Comparison of Gene and Exon Prediction Techniques for Detection of Short Coding Regions

Mahmood Akhtar

School of Electrical Engineering and Telecommunications
The University of New South Wales, Sydney 2052,
Australia
mahmood.akhtar@student.unsw.edu.au

Abstract

Background: The segments of DNA molecule, called genes are known to carry useful information in their protein coding regions (exons) and are responsible for protein synthesis. In eukaryotes, exon regions are separated by non-coding regions (introns), whereas in procaryotes these regions are continuous. Accurate prediction of exon regions is a research problem currently being addressed.
Methods: Various methods have been used to automatically distinguish exons from introns in a DNA sequence; however these have been predominantly 'frequency' domain techniques. Two new 'time' domain techniques, the Average Magnitude Difference Function (AMDF) and Time Domain Periodogram (TDP), for gene and exon prediction have recently been proposed.
Results: I present a detailed comparison of time-domain and frequency-domain techniques for the detection of both short and long coding regions that are both closely and widely spaced. Rather than performing classification, the features of the various techniques are compared using the receiver operating characteristic (ROC) curve.
Conclusions: Recently proposed time-domain techniques for exon region prediction provide superior properties for the separation of exon and intron regions as compared with their frequency-domain counterparts. Further, these time domain techniques have the additional advantage of more accurate exon region detection with smaller frame size, so that sequences with short and/or closely spaced coding regions can be predicted very easily.

Keyword: Exons; Introns; Frequency-domain; Time-domain; Receiver Operating Characteristic (ROC) curve; Auto-regressive (AR); AMDF; TDP; Peak Ratio; Area Under the Curve (AUC).

I. Introduction

Segments of the DNA molecule referred to as genes are known to carry useful information, and are responsible for protein synthesis [1]. The understanding of the nature of this information and its role in determining the particular function encoded by the gene is a research problem currently being addressed. A key step towards the solution of this problem is to find exon regions in eukaryotes. In eukaryotes, exon regions are separated by introns, whereas in procaryotes these regions are continuous.

Tsonis *et al*. [2] employed Fourier analysis of DNA coding and non-coding sequences in an attempt to identify possible patterns in gene sequences. They found that while intronic sequences show a rather random pattern, coding sequences show periodicities and in particular a periodicity of three. In order to completely understand this periodicity of three let us consider the periodic sequence: A-- A-- A-- A-- … where blanks can be filled randomly by A, T, C or G (four types of bases of a DNA

sequence). This sequence shows a periodicity of three because of the repetition of the base A. Trifonov [3] suggested that one of the initial causes of periodicity could be the universal $(RNY)_n$ pattern (R= A or G, Y= C or T, N= any base). Eskesen *et al*. [4] concluded that DNA periodicity in exons is determined by codon (triplet of three base pairs) usage frequency.

However, many techniques for gene and exon prediction based on period-3 periodicity have been used and shown to be successful. Anastassiou [5, 6] presented an optimized spectral content measure based on windowing DFT for exon detection. Vaidyanathan and Yoon [7] proposed the use of IIR anti-notch filter. Rao and Shepherd [8] proposed the use of Auto-regressive (AR) model for detection of 3-periodicity for small DNA sequences. However discussed techniques are frequency-domain techniques.

Recently, Akhtar *et al* [9] have proposed two time-domain techniques for gene and exon prediction. In this paper, I present an overview of these and selected previous gene and exon prediction techniques, along with a comparison based upon ROC curves. It is important to note that this paper concentrates on evaluating the discrimination capabilities of the various features employed by recent exon prediction techniques, while the actual classification performance will be explored in future research.

II. 'Time' and 'Frequency' Domain Approaches

In order to apply digital signal processing (DSP) techniques, the character sequences of DNA are first converted into numeric sequences. For example, for a DNA sequence $x[n]$ =ATGCAAGTTCGA… the binary indicator sequence for each base type would look like:

$$x_A[n] = 100011000001\ldots$$
$$x_T[n] = 010000011000\ldots$$
$$x_C[n] = 000100000100\ldots$$
$$x_G[n] = 001000100010\ldots$$

where n represents the base number. It is clear that sequence 111111111111… is obtained by adding all four of binary indicator sequences. Different types of indicator sequences (binary, complex, weighted complex etc) can be used.

A. Frequency Domain Techniques

The notion of frequency can be applied to DNA sequences in the sense that portions of the sequence can recur regularly at a particular frequency (especially during coding regions). This frequency of recurrence can be exploited using techniques such as the DFT, anti-notch filtering and AR methods.

Sliding Window DFT Method. The DFTs $X_A[k]$, $X_T[k]$, $X_C[k]$ and $X_G[k]$ for the above binary indicator sequences can be obtained. The periodicity of period 3 in protein coding regions of a DNA sequence suggests that the value of DFT coefficients corresponding to k = N / 3 should be large. Thus if we take the window size N to be sufficiently large (e.g. 351) and a multiple of three, we can observe a peak at the sample value k = N / 3. Calculation of DFT at this single point (k = N / 3) is sufficient. The window can then be moved by sliding one or more bases.
The plot of:

$$S[k] = \sum_m \left| X_m[k] \right|^2 \quad (m = A, T, C \text{ and } G) \qquad (1)$$

has been used as a measure of the total spectral content of DNA character strings by Tiwari et at. [10], and later with some optimization, by Anastassiou [5, 6].

Auto-regressive (AR) Method. The AR method is an alternative to calculate the spectrum of signals. It can be described as an all-pole digital filter:

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \qquad (2)$$

where a_k are the coefficients of the order p model. The power spectrum estimation expression for the AR model given in (3) can be used for the detection of protein coding regions, especially in small DNA sequences [8],

$$\hat{P}_y(k) = \frac{\sigma^2}{\left|1 + \sum_{r=1}^{p} a_k W_N^{-kr}\right|^2} \qquad (3)$$

where σ^2 is the variance of the input, $W_N = e^{-j(2\pi/N)}$, and k represent the power spectrum points.

B. Time Domain Techniques

The notion of time can be applied to DNA in a similar sense to frequency, in that the objective is to detect periodic repetitions in the DNA sequence. The approach that has been taken to applying time-domain algorithms to DNA sequences can be seen in the block diagram of Fig. 1. The sequence is first converted to four binary indicator sequences. The binary indicator sequences cannot be processed by the full range of traditional digital signal processing techniques until they are converted into numerical values. This can be achieved by using a second-order resonant filter with centre frequency set to $2\pi/3$ (similar to Vaidyanathan and Yoon [7], as shown in Fig. 2), which emphasizes the period-three behaviour of the DNA sequence, but de-emphasizes the other components.

The AMDF Algorithm. The Average Magnitude Difference Function (AMDF) is a speech processing algorithm (used in pitch determination) that can be applied to any periodic or near-periodic sequence of length greater than the period to derive an estimate of the periodicity. Practically, the AMDF function will produce a deep null if correlation exists at period 3. Because the signal has been band pass filtered before applying the AMDF, small values found throughout the non-coding and coding regions. By contrast, if a different period is tested, it will produce relatively low AMDF values during the non-coding regions (where there is relatively more non-period 3 behavior) and extremely large AMDF values during the coding regions (where the behavior is entirely period 3).

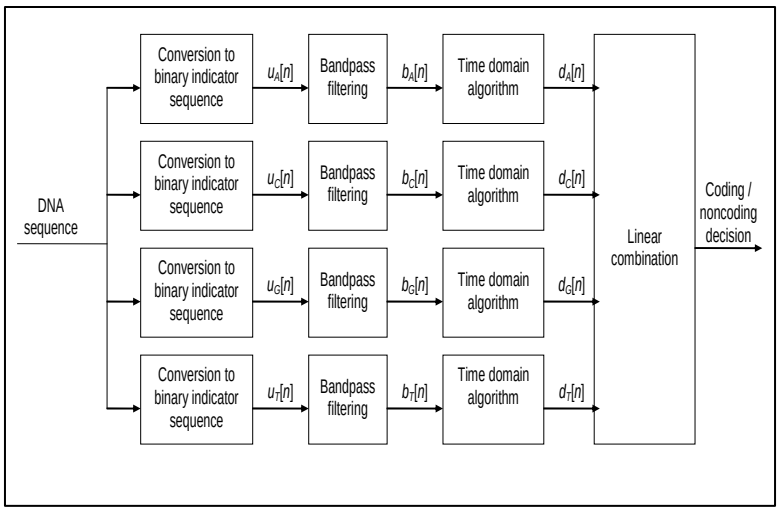

Fig. 1: Block diagram for time-domain algorithms

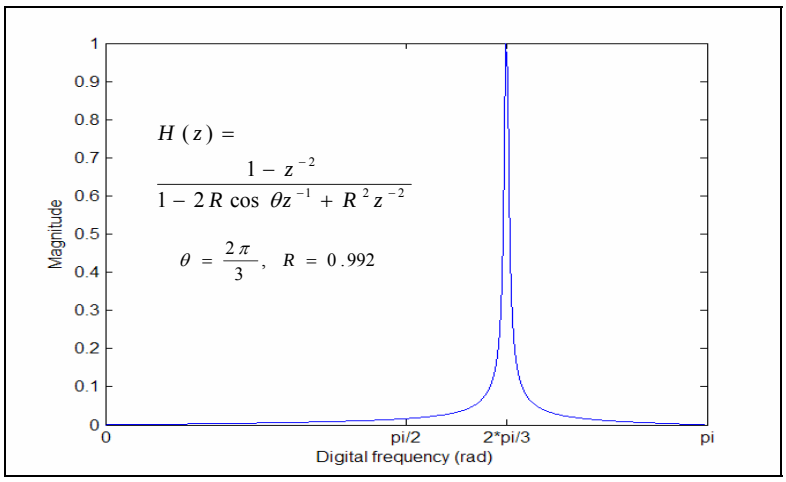

Fig. 2: Second-order resonant filter with centre frequency set to $2\pi / 3$

The TDP Algorithm. The Time Domain Periodogram (TDP) is an algorithm used for periodicity detection in sunspots and pitch detection for speech processing [11]. In contrast to the AMDF, the data are summed rather than taking differences. The data are arranged in a matrix, with rows containing sub-sequences of length equal to the period (k) being tested. The columns are then summed, and the maximum and minimum of the resulting vector are then used to derive the final estimate of the degree of periodicity at period k.

Practically, TDP algorithm will produce a peak if correlation exists at period $k = 3$.

III. Evaluation Methods

A. Receiver Operating Characteristic

In order to measure and compare the efficacy of all these techniques I have used receiver operating characteristic (ROC) curves. ROC curves were developed in the 1950's as a by-product of research into making sense of radio signals contaminated by noise and have been subsequently widely applied in a range of different applications. Considering the problem of exon and intron separation in a DNA sequence, the terms, 'TPF (True Positive Fraction)', 'FPF (False Positive Fraction)', 'FNP (False Negative Fraction)', 'TNF (True Negative Fraction)', can be defined, with the aid of Fig. 3:

- TPF: Truly predicted as coding region (exon) is part of this fraction.
- FPF: Falsely predicted as coding region (exon) is part of this fraction.
- FNF: Falsely predicted as non-coding region (intron) is part of this fraction.
- TNF: Truly predicted as non-coding region (intron) is part of this fraction.

If we assume the sum of all actual coding region points is equal to 1, then from above definitions, TPF + FNF = 1. Similarly for non-coding region points, TNF + FPF = 1. An ROC curve allows an exploration of the effect on TPF and FPF as the position of an arbitrary decision threshold is varied, and is a plot of the TPF as a function of FPF of an exon and intron separation method for varying

decision threshold values. It can be observed that if our decision threshold is very high, then there will be almost no false positives, but we won't really identify many true positives either.
The closer the ROC curve is to a diagonal, the less useful the test is at discriminating between exon and intron of a DNA sequence. The more steeply the curve moves up and then (only later) across, the better the test. A more precise way of characterizing this is to calculate the area under the receiver operating characteristics curve (AUC); the closer the area is to 0.5, the poorer the test, and the closer it is to 1.0, the better the test.

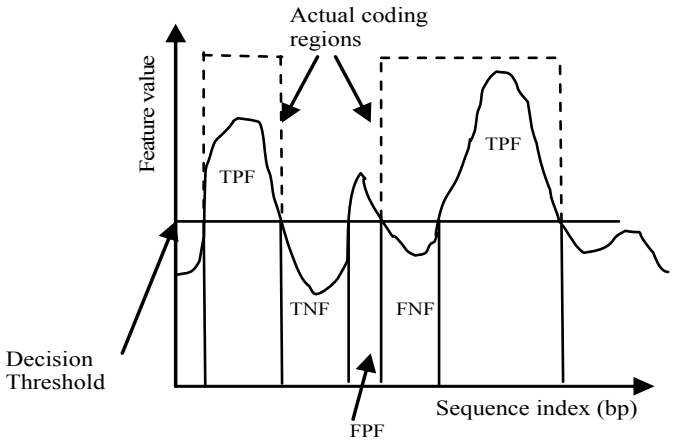

Fig. 3: Definitions of various quantities used to produce an ROC curve

When testing a number of sequences, their AUCs can be combined by weighting each AUC according to its length, to give a length-normalized average, as follows:

$$AUC_{LN} = \frac{\sum_{i=1}^{M} (L_i \times AUC_i)}{\sum_{i=1}^{M} L_i} \qquad (4)$$

where $i = 1, 2, 3\ldots, M$ is the sequence number and L_i is the length of the ith sequence. To give a fuller picture of the range of discrimination power of different techniques, the 10^{th} percentile, and 90^{th} percentile values of the AUC across a number of sequences were also calculated.

B. Peak Ratio

I have devised a second evaluation criterion, referred to here as a 'Ratio of Peaks' (*RP*):

$$RP = \frac{P_{HDCR}}{P_{LDCR}} \qquad (5)$$

where P_{HDCR} is the peak value in the highest detected coding region and P_{LDCR} is the peak value of the lowest detected coding region of a genomic sequence. In general, the detection of smaller coding regions or regions with less periodicity (e.g. the presence of very few codons having the same base at positions I and II) is less prominent [12], and also these regions would contribute very little in ROC curve values. Therefore, for a better detection of smaller coding regions it is desirable to have this ratio as small as possible. The length-normalized average (calculated similarly to equation (4)), along with 10^{th} percentile, and 90^{th} percentile values of the *RP* were calculated.

IV. Experimental Results

A. DNA Sequence Database

The frequency-domain and time-domain techniques were applied to 39 genomic sequences, including small sequences, sequences with closer coding regions and sequences with small coding regions. The results were compared by using ROC curves and the *RP* criterion described in section III (*B*). These sequences can be retrieved directly from the GenBank database, maintained by National Center for Biotechnology Information (NCBI) [13]. Following the convention of several recent authors (e.g. [7, 14]), I also show my results for the gene F56F11.4 in C-elegans (base number 7021 – 15080 in chromosome III; accession number AF099922), containing five exons.

B. Prediction Technique Settings

A constant window size of 351 was used for the sliding window DFT method for all sequences. In my AR model implementation, I varied order of the model from 10 to 120 and frame size from 51 to 600. I found an order of p=85 and frame size of 240 more suitable for my sequences. In frequency-domain techniques, I used a larger window size so that the periodicity effect could dominate the background 1/f spectrum which has a strong presence in DNA sequences. For both of the time-domain techniques a frame size of 117 was used.

C. ROC Results

Table 1 shows experimental result for all sequences using the AUC measure described in section III (*A*). Here we can observe that AUC for time-domain techniques is larger than frequency-domain techniques. Moreover, the results distribution (e.g. 10^{th} and 90^{th} percentile) is quite symmetric along their average value.

Fig. 4 and 5 shows the variation of AUC with average coding region length and the variation of AUC with average spacing between coding regions respectively. The time-domain techniques show a clear improvement throughout all coding region lengths and spacing. Figure 6 shows ROC curves for all techniques, aggregated across the entire data set. It is clear that time-domain techniques seem to discriminate more accurately exons from introns as compared to frequency-domain techniques.

Table 1: Experimental results for mixed sequences by using area under the ROC curve (AUC)

Method	Length normalized AUC	10^{th} Percentile (AUC)	90^{th} Percentile (AUC)	Gene F56F11.4 C. elegans (AUC)
FFT	0.7109	0.4620	0.7702	0.9255
AR	0.5043	0.4551	0.5500	0.4704
AMDF	0.7830	0.5156	0.8818	0.9730
TDP	0.7753	0.5030	0.9067	0.9695

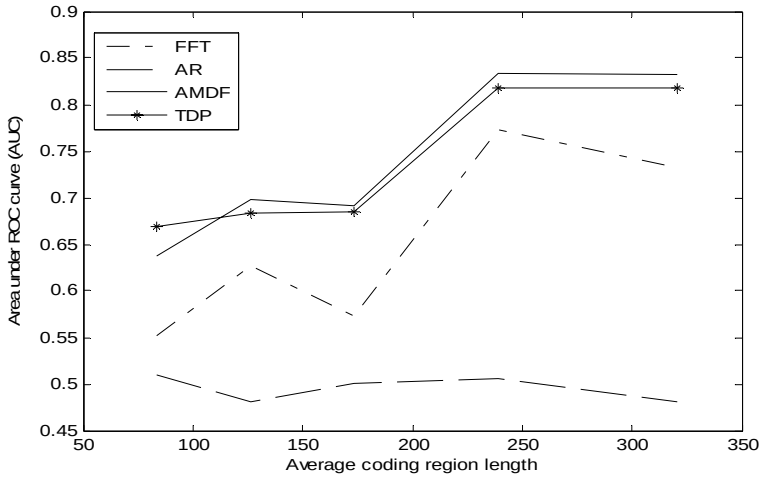

Fig. 4: AUC vs. Average coding region length curve for all techniques

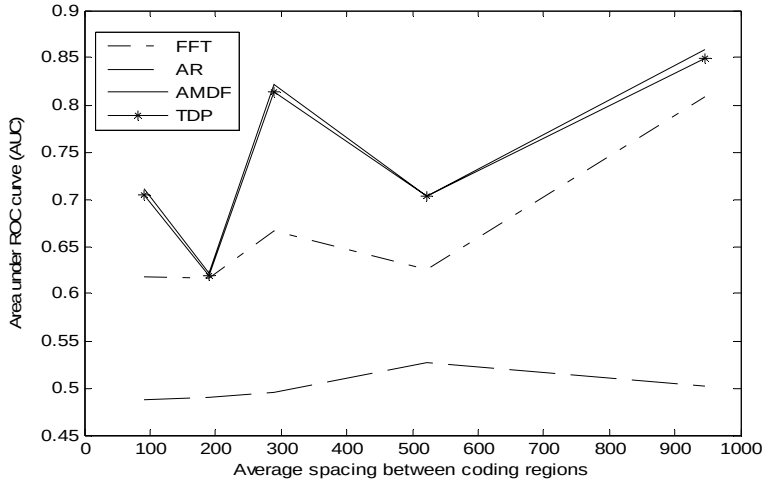

Fig. 5: AUC vs. Average spacing between coding regions curve for all techniques

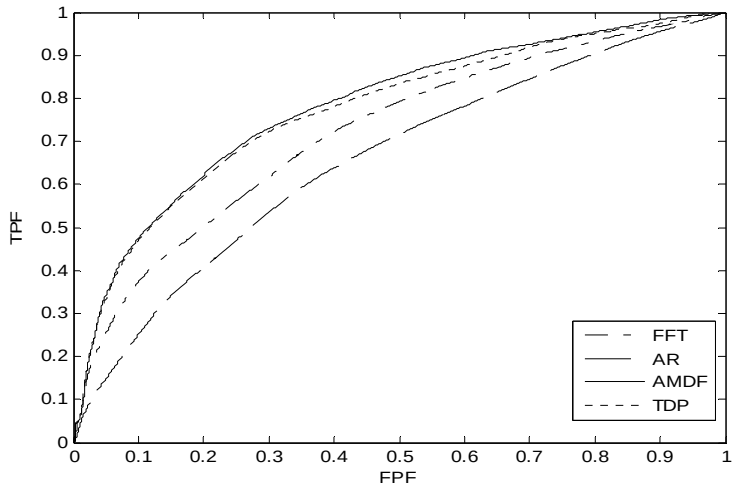

Fig. 6: ROC curves aggregated across the entire data set, for all techniques

D. Peak Ratio Results

Table 2 shows experimental results using the RP criterion, defined in section III (*B*), across all sequences. Fig. 7 and 8 shows the variation of RP with average coding region length and the variation of RP with average spacing between coding regions respectively. According to these results, the peak ratio for time-domain techniques is much smaller as compared to their frequency-domain counterparts throughout all coding region lengths and spacing.

Table 2: Experimental results for mixed sequences using a peak ratio (RP) criterion

Method	Length normalized RP	10^{th} Percentile (RP)	90^{th} Percentile (RP)	Gene F56F11.4 C. elegans (RP)
FFT	7.6010	1.7406	11.8640	3.3906
AR	8.8057	1.4413	9.6549	4.5355
AMDF	2.4215	1.3044	2.4977	1.8005
TDP	2.7493	1.2829	3.1832	1.8699

E. Comparison

Fig. 9 shows the comparison of four techniques. It can be observed that two time do-main algorithms perform with high accuracy compared with the frequency domain techniques, and have the additional advantages of computational efficiency and accurate estimation at smaller frame sizes.

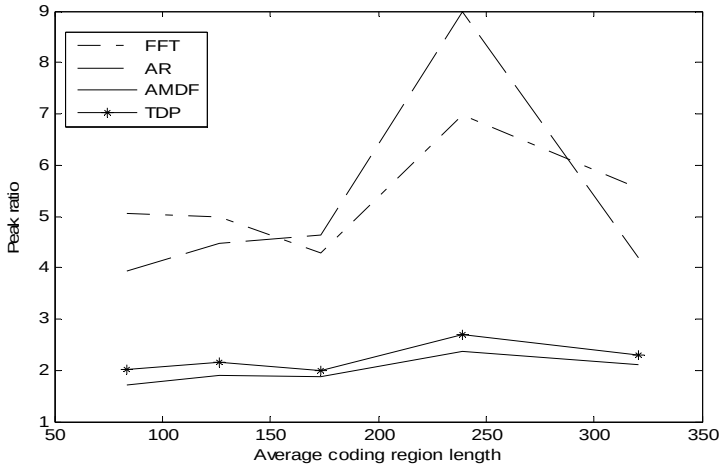

Fig. 7: Peak ratio vs. average coding region length curve for all techniques

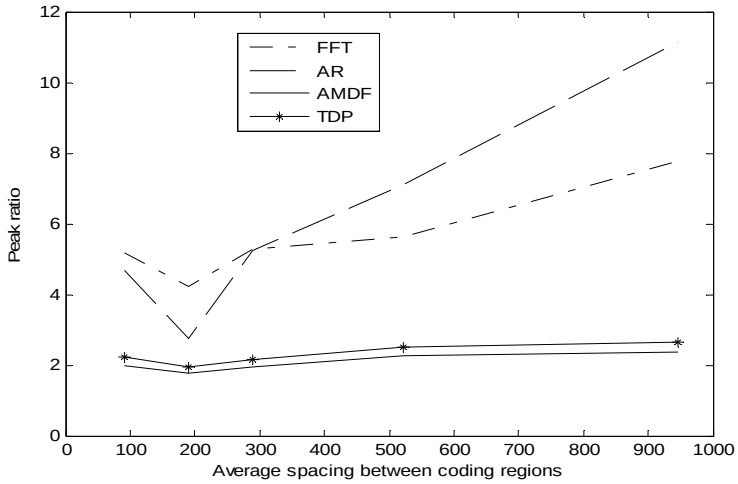

Fig. 8: Peak ratio vs. average spacing between coding regions curve for all techniques

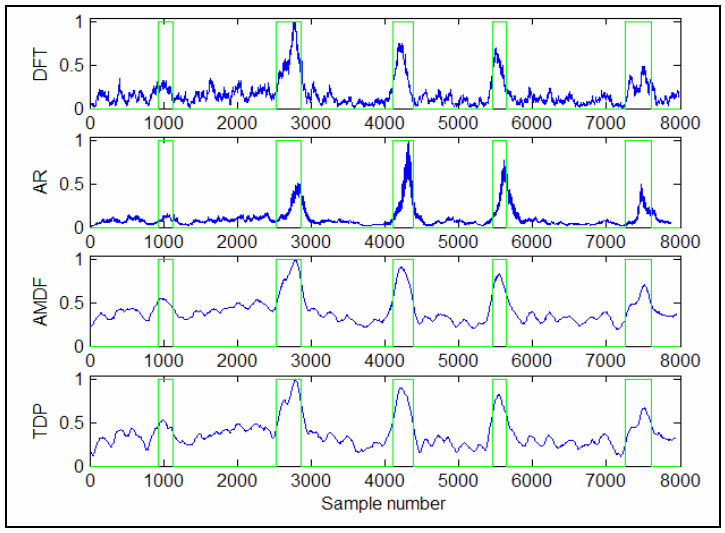

Fig. 9: Comparison of time-domain and frequency-domain techniques for gene F56F11.4 in C-elegans

It is evident that the time-domain algorithms perform with higher accuracy compared with the frequency-domain techniques, for all conditions. Although the AR spectral estimate theoretically provides higher 'time' (i.e. base pair) resolution, via use of a shorter time window, this does not seem to produce any advantage in terms of discriminating shorter and/or more closely spaced coding regions. This is possibly due to its modeling spurious detail in the form of spectral peaks. It is very difficult to get the model order *p* roughly correct for all type of sequences, before starting the analysis. This could be one of the reasons for the very weak results obtained for the AR method in these experiments.

The GENSCAN program has limitations for many organisms and smaller sequences, especially c. elegans sequences. Future work will focus on these sequences, further evaluations using larger and more diverse sequence data, classification techniques and finding the exact start and stop codons of exons using time-domain techniques.

References

[1] Mount, D. W., "*Bioinformatics Sequence and Genome Analysis*", Cold Spring Harbor Laboratory Press, New York, 2001.

[2] Tsonis A.A., Elsner J.B. and Tsonis P.A., "Periodicity in DNA coding sequences: Implications in Gene Evolution", *J Theor Biol* 1991, 151(3), pp. 323-331.
[3] Trifonov EN, "Elucidating sequence codes: three codes for evolution", *Ann NY Acad Sci* 1999, 870: pp. 330-338.
[4] Eskesen S.T., Eskesen F.N., Kinghom B., and Ruvinsky A., "Periodicity of DNA in exons", *BMC Molecular Biology*, August 18, 2004. 5(1):12.
[5] Anastassiou, D., "Frequency-domain analysis of biomolecular sequences", *Bioinformatics*, Vol. 16, no. 12, pp. 1073-1082, Dec. 2000.
[6] Anastassiou, D., "Genomic signal processing", *IEEE Signal Proc. Mag.*, July 2001, pp.8-20.
[7] Vaidyanathan, P. P., and Yoon, B.-J., "Gene and exon prediction using allpass-based filters", in *Proc. IEEE Workshop on Gen. Sig. Proc and Stat*., 2002.
[8] Rao, N. and Shepherd, S.J., "Detection of 3-periodicity for small genomic sequences based on AR technique", *International Conference on Communications, Circuits and Systems ICCCAS*, 2004, Vol. 2, pp. 1032-1036.
[9] Akhtar, M., Ambikairajah, E., and Epps, J., "Gene and Exon prediction using time-domain techniques", in *Poster Proc. Asia-Pacific Conf. on Bioinformatics (APCB)*, (Singapore), January 2005, p43.
[10] Tiwari, S., Ramachandran, S., Bhattacharya, A., and Ramaswamy, R., "Prediction of probable genes by Fourier analysis of genomic sequences", *CABIOS*, Vol. 113, pp. 263-270, 1997.
[11] Ambikairajah, E., and Carey, M., "The Time Domain Periodogram Algorithm", *Signal Processing*, vol. 5, 1983, pp. 491-513.
[12] Gutierrez G., Oliver J.L. and Marin A., "On the origin of the periodicity of three in protein coding DNA sequences", *J Theor Biol* 1994, 167(4), pp. 413-414.
[13] NCBI GenBank database, online access: *http://www.ncbi.nlm.nih.gov/Genbank/*
[14] Guan, R., and Tuqan, J., "IIR Filter Design for Gene Identification", in *Proc. IEEE Workshop on Gen. Sig. Proc and Stat*., 2004.

Mahmood Akhtar received the B.Sc. degree (Honors) in Electrical Engineering from University of Engineering and Technology (UET), Lahore, Pakistan, in 2002. He completed MS Computer Engineering from National University of Sciences and Technology (NUST), Rawalpindi, Pakistan, in 2004. He is currently pursuing the Ph.D. degree in the School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, Australia. His research area is Genome Signal Processing.