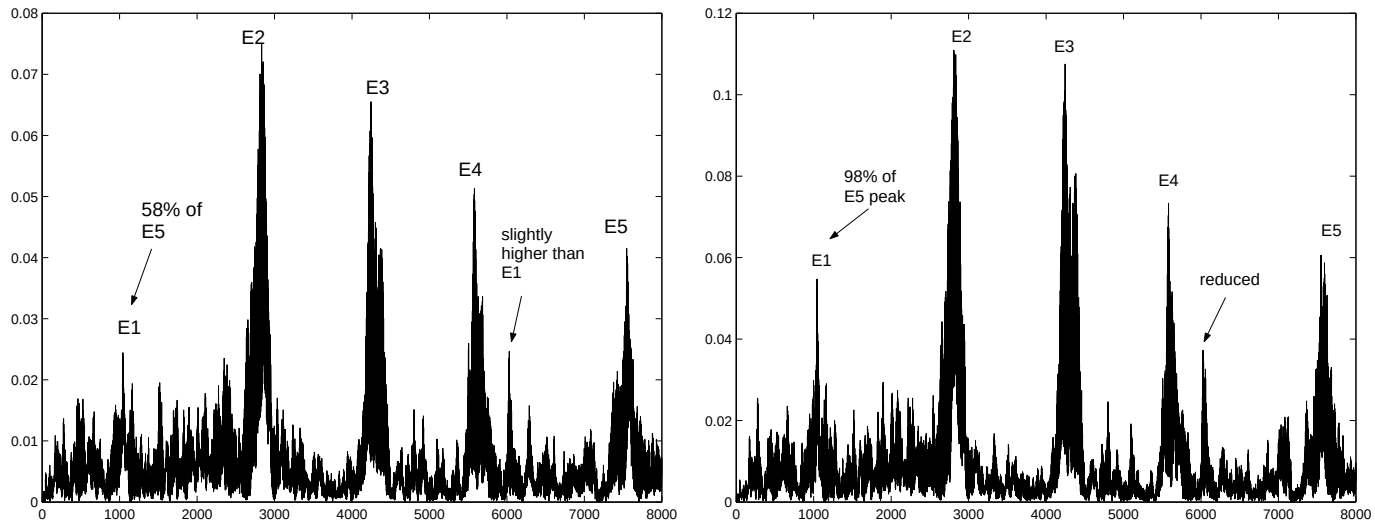
matrices is the probability of the nucleotide i following the nucleotide pair jk. It is formed by counting the occurrences of all nucleotides in exon regions and intron regions respectively and normalising them to the length. A trinucleotide probability indicator matrix f_t may be formed by dividing $f_e(4,16)$ by $f_i(4,16)$ element by element. (A format for trinucleotide bias table is shown as Table II). The elements of f_t may now be used to form the probability indicator sequences $p_X[n]$ and may be utilized as in the previous cases for locating genes.

D. Reason for improvement

The improvement shown with the incorporation of coding statistics, may be explained as follows. In this method, those nucleotides, which are more in an exonic area, compared to an intronic area get a relatively higher weightage. By placing the relative frequency indicators in the probability indicator sequences at the corresponding positions in which the nucleotides exist, we take into account the nucleotide bias difference in exons and introns and this makes the differentiation better. The dinucleotide and trinucleotide matrices bring in the nucleotide dependency differences among exons and introns. The authors found that the use of normalized frequencies of nucleotides in exons alone by taking a random distribution of nucleotides in introns (that is to use a background probability of 0.25) is not sufficient enough to give a better performance.

IV. RESULTS

Here we compare the results obtained by applying the model independent filtering technique (MIFT) using binary indicator sequences and the model dependent filtering technique (MDFT) using probability indicator sequences. For finding the indicator sequences, the counts in exons and introns inside the genes are taken. These may be replaced by average values from the whole chromosome/genome of interest. Shown below is a comparison of the results of filtering technique with binary sequence indicators and with single nucleotide bias. The genes used for experimentation are taken from the data set constructed



(a) with MIFT

(b) With MDFT

Fig. 1. Spectrum of F56F114a

by Burset and Guigo.

1.Gene F56F114a (a gene from C.elegans)

Referring to Fig.(1),

a. MIFT: All the five exons of the gene are distinguishable but E1 peak is small and is only 58% of E5 (exon with the next higher peak) peak. Also there is a peak at position 6031, which is slightly higher than E1, which may lead to a wrong conclusion that there exists an exon around 6031.

b. MDFT: E1 peak is increased to 98% of E5. Also the peak at 6031 is now only 58% of E1 and so the doubt of an exon around 6031 doesn't arise.

2. Gene HUMBETGLOA (human beta globin A)

Refering to Fig.(2),

a. MIFT: along with the 3 exons, a false one appears around position 500, which has a peak greater than that of E1.

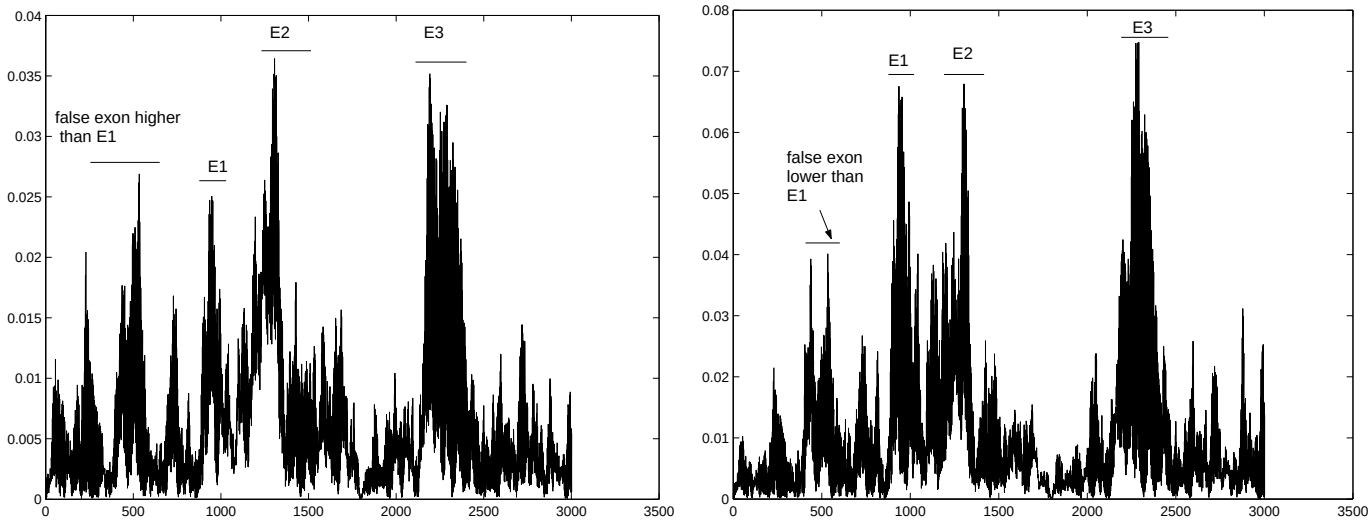
b. MDFT: The false exon peak has reduced to just 58% of E1 and now only the three (true) exons are visible.

3. Gene MMHOX13 (Mouse Hox 1.3 gene)

Refering to Fig.(3),

a. MIFT: a false exon with a greater peak than E2 is seen around position 2207

b. MDFT: the maximum strength at position 2207 (false exon region) is lesser than (67% of) the peak at



(a) with MIFT (b) With MDFT

Fig. 2. Spectrum of HUMBETGLOBA

E2.

4. Gene HUMCBRG (homosapiens carbonyl reductase gene)

Refering to Fig.(4),

a. MIFT: three false exons with peaks greater than E2 are seen around positions 1600, 2100 and 2500.

b. MDFT: the maximum strength around 1600, 2100 and 2500 (false exon regions) are lesser than the peak at E2.

5. Gene HUMMIF (homosapiens macrophage migration inhibitory factor)

Refering to Fig.(5),

a. MIFT: a false exon with a greater peak than E3 is seen at position 960

b. MDFT: the maximum strength at position 959 (false exon region) is lesser than the peak at E3.

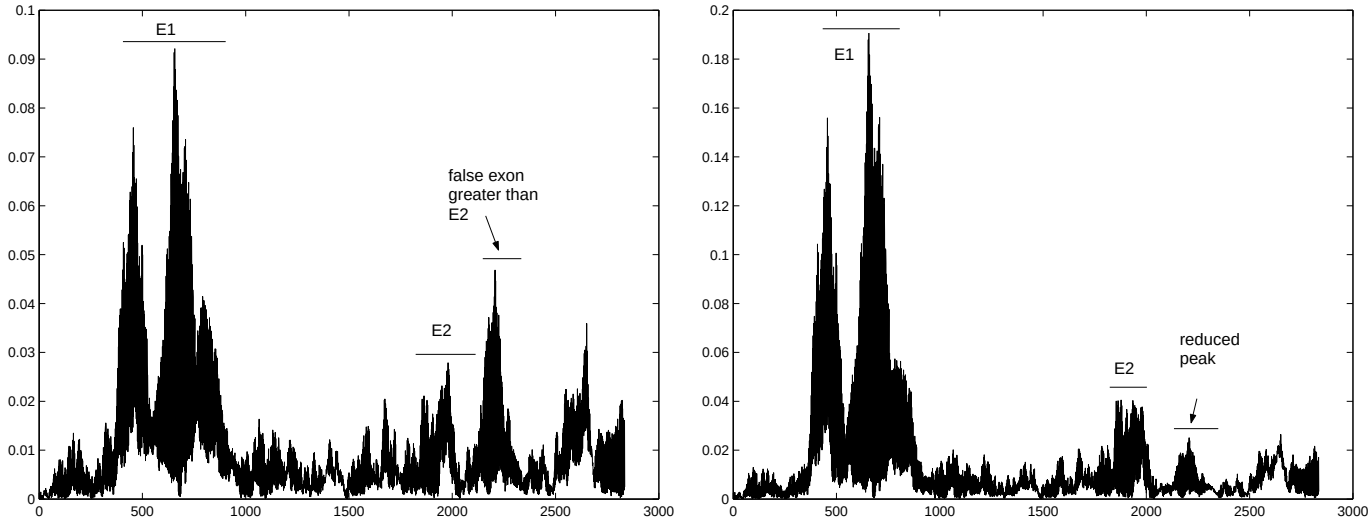
6.Gene HSODF2 (homosapiens ODF2 gene)

Refering to Fig.(6),

a. MIFT: a false exon with a peak of strength 75% of E1 is seen around position 726

b. MDFT: the maximum strength at position 726 (false exon region) is now reduced to 40% of the peak at E3.

TableIII summarizes the experimental results by comparing the peaks obtained in exon and intron regions using binary indicator sequences and sequences with single, dinucleotide and trinucleotide biases. It can be noted that the peaks in exon regions are boosted up and the peaks in the intron regions are reduced



(a) with MIFT (b) With MDFT

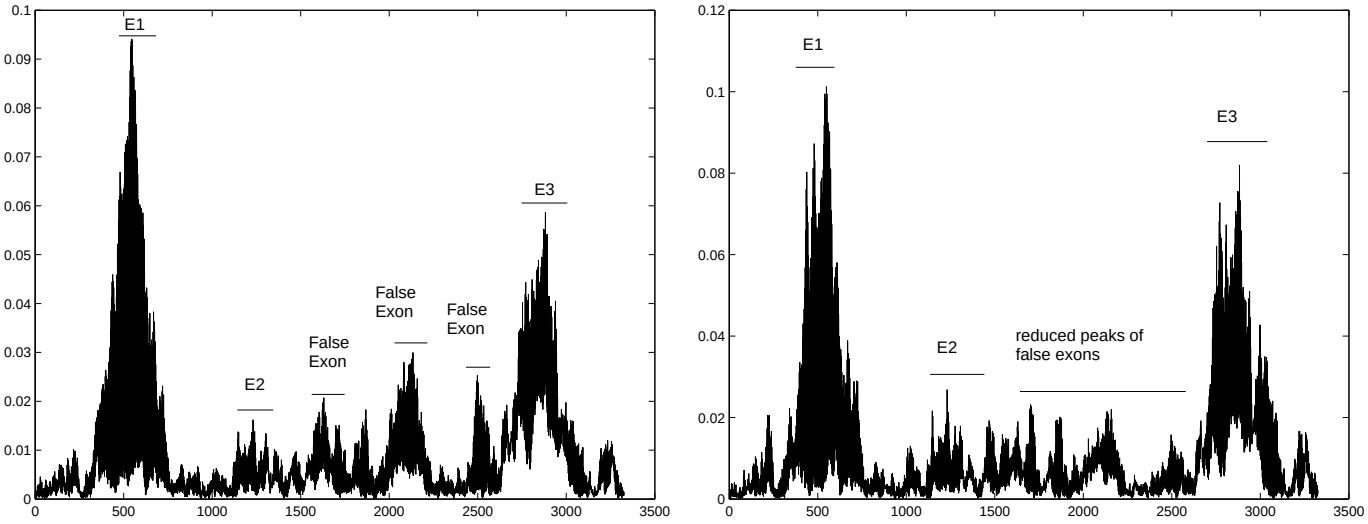
Fig. 3. Spectrum of MMHOX13

thus avoiding the detection of false exons in intron regions in most of the cases when new technique is applied.

If the coding statistics in the chromosome/genome is unknown for a sequence whose exons are to be located the above mentioned results may be used as follows. Initially MIFT may be used to locate the more prominent exons in the sequence as the method does not need any coding statistics. The frequencies of nucleotides in the exons and introns thus obtained may be utilized to build a coding statistics. Using that MDFT may be performed for better discrimination and to identify the exons not visible by MIFT. Improving the reliability of coding statistics through iterations may also be considered. Finally, Genomic signals (start and stop codons, exon-intron boundary patterns, promoter sequences,poly-A sequence etc) may the be used to find the exact nucleotide positions of the exons around the peaks found by the method described above.

V. Conclusion

We have incorporated coding statistics into the filtering technique for locating exons, in the form of single nucleotide, dinucleotide and trinucleotide biases and have found that it improves the power of discrimination of the technique substantially. But the authors have also found many cases where the results are not very much appreciable. So incorporating other coding measures into DSP techniques can also be experimented with.



(a) with MIFT (b) With MDFT

Fig. 4. Spectrum of HUMCBRG

REFERENCES

[1] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. *Biological Sequence Analysis.*, Cambridge University Press, 1998.

[2] J.M. Claverie, *Computational methods for the identification of genes in vertebrate genomic sequences,* Human Molecular Genetics, Vol. 6, No. 10 1735-1744,1997

[3] Fickett, J. W., and Tung, C. -S., *Assessment of protein coding measures,* Nucleic Acids Research. 20(24): pp6441-6450,1992

[4] Fickett,J.W.,*Finding genes by computer: the state of the art.* Trends Genet. 12, pp316-320.,1996

[5] Silverman, B. D., and Linsker R, *A Measure of DNA Periodicity.* J. theor. Biol., 118, pp295-300.1986

[6] S.Tiwari, S.Ramachandran, A.Bhattacharya, S.Bhattacharya and R.Ramaswamy,*Prediction of probable genes by Fourier analysis of genomic sequences,*CABIOS,vol.13,no.3,pp.263 -270,1997.

[7] D.Anastassiou,*Genomic signal processing,* IEEE Signal Processing Magazine,pp.8 -20,July 2001.

[8] Eivind Coward,*Equivalance of two Fourier methods for biological sequences,* Math.Biol,36:pp64-70,1997

[9] P.P.Vaidyanathan and B-J.Yoon,*Digital filters for gene prediction applications,* IEEE Asilomar Conference on Signals, Systems, and Computers,Monterey, CA, Nov.2002.

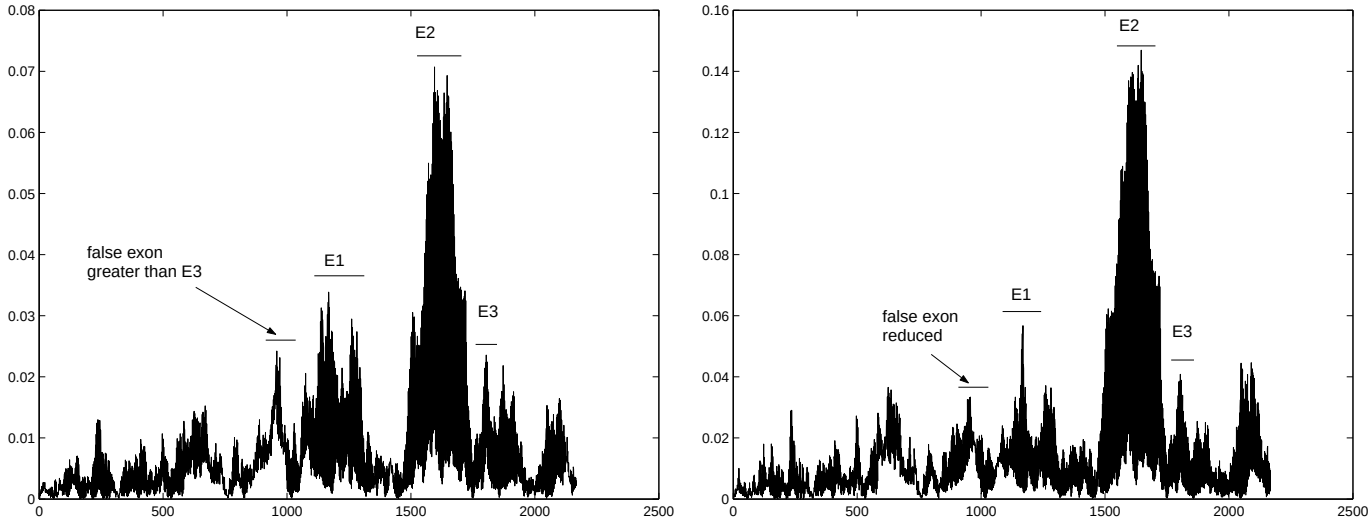
[10] P.P.Vaidyanathan and Byung-Jun Yoon,*The role of signal-processing concepts in genomics and proteomics,* Journal of the Franklin Institute, special issue on Genomics, 2004.

[11] P.P.Vaidyanathan and B-J.Yoon,*Gene and exon prediction using all pass filters,* Workshop on Genomic Sig. Proc. and Stat., Raleigh,NC, Oct.20002

[12] A.A.Tsonis et al,*Periodicity in DNA coding sequences:Implications in gene evolution,* J.Theor.Biol,151,pp323-331,1991

[13] W.Wang and D.H. Johnson,*Computing linear transforms of symbolic signals.,* IEEE Trans. Signal Processing, pp628-634,March 2002

[14] Vera Afreixo et al, *Spectrum and symbol distribution of nucleotide sequences,* Phy.Rev.E,Vol 70,no3,031910,Sept.2004



(a) with MIFT (b) With MDFT

Fig. 5. Spectrum of HUMMIF

[15] Vera Afreixo et al,*Fourier Analysis of symbolic data, A brief Review*, Digital Signal Process, Vol 14,no6,pp524-530,Nov 2004

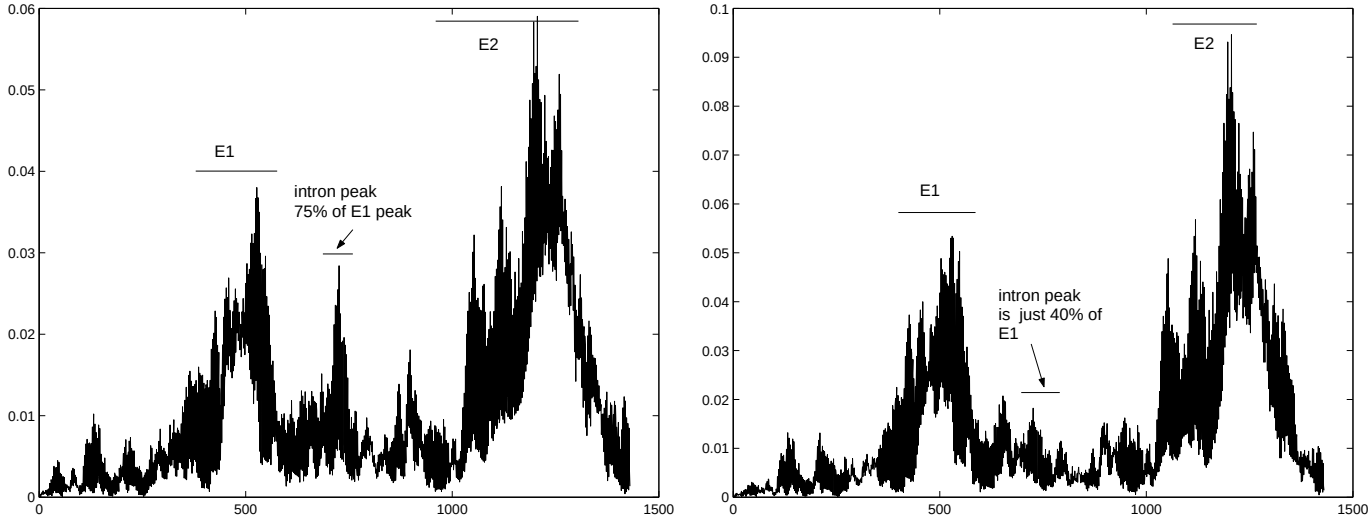
[16] Xin-Yu zhang,*Signal processing techniques in Genomic Engineering*, Proceedings of IEEE, Vol90 No 12, pp1822-1832,Dec 2002

[17] Biju Issac et al, *Locating probable genes using Fourier Transform approach*, Bioinformatics, Vol18,no 1, pp 196-197, 2002

[18] Guy Dodin et al,*Fourier and Wavelet Transform Analysis, a tool for visualising regular patterns in DNA sequences*,J.Theor.Biology,206, pp3230326,2000

[19] Ming Yan et al,*A new Fourier Transform approach for protein coding measure based on the format of Z curves*, Bioinformatics, Vol 14, no 8, pp 685-690, 1998

[20] Pietro Lio, *Wavelets in bioinformatics and computational biology: state of art and perspectives*, Bioinformatics Review, Vol 19 no1, pp2-9, 2003



(a) with MIFT

(b) With MDFT

Fig. 6. Spectrum of HSODF2

TABLE III

A COMPARISON OF RESULTS

gene	segment	region	binary	single bias	dinu.bias	trinu.bias
F56F114a	929-1135	exon	0.0244	0.0547	0.0524	0.0664
	2528-2857	exon	0.0750	0.1109	0.1402	0.2227
	4114-4377	exon	0.0655	0.1075	0.1488	0.2158
	5465-5644	exon	0.0514	0.0733	0.1027	0.1052
	7255-7605	exon	0.0415	0.0606	0.0827	0.1294
	6000-6100	intron	0.0246	0.0373	0.0341	0.0397
humbetgloa	866-957	exon	0.0250	0.0560	0.0867	0.0964
	1088-1310	exon	0.0364	0 .0577	0.1203	0.1539
	2161-2289	exon	0.0352	0.0630	0.1176	0.1411
	450-600	intron	0.0269	0.0344	0.0466	0.0390
	2850-2900	intron	no peak	no peak	0.0604	no peak
mmhox13	238-799	exon	0.0922	0.1531	0.1839	0.2473
	1760-2010	exon	0.0279	0.0329	0.0397	0.0406
	2100-2300	intron	0.0468	0.0292	0.0349	0.0436
humcbrg	277-566	exon	0.0941	0.0958	0.1058	0.1538
	1112-1219	exon	0.0162	0.0232	0.0285	0.0401
	2608-3044	exon	0.0587	0.0803	0.0840	0.1148
	1600-1650	intron	0.0207	0.0167	0.0150	0.0269
	2100-2150	intron	0.0300	0.0206	0.0244	0.0369
	2450-2500	intron	0.0253	0.0177	0.1360	0.0186
hummif	1173-1280	exon	0.0339	0.0451	0.0631	0.0741
	1470-1642	exon	0.0707	0.0965	0.1264	0.0741
	1738-1804	exon	0.0238	0.0332	0.0275	0.0741
	940-1000	intron	0.0242	0.0264	0.0415	0.0668
hsodf2	280-599	exon	0.0380	0.0464	0.1055	0.1364
	843-1275	exon	0.0590	0.0840	0.1960	0.4637
	700-750	intron	0.0284	0.0200	0.0269	0.0467