A DSP-Based Approach for Gene Prediction in Eukaryotic Genes

D. K. Shakya[1], Rajiv Saxena[2], and S. N. Sharma[3]

[1]Department of Biomedical Engineering,
Samrat Ashok Technological Institute, Vidisha, 464001, India
[2]Department of Electronics and Communication Engineering,
Jaypee University of Engineering and Technology, Raghogarh, Guna, 473226, India
[3]Department of Electronics and Instrumentation Engineering,
Samrat Ashok Technological Institute, Vidisha, 464001, India
devendrashakya@rediffmail.com, rsaxena2001@yahoo.com, sanjeev_n_sharma@rediffmail.com

Abstract: A simple algorithm to improve the identification accuracy of protein coding regions (exons) in Deoxyribonucleic Acid (DNA) sequences exploiting period-3 property is proposed. Three base periodicity is quite pronounced in exons and is commonly used in Digital Signal Processing (DSP) based methods to locate the exonic regions. Improvement in the accuracy of the protein coding regions has been achieved by extracting the background noise that comes from long range correlation present in DNA sequences and then eliminating this noise from the period-3 power spectrum. Proposed algorithm is data independent as it does not requires the empirical determination of any parameter for increasing the discrimination between coding and non-coding regions of a DNA sequence. Performance of the algorithm has been evaluated on F56F11 C.elegans chromosome-III nucleotide sequences. Performance of this algorithm has been compared with the spectral content method and an improvement in the correlation coefficient (CC), the performance metric used in this work, is observed.

KeyWords: Deoxyribonucleic Acid (DNA), Protein coding regions, Period-3 property, Discrete Fourier Transform (DFT), IIR Digital Filters, Genomic Signal Processing.

1. Introduction

DNA sequences are of fundamental importance in understanding living organisms, since all the information of the hereditary and species evolution is contained in these macromolecules. The DNA sequence comprises of four key chemicals, adenine (A), thymine (T), guanine (G), and cytosine (C). One of the present challenges of analyzing the DNA sequences is to determine the protein coding regions (exons) in eukaryotic gene structures [1, 2]. The difficulty of the problem is mainly due to the noncontiguous and non-continuous nature of genes (i.e., DNA consists of genic and intergenic regions, and eukaryotic genes are further divided into relatively small protein coding segments known as exons, interrupted by non-coding spacers known as introns). Furthermore, often the intergenic and intronic regions make up most of the genome. Figure 1 shows a DNA sequence.

In eukaryotes, exon regions are separated by introns, whereas in procaryotes these regions are continuous. Base sequences in the protein-coding regions have a strong period-3 component due to codon structure involved in the translation of the base sequences into amino acids [3]. Fourier analysis of DNA sequences is used to identify possible patterns in coding and non-coding regions. While intronic sequences show a rather random pattern, exonic sequences show periodicities of 3, 10.5, 200, and 400 [4]. Three base periodicity is quite pronounced and is commonly used in Digital Signal Processing (DSP) based methods to locate the exonic regions. Periodicity of three is present in the example periodic sequence: A-- A-- A-- A-- ..., where blanks can be filled randomly by A, T, C or G. This sequence shows a periodicity of three because of the repetition of the base A. Based on the period-3 property a number of algorithms have been developed to identify the protein coding

Received: June 14th, 2011. Accepted: December 13th, 2011

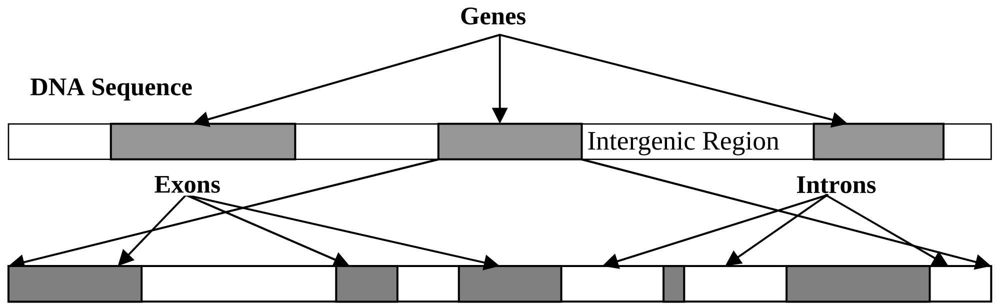


Figure 1. Various regions in a DNA molecule

regions [5, 6]. In the period-3 based methods like anti-notch and multistage filtering [7], quadratic windowing [8], emphasis has been on suppressing the signal in the non-coding regions [9], and boosting the coding region signal. In this work a simple algorithm to achieve better discrimination between coding and non-coding regions is proposed. Noise present in the coding and non-coding region has been captured using a notch filter and Discrete Fourier Transform (DFT). This noise has then been used to reduce the signal level in the non-coding regions without affecting the signal values in the coding-regions significantly. Genomic sequences comprises of four characters, so mapping of these sequences to numerical sequences is mandatory prior to sequence analysis. As with the commonly used binary or Voss representation scheme for DNA mapping [4] computational requirements are high, EIIP (electron-ion-interaction-potential) [2] values associated with each nucleotide are used in this work to map DNA character strings into numerical sequences for computational work. Numerical representation of DNA sequence and its spectrum analysis is discussed in the next section.

2. DNA Numerical Representation and Spectrum

DFT is used for spectrum analysis of biological data and comprises of three steps. (a) DNA sequence is mapped into a numeric sequence. (b) Spectrum of finite-length windowed DNA numerical sequences is computed. (c) Window function is translated by one or more bases and power spectrum is calculated along the length of the investigating DNA sequence. The DFT of a length-N block of x(n) is defined as

$$X[k] = \sum_{n=0}^{N-1} x(n)\,.\,w(n).e^{-j2\pi kn/N}, \qquad 0 \le k \le N-1 \tag{1}$$

where, w(n) is a window function [10].

In the EIIP mapping, the electron-ion-interaction-potential associated with each nucleotide is used to map DNA character string into numerical sequence, x(n). The EIIP values for the nucleotides are as follows [2] - A = 0.1260, G = 0.0806, T = 0.1335, and C = 0.1340. For example, for a DNA sequence CGATGACGAA, the EIIP indicator sequence will be x(n)= [0.1340 0.0806 0.1260 0.1335 0.0806 0.1260 0.1340 0.0806 0.1260 0.1260]. Merit of EIIP mapping scheme is that only one numeric sequence need to be processed instead of four as in binary representation, to extract the hidden information. Because of the period-3 property, magnitude of the DFT coefficient corresponding to k=N/3 is large in coding region. These coefficients are obtained using (1) and are then used to obtain the spectral content (SC) measure [3], as follows-

$$S[k] = \left| X[k] \right|^2 \tag{2}$$

The window is then slided by one or more bases and S[N/3] is recalculated for the entire genomic sequence. From the plot of S[N/3] versus nucleotide position, coding and non-coding regions are identified by applying a simple decision threshold. Codons with S[N/3] value above the threshold are considered to be in coding regions otherwise they are recorded as non-coding regions codons.

3. Proposed Algorithm (PA)

The proposed algorithm (PA) is shown in Figure 2. The DFT magnitude values at k = N/3 for DNA signal are obtained using (1).

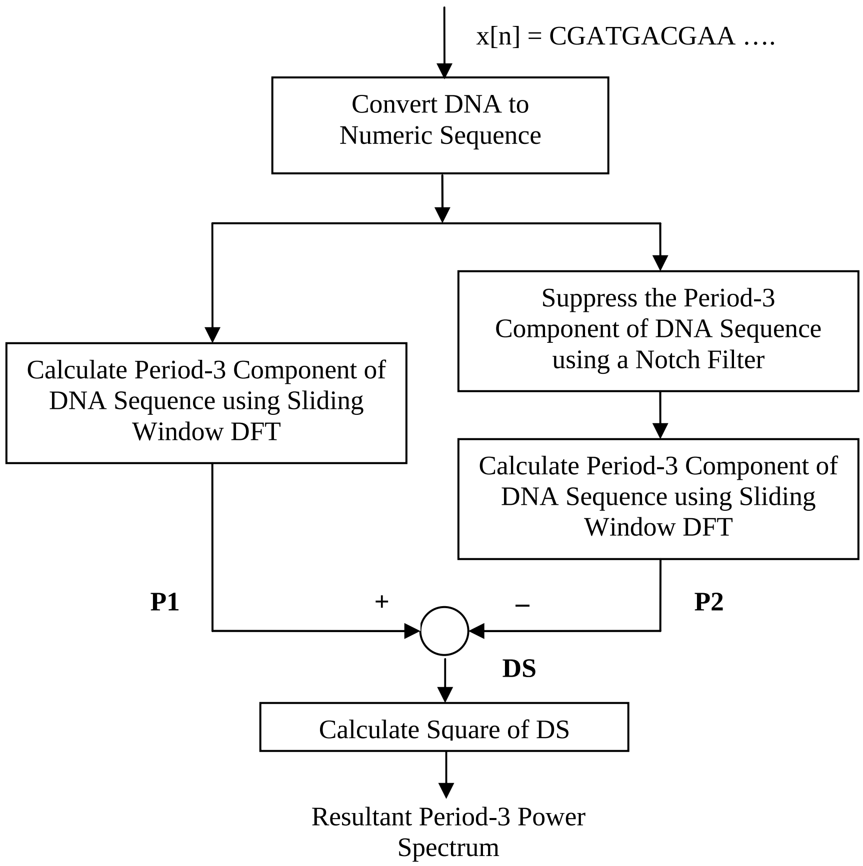


Figure 2. Proposed Algorithm for Gene Prediction

In (1) Bartlett window of length 351 has been used as it provides optimal window shape for processing genomic sequences in [11]. By sliding the window by one sample the process is carried out over the entire DNA sequence and the resultant signal obtained is shown in the algorithm by P1. In signal P1, that represents period-3 magnitude components of DNA data, non-coding region signals representing the noise are not suppressed effectively.

To capture this background noise which comes due to long-range correlation exhibited by DNA sequences both in the genic regions and intergenic regions, and eliminate it from P1, the numeric DNA sequence is first passed through a second order all pass Infinity Impulse Response (IIR) notch filter [7]. IIR filters require less computation and memory than FIR filters and can be very efficient here. Such filters can be built from second order allpass filters. The transfer function of the filter with pole at $\mathrm{Re}^{\pm j\theta}$ is given by (3).

$$A(z) = \frac{R^2 - 2R\cos\theta Z^{-1} + Z^{-2}}{1 - 2R\cos\theta Z^{-1} + R^2 Z^{-2}} \tag{3}$$

where, R < 1 for stability and |A(z)| =1 for all pass filter. The transfer function for notch filter is given by (4).

$$G(z) = \frac{1 + A(z)}{2} \tag{4}$$

The filter notch is centered at frequency 2π/3 and value of R is selected as 0.992. Notch filtering process removes the period-3 component of genomic sequence. The filtered signal is then subjected to a similar sliding window based DFT operation to obtain the spectral output P2 using (1). This spectral output P2 approximates the noisy component present in coding and non-coding regions of P1. The two spectral values P1 and P2 are plotted in Figure 3 for the gene F56F11.4 in *C.elegans* chromosome III.

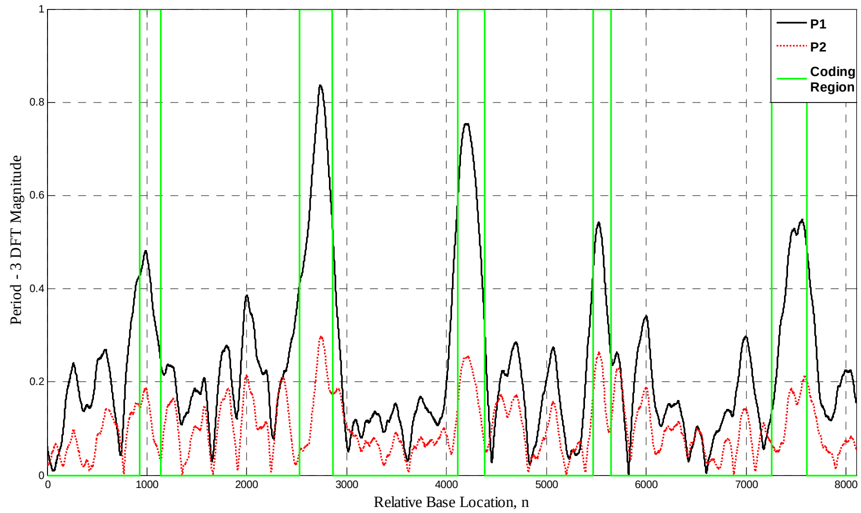


Figure 3. Capturing Noise for Gene F56F11.4

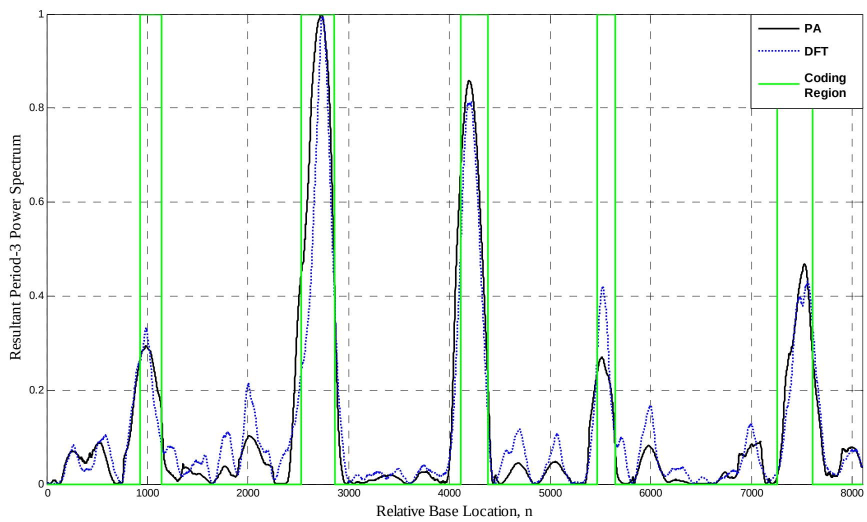


Figure 4. Comparative Results for F56F11.4

In the difference signal (DS), noise component in non-coding region is suppressed without affecting the coding region signals significantly. By squaring DS, resultant period-3 power spectrum, SR (N/3) is obtained for this algorithm.

4. Comparative Performance Evaluation

The performance of proposed algorithm is compared with the DFT spectral content method [3]. Comparative performance is illustrated in Figure 4 to Figure 6.These figures illustrate the suppression of noise present in non-coding regions. For evaluation of gene structure prediction programs different measures of prediction have been discussed in [12, 13] and can be explained with the aid of Figure 7. True positive (TP) is the number of coding nucleotides correctly predicted as coding. False negative (FN) is the number of coding nucleotides predicted as non-coding. True negative (TN) is the number of non-coding nucleotides correctly predicted as non-coding. False positive (FP) is the number of non-coding nucleotides predicted as coding. Sensitivity (S_n) is the probability of a nucleotide being predicted as coding given that it is actually coding and specificity (S_p) is the probability of a nucleotide being actually coding given that it has been predicted as coding. Both S_n and S_p can be viewed as conditional probabilities. Neither S_p nor S_n alone constitutes good measures of global accuracy.

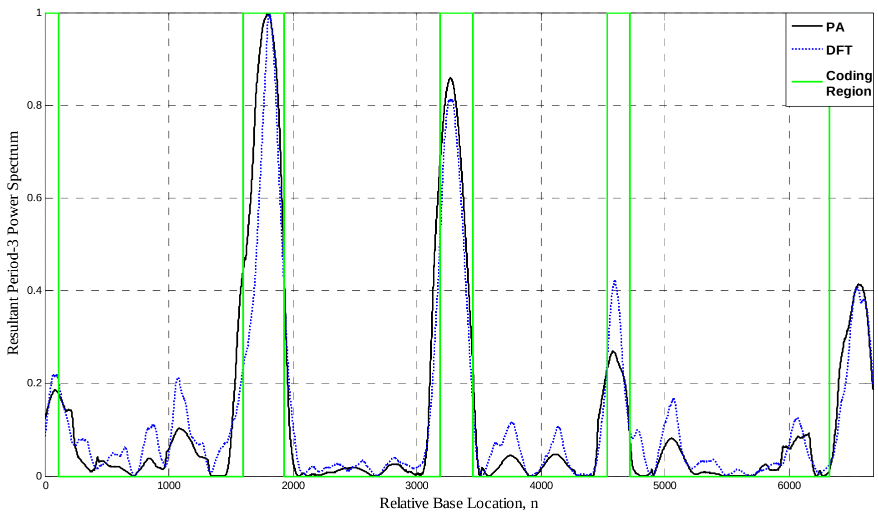


Figure 5. Comparative Results for F56F11.4a

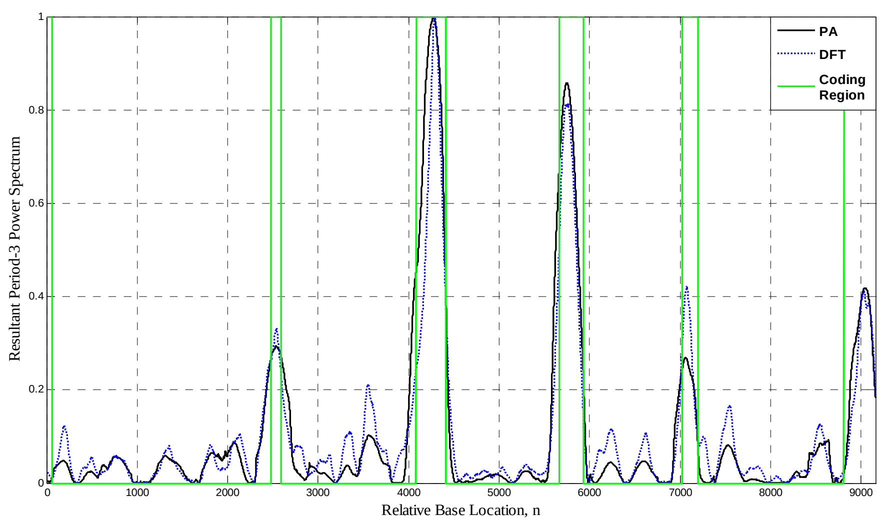
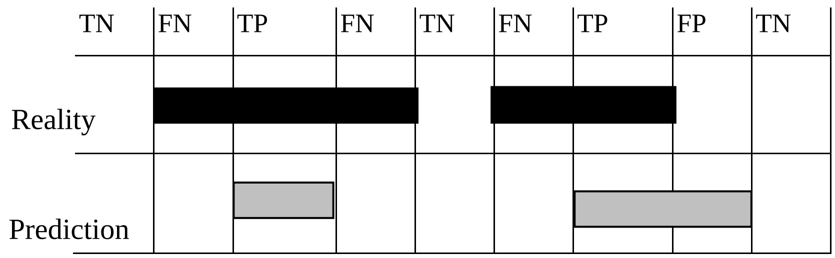


Figure 6. Comparative Results for F56F11.4b



TP=True Positive, FP=False Positive, TN=True Negative, FN=False Negative

$$S_n = \frac{TP}{TP + FN} \qquad S_p = \frac{TP}{TP + FP}$$

(Sensitivity) (Specificity)

Figure 7. Nucleotide Level Measurement

The suggested measure for gene structure prediction is correlation coefficient (CC) as it includes the aspects of both sensitivity and specificity [13]. Value of CC can be calculated using (5).

$$CC = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}} \tag{5}$$

Using simple decision thresholds DNA sequences are classified into coding/non-coding regions. Comparative values of S_n, S_p, and CC for complete F56F11 DNA sequence with decision thresholds varying from 10% to 90% are computed and shown in Table-1. Improvisation in the CC values is observed with the proposed algorithm. This method is data independent and does not require any prior information like data driven methods to train a model. Data independent methods are less accurate than the data dependent ones but they do not require any training or testing and can be quite useful for newly discovered genomic sequences for which no information is available initially.

Table 1. Comparative Analysis of S_n, S_p and CC Values.

Threshold	**Proposed Algorithm**			**DFT**		
	S_n	**S_p**	**CC**	**S_n**	**S_p**	**CC**
0.1	0.7627	0.5474	0.5229	0.7511	0.4898	0.4617
0.2	0.6793	0.6498	0.5651	0.6188	0.6721	0.5481
0.3	0.5230	0.6942	0.5084	0.5032	0.7698	0.5425
0.4	0.4279	0.7256	0.4696	0.3182	0.7851	0.4255
0.5	0.3339	0.7427	0.4165	0.2233	0.7968	0.3558
0.6	0.2443	0.7364	0.3484	0.1695	0.8120	0.3125
0.7	0.1943	0.7650	0.3189	0.1260	0.8278	0.2722
0.8	0.1495	0.7881	0.2852	0.0829	0.8049	0.2142
0.9	0.0740	0.7549	0.1910	0.0425	0.7585	0.1446

Conclusion

A very simple algorithm based on direct capturing of noise and removing it from resultant power spectrum has been developed for improving accuracy in the detection of protein coding regions by DFT spectral content method. The algorithm does not require empirical determination of any parameter for noise suppression and is thus data independent. Performance of the algorithm has been evaluated on F56F11. Improvement in period-3 detection in terms of CC that includes the aspects of both sensitivity and specificity is observed. The algorithm suppresses the extraneous peaks introduced by pseudo periodicities and thus enhances the probability of correct prediction of the exons. This algorithm can also be used for improving the accuracy in the detection of other periodicities of significance present in the DNA data. Further improvements will be carried out in future work by avoiding the suppression of coding region signals. Also the proposed algorithm will be generalized to improve identification accuracy using other transforms like wavelet.

References

[1] J. Tuqnan and A. Rushdi, “A DSP Approach for finding the codon bias in DNA sequence,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 343-356, June 2008.

[2] K. D. Rao and M. N. S. Swamy “Analysis of genomics and proteomics using DSP techniques,” *IEEE Transactions on Circuits and Systems-1*, vol. 55, no. 1, pp. 370-378, February 2008.

[3] 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.

[4] M. Akhtar, J. Epps, and E. Ambikairajah, “Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 310-321, June 2008.

[5] D. Anastassiou, “Genomic signal processing,” *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8-20, July 2001.

[6] D. Sussillo, A. Kundaje, and D. Anastassiou, “Spectrogram analysis of genomes,” *EURASIP Journal on Applied Signal Processing*, vol. 1, pp. 29-42, 2004.

[7] P. P. Vadyanathan and B. J. Yoon, “Digital filters for gene prediction applications,” *in Proceedings 36th Asilomer Conference on Signals Systems and Computers, Monterey*, CA, November 2002.

[8] T. W. Fox and A. Carreira, “A digital signal processing method for gene prediction with improved noise suppression,” *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 1, pp. 108-114, 2004.

[9] D. K. Shakya, Rajiv Saxena, and S. N. Sharma, “A Simple Algorithm for Gene Prediction with Improved Noise Suppression”, *Proceedings of the 10th IEEE International Conference on Signal Processing*, Beijing, China, 2010, pp.1765-1768

[10] F. J. Harris, “On the use of windows for harmonic analysis with the discrete fourier transform,” *Proc. IEEE*, vol. 66, pp. 51–83, 1978.

[11] T. S. Gunawan, “On the optimal window shape for genomic signal processing”, *Proceeding of the International Conference in Computer and Communication Engineering*, pp. 252-255, May 13-15, 2008.

[12] C. Burge, “Identification of genes in human genomic DNA”, Ph.D. dissertation, Stanford University, Stanford, CA, 1997.

[13] M. Burset and R. Guigo, “Evaluation of gene structure prediction program”, Genomic, vol. 34, pp. 353-367, 1996.

D. K. Shakya received the B.E. degree in Electronics and Instrumentation Engineering from Barkatullah University, Bhopal, M.P., India, in 1999, and the M.E. in Digital Technique and Instrumentation from Rajiv Gandhi Proudyogiki Vishwavidyalaya Bhopal, M.P., India, in 2002. He is currently working as an Assistant Professor in the Department of Biomedical Engineering, Samrat Ashok Technological Institute, Vidisha, M.P., India, and pursuing Ph.D. degree. His teaching and research interests include Genomic and Proteomic Signal Processing, Digital Filter Design, Bio-signal and Image Processing.

Rajiv Saxena received the B.E. degree in Electronics and Telecommunication Engineering in the year 1982 from Jabalpur University, M.E. degree in Digital Techniques & Data Processing from Jiwaji University, Gwalior in 1990, and the Ph.D. degree from University of Roorkee (IIT Roorkee), India, in year 1996 in Electronics & Computer Engineering. He is currently working as a Professor and Head and in the Department of Electronics and Communication Engineering, Jaypee University of Engineering and Technology, Raghogarh, Guna, M.P., India. His teaching and research interests include Digital Signal Processing, Communication Engineering, Integral Transforms, Digital Image Processing and Mobile Communication System. He has 27 years of teaching experience. He has guided 10 Ph.D. scholars and published more than 40 research papers in various international and national refereed journals and conferences. He received the best paper award by IETE, New Delhi, in the year 2008.

S. N. Sharma received the B.E. degree in Electronics and Instrumentation Engineering from Barkatullah University, Bhopal, M.P., India, in 1991, M.E. degree in Measurements & Instrumentation Engineering from University of Roorkee (IIT Roorkee), India, in 1993, and the Ph.D. degree in Electronics and Communication Engineering, in 2007, with specialization in Signal Processing from Thapar University, Patiala. He is currently working as an Associate Professor in the Department of Electronics and Instrumentation Engineering, Samrat Ashok Technological Institute, Vidisha, M.P., India, His teaching and research interests include Digital Signal Processing, Genomic Signal Processing, Fractional Fourier transform, Digital Filter Design and Bio-signal Processing. He has 16 years of teaching experience. He is having 10 publications in reputed journals and refereed conferences.