Evaluation of DNA Mapping Schemes for Exon Detection

S. D. Sharma
Department of E&I
Inst. of Technology & Management
Gwalior (M.P.), India
suneel.sati@gmail.com

K. Shakya
Biomedical Engg. Department
Samrat Ashok Technological Inst.
Vidisha (M.P.), India
devendrashakya@rediffmail.com

S. N. Sharma
Department of E&I
Samrat Ashok Technological Inst.
Vidisha (M.P.), India.
sanjeev_n_sharma@rediffmail.com

Abstract—**Identification of protein coding regions (exons) in eukaryotic genomic sequences is an active area of research at present. Mapping of symbolic genomic sequences to numeric sequences is the first step required for processing them using digital signal processing (DSP) tools. For DFT-based methods paired numeric and frequency of nucleotide are reported as the best mapping schemes. In this work performance of a wavelet-based method for exon detection is evaluated with different symbolic-to-numeric representations. Optimum performance is obtained by using Z-Curve for DNA mapping. For performance evaluation Receiver Operating Characteristics (ROC) curves are used and the study is conducted on HMR195 data set. This work in general highlights that exon prediction accuracy and computational complexity of the DSP-based algorithms is dependent on the scheme used to map DNA nucleotides into numerical sequences, and so the optimum performance of any algorithm can only be attained with a particular mapping scheme.**

Keywords - DNA; Protein-coding regions; Discrete Fourier Transform; Gabor Wavelet;

I. INTRODUCTION

Deoxyribonucleic acid (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. Organisms can be categorized into prokaryotes and eukaryotes. In prokaryotes DNA is free in the cell, whereas in eukaryotes DNA is kept inside the nucleus and is separated from the rest of the cell by a nuclear memebrane. The DNA sequence comprises 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]. Methods used to identify protein-coding regions or exons in eukaryotic cells are either probabilistic or deterministic [3]. Probabilistic methods provide high accuracies but are model-dependent and require suitable training data for prediction. On the other hand prediction accuracy of deterministic methods is relatively less but they are model-independent, and are more suitable for the analysis of uncharacterized genomic sequences where the previous knowledge of the species under analysis is not available.

Base sequence in the protein-coding region has a strong period-3 component due to codon structure involved in the translation of the base sequence into amino acids [4]. Most of the deterministic techniques rely on spectral analysis of the DNA sequences using the discrete Fourier transform (DFT) to identify this period-3 component. Based on the period-3 property a number of algorithms have been developed to identify the protein coding regions [5, 6]. The performance of the DFT based methods is dependent on the window length [7]. Recently, a method based on modified Gabor-wavelet transform (MGWT) for the identification of protein-coding regions has been introduced [8]. The performance of the MGWT is independent of the window length and also better than the DFT-based methods.

The first step in genomic signal processing is the conversion of symbolic genomic sequences to numeric sequences. DNA numerical representations for DNA sequence analysis are discussed in [9, 10]. In [1], performance of DNA numerical representations for period-3 based exon prediction using DFT is compared. Performance of MGWT has been studied in [8] using only Voss representation. In this paper prediction accuracy and computational complexity of MGWT using other mapping schemes has been evaluated and compared. As in [1], best mapping scheme for MGWT has been proposed here on the basis of area under the ROC curves (AUC) and the processing requirements.

II. MAPPING SCHEMES

DNA sequences comprise of four letters A, T, C, and G. Processing of DNA sequences using DSP methods requires their conversion from a character string into numerical sequences as a first step. In recent years, a number of schemes have been introduced to map DNA characters into numeric values [9, 10]. These DNA representation schemes and their complexity in terms of the number of sequences to be processed are summarized in Table 1. Akhtar *et.al.* have compared the DNA representations for the exon detection problem for DFT-based methods [1]. In [1] it has been concluded that paired numeric and frequency of nucleotide occurrence methods reveal improved DFT-based gene and exon prediction with 75% less downstream processing. In identification of protein-coding regions using MGWT [8], Voss mapping scheme has been used.

978-1-4244-9394-4/11/$26.00 ©2011 IEEE

However, performance of MGWT with other existing mapping schemes is missing in the current literature and is presented in this paper.

Table 1. DNA Numerical Representation Schemes

S.No.	Mapping Schemes	DNA Representation	S(n)= [CGAT]	Complexity
1	Voss	$X_n = 1$ for S(n) = x $X_n = 0$ for S(x) ≠ x X_n applies to any C_n, G_n, A_n, and T_n	C_n = [1, 0, 0, 0] G_n = [0, 1, 0, 0] A_n = [0, 0, 1, 0] T_n = [0, 0, 0, 1]	4
2	Tetrahedron	$x_r(n) = \frac{\sqrt{2}}{3}[2T_n - C_n - G_n]$ $x_g(n) = \frac{\sqrt{6}}{3}[C_n - G_n]$ $x_b(n) = \frac{1}{3}[3A_n - T_n - C_n - G_n]$	$x_r(n) = \frac{\sqrt{2}}{3}[-1, -1, 0, 2]$ $x_g(n) = \frac{\sqrt{6}}{3}[-1, -1, 0, 0]$ $x_b(n) = \frac{1}{3}[-1, -1, 3, -1]$	3
3	Z-Curve	$\begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} = 2 * \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} x_A[n] \\ x_C[n] \\ x_G[n] \\ x_T[n] \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} = \begin{bmatrix} -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ -1 & -1 & 1 & 1 \end{bmatrix}$	3
4	Complex	A = 1+j, C = -1+j, G = -1-j, T = 1-j	[-1+j, -1-j, 1+j, 1-j]	1
5	EIIP	A= 0.1260, C= 0.1340, G= 0.0806, T= 0.1335	[0.1340, 0.0806, 0.1260, 0.1335]	1
6	Paired Numeric	A or T = 1, C or G = -1	[-1, -1, 1, 1]	1
7	DNA walk	C or T = 1, A or G = -1	[1, 0, -1, 0]	1
8	Frequency & Nucleotide Occurrence	C=0.2831, G=0.2134, A=0.2275, T=0.2760	[0.2831, 0.2134, 0.2275, 0.2760]	1
9	Atomic Number	A = 70, C = 58, G = 78, T = 66	[58, 78, 70, 66]	1
10	Real Number 1	C=1,G=3, A=2,T=0	[1, 3, 2, 0]	1
11	Real Number 2	C=2,G=1,A=0,T=3	[2, 1, 0, 3]	1
12	Real Number 3	C=0.5,G=-0.5,A=1.5,T=-1.5	[0.5, -0.5, 1.5, -1.5]	1

III. EXON DETECTION USING MGWT

Performance of DFT-based gene prediction methods is dependent on window length [7]. In these methods both short and long exonic regions in a gene are subjected to the analysis using same window length resulting in reduced prediction accuracy. To alleviate this shortcoming of DFT based schemes a method based on Gabor wavelet has been introduced in [8]. To capture the period-3 components with varying window lengths the Gaussian standard deviation of the Gabor wavelet is varied, while the complex exponential frequency is kept constant. In [8] Modified Gabor Wavelet (MGW) has been defined as-

$$\psi_{MGWT}(t,n,a) = e^{\frac{-(t-n)^2}{2a^2}} e^{j\omega_0(t-n)} \quad (1)$$

In (1), n is the position along the DNA sequence, a is the scaling parameter, and ω_0 is the basic frequency of wavelet. The MGWT of a signal $u(t)$ is given as-

$$U(n,a) = \int u(t) e^{\frac{-(t-n)^2}{2a^2}} e^{j\omega_0(t-n)} dt \quad (2)$$

In (2), ω_0 is fixed at $L/3$ to capture the period-3 component, where L is the length of the DNA segment analyzed. Using (2), the MGWT of the DNA numerical sequence is calculated. The spectrum of the sequence is obtained by computing the squared complex modulus of the MGWT coefficients as-

$$M(n,a) = |U(n,a)|^2 \quad (3)$$

This spectrum is then projected onto the position axis in order to detect possible coding regions, which correspond to the local maxima regions of the projection. For a DNA sequence of length N this projection spectra is obtained using following relationship, and is the period-3 measure of MGWT-

$$MGWT(n) = \sum_{a} M(n,a), \qquad \nu = 0, \_ \_ \_ N\text{–}1 \qquad (4)$$

For computing MGWT, 40 analyzing functions corresponding to 40 scale values exponentially separated between 0.2 and 0.7 are used. The length of these functions is restricted to 1,200 sequence points [8]. The results for gene F56F11.4 of C.*elegans* (Gen Bank accession numberAF099922 and positions 7021-15020) using MGWT are shown in Fig.1.

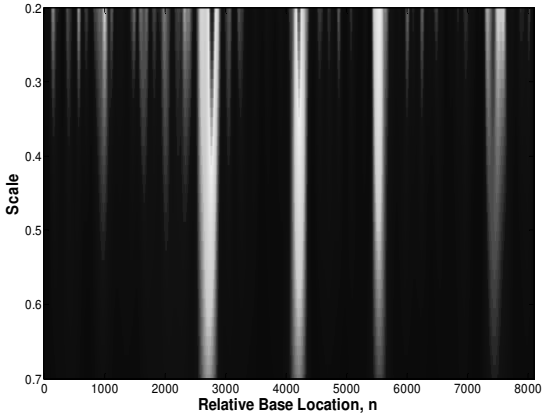

Figure 1. Spectrogram for the Sequence F56F11.4

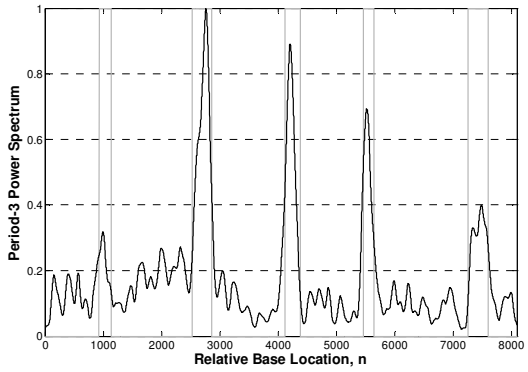

Figure 2. Projection of transform coefficients of the sequence F56F11.4

IV. EVALUATION MEASURES AND DATASET

For evaluation of gene structure prediction using MGWT with different mapping schemes, Receiver Operating Characteristic (ROC) Curves have been plotted. The prediction accuracy measures can be explained with the aid of Fig 2. Sensitivity, Sn, is the probability of a nucleotide being predicted as coding, given that it is actually coding, and specificity, Sp, is the probability of a nucleotide being actually coding given that it has been predicted as coding. In the exon-intron separation problem, an ROC curve explores the effects on TP and FP as the position of an arbitrary decision threshold is varied. The period-3 power spectrum curves are normalized with values between 0 and 1 and the decision threshold is varied to obtain values of TP and FP.

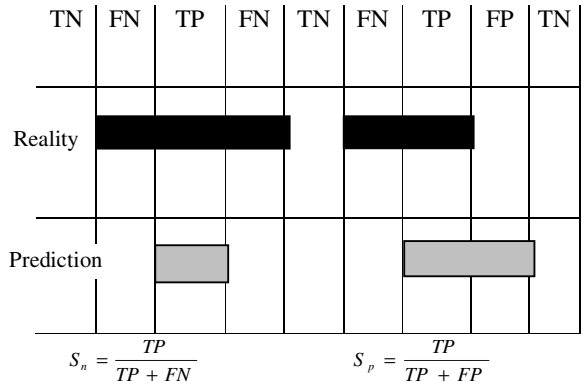

Figure 3. Nucleotide level measurement

The curve between TP and FP can be characterized as a single number using the area under the ROC curve (AUC), with larger areas indicating more accurate detection method.

The experiment for performance evaluation has been done on HMR195 [11] dataset. The HMR195 data set contains 195 mammalian sequences with exactly one complete single-exon or multi-exon genes. The ratio of Human:Mouse:Rat sequences is 103:82:10. Each sequence starts with an 'ATG' start codon and ends with a stop codon (TAA, TAG, TGA) with no in frame stop codons.

V. PERFORMANCE EVALUATION OF DIFFERENT DNA REPRESENTATIONS WITH MGWT

To evaluate the exon prediction accuracy of DNA representation schemes with MGWT, ROC curves have been plotted using the HMR195 dataset for different mapping schemes discussed in Section II. Values of AUC obtained using ROC curves, their computational complexity in terms of the number of sequences to be processed, and the exonic nucleotide detection rates for p% false positives, using HMR195 dataset are summarized in Table-2. In exon detection problem, the behavior of ROC curves at low false positive rates is more significant due to the high occurrence of false positives because of the low exonic fraction in the eukaryotic genomes [12]. Hence, in Table-2 results only up to 30% of false positives are recorded. These results indicate that the maximum AUC is obtained by using Voss and Z-curve mapping schemes. In [8], Voss mapping scheme has been used. However, computational complexity associated with Z-curve is less than the Voss scheme, as only three sequences need to be processed with the former one.

Mapping Scheme	AUC	Complexity	Data Driven	% of Exonic Nucleotides Detected as False Positive		
				10%	*20%*	*30%*
Voss	**0.8358**	**4**	**N**	**57.51**	**72.41**	**81.42**
Tetrahedron	0.7982	3	N	50.67	67.67	77.10
Z-Curve	**0.8358**	**3**	**N**	**57.50**	**72.41**	**81.42**
Complex	0.6438	1	N	26.14	40.29	50.91
EIIP	0.7501	1	N	42.00	57.39	67.63
Paired Numeric	0.8157	1	N	55.04	69.21	77.73
DNA walk	0.7802	1	N	49.27	63.17	72.59
Freq. of Nucleotide Occurrence	0.8129	1	Y	54.17	68.63	77.32
Atomic Number	0.6827	1	N	33.53	46.88	57.14
Real Number 1	0.8111	1	N	53.84	68.55	77.68
Real Number 2	0.6695	1	N	29.06	44.60	55.78
Real Number 3	0.6696	1	N	29.05	44.55	55.78

TABLE 2. SUMMARY OF RESULTS FOR WAVELET BASED EXON DETECTION

V. CONCLUSION

In this work almost all reported DNA representations have been compared for exon prediction using MGWT. In the introductory paper of MGWT [8], Voss representation has been used. Generally Voss is preferred as in [8], with the assumption that three base periodicity is lost with other mapping schemes. This study however indicates that the use of Z-curve mapping scheme with MGWT results in same prediction accuracy but with 25% less processing as compared to Voss. For DFT-based exon prediction, paired numeric and frequency of nucleotide occurrence are reported to be the best mapping schemes [1]. Hence, using the inferences drawn from this and an earlier work [1], it can be concluded that the Voss is not always the best mapping scheme and the performance of DSP-based algorithms for identification of protein-coding regions is dependent on the scheme used to map symbolic genomic data into numerical signals. As conflicting conclusions may arise due to the choice of the mapping scheme used for numerical representation of genomic data, a framework for the analysis of the equivalence of the mappings is developed in [13].

REFERENCES

[1] 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.

[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] R.Guigo, "DNA composition, codon usage and exon prediction," *Genetic Databases*, pp. 53-80, Academic Press, 1999.

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

[5] D.Anastassiou, "Genomic signal processing," *IEEE Signal ProcessingMagazine,* 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] M.Akhtar, E.Ambikairajah, and J.Epps, "Optimizing period-3 methods for eukaryotic gene prediction", Proc. *IEEE 33rd International Conference on ASSP*, Phoenix, Arizona, USA, pp.621-624, March-April 2004.

[8] J.P.Mena-Chalco, H.Carrer, Y.Zana, and R.M.Cesar Jr., "Identification of protein coding regions using the modified Gabor-wavelet transform," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol.5, no.2, pp.198-207, April-June 2008.

[9] H.K.Kwan and S.B.Arniker, "Numerical representation of DNA sequences," Proceeding of IEEE International conference of Electro/Information Technology, Windsor, Canada, pp.307-310, June 2009.

[10] M.Akhtar, J.Epps, and E.Ambikairajah, "On DNA numerical representations for period-3 based exon prediction," IEEE 5th International Workshop on Genomic Signal Processing and Statistics (GENSIPS '07), Tuusula, Finland, pp.1-4, June 2007.

[11] S. Rogic, A.K. Mackworth and B.F. Ouellette, "Evaluation of gene finding program on mammalian sequence", *Genomic Research*, vol. 11, no. 5, pp. 817-832, 2001.

[12] M.Akhtar, E.Ambikairajah, and J.Epps, "Paired spectral content measure for gene and exon prediction in eukaryotes," *Proc. of IEEE International Conference on Information and Emerging Technologies*, Karachi, Pakistan, pp. 127-130, July 6-7, 2007.

[13] L.Wang and D.Schonfeld, "Mapping equivalence for symbolic sequences: theory and applications", IEEE Trans. On Signal Processing, vol.57, no.12, pp. 4895-4905, Dec. 2009.