Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Prediction Model**

Multi-scale parametric spectral analysis for exon detection in DNA sequences based on forward-backward linear prediction and singular value decomposition of the double-base curves

Miew Keen Choong[1,] * and Hong Yan[1, 2]

[1]School of Electrical and Information Engineering, University of Sydney, NSW 2006; [2]Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong;
Miew Keen Choong* - E-mail: miewkeen@ee.usyd.edu.au; * Corresponding author

received September 20, 2007; accepted October 19, 2007; published February 12, 2008

Abstract:
This paper presents a new method for exon detection in DNA sequences based on multi-scale parametric spectral analysis. A forward-backward linear prediction (FBLP) with the singular value decomposition (SVD) algorithm FBLP-SVD is applied to the double-base curves (DB-curves) of a DNA sequence using a variable moving window sizes to estimate the signal spectrum at multiple scales. Simulations are done on short human genes in the range of 11bp to 2032bp and the results show that our proposed method out-performs the classical Fourier transform method. The multi-scale approach is shown to be more effective than using a single scale with a fixed window size. In addition, our method is flexible as it requires no training data.

Keywords: spectral estimation; autoregressive model; double-base curve; DNA sequence analysis; gene identification

Background:

Genome sequences contain the genetic information of living organisms. This information, which is common to all life, is coded in the deoxyribonucleic acid (DNA) sequences. Understanding the codes will enable us to understand the life processes of a particular organism. As such, even with the genome sequence in hand, much work remains to be done to lay open the genetic secrets of a particular species. Decoding the meaning of the nucleotide characters A, T, C, and G is becoming even more pressing with the final release of the sequencing of the human genome.

Gene identification is of great importance in the study of genomes, to determine which open reading frames (ORFs) in a given sequence are coding sequences for prokaryotic, and to determine the exons and introns, and the boundaries between them in a given gene for eukaryotic DNA sequences. There are a number of identification methods being used, either with training datasets, or without any database information. Genescan **[1]** use a semi-hidden Markov model, and FEX **[2]** use a linear discriminant function to determine genes, are examples of gene or exon finding algorithms based on database information. Examples of algorithms without database information are statistical correlation analysis **[3]**, statistical regularity to detect coding regions **[4]**, and Fourier analysis **[5]**.

Among the various methods, the most prominent distinctive feature of coding and non-coding regions is the 3 base pairs (bp) periodicity or 1/3 frequency, which has been shown to be present in coding sequences **[6]**. The periodicity is caused by the coding biases in the translation of codons into amino acids. Eskesen *et al.* **[7]** has shown using simulated sequences that the DNA periodicity in coding region is determined by codon usage frequencies, which is lack in introns. This signature of 3bp periodicity in coding regions has been used and proved successful. Kotlar and Lavner **[8]** presented a method based on spectral rotation measure. Yan *et al.* **[9]** proposed the lengthen-shuffle Fourier transform.

The Fourier transform analysis has been widely used for sequence processing **[5, 9]**. However, Fourier transform contains the problems of windowing or data truncation artifacts and spurious spectral peaks, and thus, the spectral obtained using the Fourier transform will exhibit the same problems. This problem has been studied extensively in digital signal and image processing, where autoregressive (AR) models are used to achieve a high spectral resolution. The AR model or linear prediction (LP) process is a relatively new approach to spectral analysis to overcome the limitation of Fourier methods.

In this paper, we concentrate on the periodicity of 3bp to distinguish coding and non-coding regions. Our method is developed based on the AR model using forward-backward linear prediction (FBLP) and the singular value decomposition (SVD) (FBLP-SVD) algorithm **[10]**. We apply a moving window analysis to the double-base curves (DB-curves) **[11]** representation of a DNA sequence to identify very short human genes. Since different window sizes for spectral analysis will induce different results **[8]**, we have developed a multi-scale method **[12]** to solve the problem. Using this new approach, we are able to improve the results considerably.

Methodology:

We adopt the DB-curve mapping rule **[11]** which uses two bases out of the four DNA nucleotides. By ignoring the base order, there are six combinations: AC, AC, AG, TC, TG, and CG. The DB-curves are constructed by taking the cumulative occurrence of the different combinations using equation (1) and (2) (under supplementary material) and FBLP can be written as given in equation (3 under supplementary material). Then, spectral analysis using FBLP-SVD is done.

There are several methods to estimate the LP coefficients, $a^t = (a_1, a_2, \ldots, a_M)^t$. We adopt least squares (i.e. conditional maximum likelihood) estimates by solving equation (3) (see supplementary material), by using the pseudo inverse.

A well-known deficiency of the AR method is the high bias if a low prediction order is used and the occurrence of spurious peaks if high prediction order is used. The problem is solved by Tufts and Kumaresan **[10]** with the use of singular value decomposition (SVD). The use of SVD in approximating a noisy version of a signal matrix which is constructed from a linear model will produce a better approximate of the signal matrix. The SVD based algorithm is able to increase the signal-to-noise-ratio (SNR) in the data.

The power spectrum density is estimated by equation (4) (shown under supplementary material). The order selection criteria used in this paper is combined information criterion (CIC) **[13]**, and the candidate order is in the range of $N/3$ to $N/2$. This range is chosen as Lang and McClellan **[14]** recommended that, for a fixed number of data samples, the number of coefficients should be between $N/3$ and $N/2$ for best frequency estimation.

For a DNA sequence of length N, the numeric sequence using a DB-curve is given in equation (5) (supplementary material). The power spectrum as constructed using (4) can be given by $|P(f)|^2$ as in equation (6) (shown in supplementary material).

We employ a sliding window with step size of 1 along the DNA sequence to calculate the local spectrum density. As may already be obvious, different window sizes for spectral analysis will produce different results. A short analysis frame may detect short exons and introns, but causes more statistical fluctuations **[8]**. A larger window size may miss the short exons and introns, but cause fewer false negatives and false positives. Thus, we make use of multiple window sizes, with the aim of reinforcing the advantages of both short and long window sizes but overcome the disadvantages that are caused by them. It has been shown that different window sizes in spectral analysis are equivalent to different scales used in wavelet analysis **[12]**. We select the window size within the range of 60bp-360bp. In this paper, we chose four windows, which are 60bp, 90bp, 180bp and 360bp. The P_{ratio} combination of the windows, P_{multi} is defined as in equation (7) (see supplementary material).

Results and discussion:

The analysis of the proposed algorithm is conducted on DNA sequences of human genome downloaded from the NCBI GenBank database. The selection of the genes used for the simulation is done based on the paradigm that the gene contains short exons. There are a total of 96 genes, with 692 exons. The lengths of the exons in the genes are within the range of 11bp to 2032bp.

The Receiver Operating Characteristics (ROC) graphs and area under the ROC curve (AUC) are used as the evaluation criteria. The ROC is an important comparison method as it can be used to depict the tradeoff between hit rates and false alarm rates of the detector. An AUC value of 1.0 indicates a perfect test and a score of 0.5 means a random classifier.

Firstly, comparison is done on the proposed multi-scale analysis and fixed window spectral analysis. We labeled every nucleotide in the coding region as “positive” and all other regions as “negative”. The result is shown in the first row of Table 1 (under supplementary material). We can observe that the AUC for multi-scale approach is larger than all single scales. As we have mentioned before, short window size may detect the short exons whereas large window sizes cause fewer statistical fluctuations. The result of the analysis shows that the advantages of short and large analysis frames can be maintained while suppressing the disadvantages by the combination of different window sizes.

An assessment is also done on the numerical representation of sequence. We compare the DB-curve representation which originated from Wu *et al.* **[11]**, and the binary representation which is one of the most frequently used conversions. By comparing the first two rows of Table 1 (see supplementary material), we conclude that the DB-curve representation out-performs the classical binary conversion. Unlike the binary conversion which takes the conversion of the four nucleotides independently, the DB-curve representation takes two bases at a time. In this way, a DB-curve can conserve the biological feature. For example, the DB-curve of AC will represent the distribution of purine/pyrimidine and strong/weak hydrogen bonds along the sequences. Table 2 (see supplementary material) shows the different biological meanings of the four nucleotides respectively. Jiang and Yan (unpublished data) have shown that the DB-curve enhances the spectral contents in coding regions using Fourier transform incorporating information of three codon strand and phase compensation.

Another major point of this experiment is to show that the use of the FBLP-SVD algorithm improves the performance of the detection as compared to the Discrete Fourier

Transform (DFT). Evaluation is done on the proposed algorithm and the DFT analysis using both multi-scale and single scale spectral analysis. Improvement of results is noticed as the AUC of the FBLP-SVD method is larger than the AUC of the DFT in the multi-scale method and all single scales except at window size 360bp, in which there is a slight drop in the AUC. The drop is negligibly small compared to the increment. Figure 1(a) shows the ROC curve of the multi-scale method using FBLP-SVD (DB-curve and binary representation) and DFT (DB-curve). It is clear that the multi-scale FBLP-SVD using DB-curve representation is able to distinguish more exons from non-coding regions.

An example of the spectrum generated by our method and the DFT is illustrated in Figure 1(b). Both methods use multi-scale method and the spectra are smoothed with a Gaussian window. There are 18 exons with length in the range of 90bp-389bp embedded in the non-coding region of length 80bp-2113bp. As can be seen from the upper graph, our method produces much smaller responses in non-coding regions and the spectrum differences of the coding and non-coding regions are greater. The DFT produces a relatively less significant spectrum at the 1/3 frequency. Thus, it easier to identify the exons using FBLP-SVD compared to DFT.

We then try to separate the coding and non-coding region. To determine the threshold value that discriminates the coding regions from the non-coding regions, the cumulative distribution function is plotted (not shown). From the graph, the intersection of the two curves, which is 0.06214, is selected as the threshold. We obtain a sensitivity of 0.9090 and specificity of 0.6041.

An example of the splicing result showing five exons in gene X62654 is presented in Figure 1(c). The exons are seven exons of length 77bp-189bp which are embedded in the non-coding regions of variable length in the range of 94bp-935bp. The multi-scale FBLP-SVD using DB-curves is able to produce high P_{ratio} value at the 1/3 frequency for all the short exons and produce low P_{ratio} value for non-coding regions. In other words, it is able to discriminate the exon regions from the non-coding regions precisely.

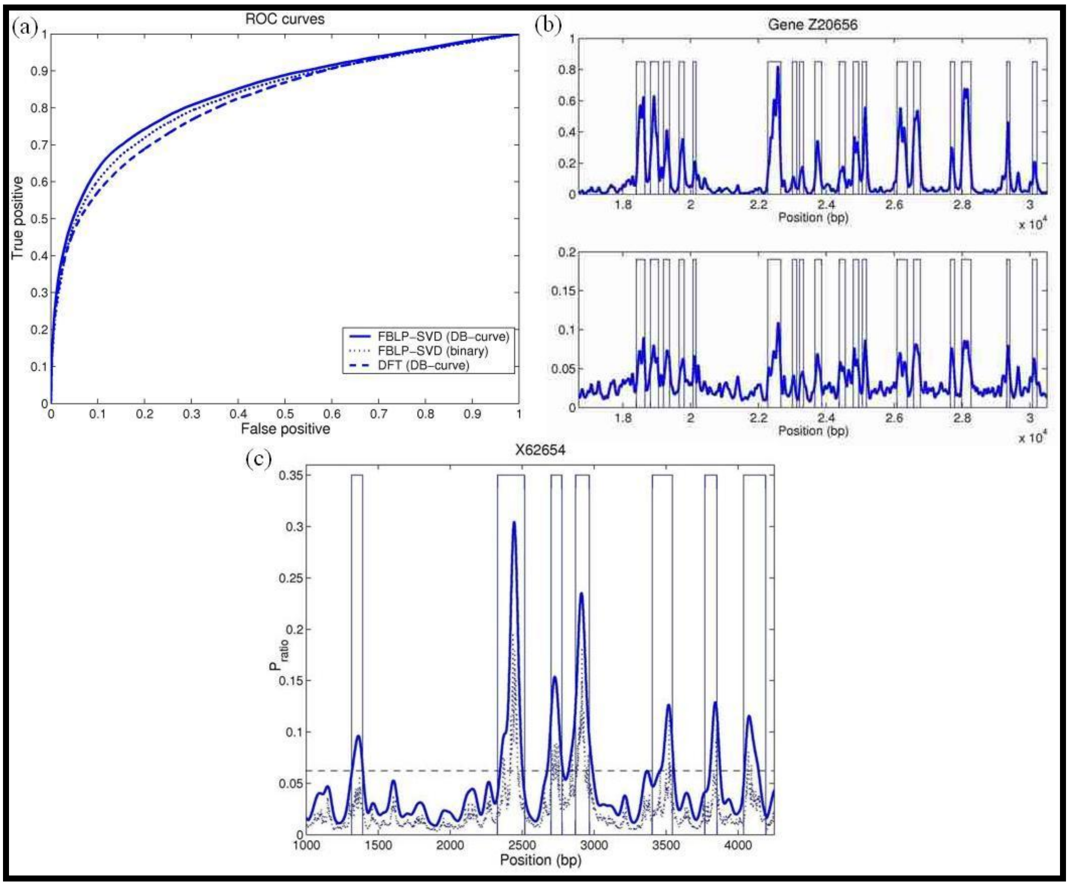

Figure 1: (a) ROC curves for the multiple-size moving window using FBLP-SVD using the DB-curve (solid line), FBLP-SVD using binary representation (dotted line) and the DFT using the DB-curves (dashed line). (b) Comparison of the FBLP-SVD and the DFT algorithm of gene Z20656. The length of the exons shown is in the range of 90bp-389bp embedded in non-coding regions of length in the range of 80bp-2113bp. The actual exon locations are marked with straight vertical lines. Upper diagram is the spectrum generated using FBLP-SVD whereas the lower diagram is generated from the DFT. (c) Graph of P_{ratio} along the gene X62654. The exons are 77bp in length starting at position 1316, 189bp at position 2328, 75bp at position 2701, 96bp at position 2870, 141bp at position 3403, 83bp at position 3768, and 153bp at position 4038. The dotted line represents the original P_{ratio} signal

along the gene. The solid line shows the Gaussian-smoothed signal. The actual exon locations are marked with straight vertical lines. The horizontal dashed line is the threshold value of 0.06214.

Conclusion:

We have proposed the multi-scale FBLP-SVD algorithm for exon detection in DNA sequences and carried out comparisons of multi-scale analysis with single scale methods, assessment of DB-curve representation and binary conversion, and evaluation of the FBLP-SVD algorithm and the DFT. The results have shown that the multi-scale FBLP-SVD algorithm with the DB-curve representation has a superior performance. Spectral analysis for the 3bp periodicity for exon detection is mostly based on Fourier transform in conventional methods. Our simulation results show that autoregressive model outperforms the Fourier transform. Besides, the majority of the methods proposed by other researchers are based on a single fixed window, or a single scale. The multi-scale approach is similar to wavelet analysis and offers a powerful means for detection of spectral components embedded in noise. In addition, we focus on the comparison of different periodicity estimation methods using spectral analysis algorithms without incorporating other criteria. Our method is based on the 3bp periodicity which is independent of any training datasets or database information. Thus it is more flexible and can be applied to DNA sequences obtained from different sources.

Acknowledgement:

The authors are thankful to Ms. June Jiang for useful discussions and for introducing some of the datasets for our simulations. This project is partially supported by a grant from the Hong Kong Research Grant Council (Project CityU 122005).

References:

[01] C. Burge & S. Karlin, *J Mol Biol.*, 268: 78 (1997) [PMID: 9149143]
[02] V. V. Solovyev, *et al.*, *Nucleic Acids Res.*, 22: 5156 (1994) [PMID: 7816600]
[03] W. C. Liew, *et al.*, *Int J Bioinformatics Research and Applications*, 1: 181 (2005)
[04] M. J. Shulman, *et al.*, *J Theor Biol.*, 88: 409 (1981) [PMID: 6456380]
[05] S. Tiwari, *et al.*, *Comput Appl Biosci.*, 13: 263 (1997) [PMID: 9183531]
[06] J. W. Fickett, *Nucleic Acids Res.*, 10: 5303 (1982) [PMID: 7145702]
[07] S. Eskesen, *et al.*, *BMC Mol Biol.*, 5: 12 (2004) [PMID: 15315715]
[08] D. Kotlar & Y. Lavner, *Genome Res.*, 13: 1930 (2003) [PMID: 12869578]
[09] M. Yan, *et al.*, *Bioinformatics*, 14: 685 (1998) [PMID: 9789094]
[10] D. W. Tufts, *et al.*, *Proc IEEE*, 70: 684 (1982)
[11] Y. Wu, *et al.*, *Chem Phys Lett.*, 367: 170 (2003)
[12] P. Yiou, *et al.*, *Physica D*, 142: 254 (2000)
[13] P. M. T. Broersen, *IEEE Trans Signal Process*, 48: 3550 (2000)
[14] S. W. Lang & J. H. McClellan, *IEEE Trans Acoust*, ASSP-28: 716 (1980)

Edited by P. Kangueane
Citation: Choong & Yan, Bioinformation 2(7): 273-278 (2008)

Supplementary material

Equations:

$$u_{DB}[n] = \sum_{i=1}^{n} x_{DB}[i] \quad n = 1,2,\ldots,N \qquad \longrightarrow \qquad (1)$$

where,

$$x_{DB}[n] = \begin{cases} +1 & \text{if } x[n] = \beta_1 \\ -1 & \text{if } x[n] = \beta_2 \\ 0 & \text{otherwise} \end{cases} \quad \begin{matrix} (DB \in \{AT, AC, AG, TC, TG, CG\}) \\ (\beta \in \{A, C, G, T\}) \end{matrix} \qquad \longrightarrow \qquad (2)$$

Instead of taking the cumulative x_{DB}, we take x_{DB}. Therefore, a DNA sequence is decomposed into six numerical seri consisting of 1 or 1.

The FBLP can be written as:

$$\begin{bmatrix} x[M] & x[M-1] & \ldots & x[1] \\ x[M+1] & x[M] & \cdots & x[2] \\ \vdots & \vdots & & \vdots \\ x[N-1] & x[N-2] & \cdots & x[N-M] \\ x^*[2] & x^*[3] & \cdots & x^*[M+1] \\ x^*[3] & x^*[4] & & x^*[M+2] \\ \vdots & \vdots & & \vdots \\ x^*[N-M+1] & x^*[N-M+2] & \cdots & x^*[N] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_M \end{bmatrix} = \begin{bmatrix} x[M+1] \\ x[M+2] \\ \vdots \\ x[N] \\ x^*[1] \\ x^*[2] \\ \vdots \\ x^*[N-M] \end{bmatrix} \qquad \longrightarrow \qquad (3)$$

where "*" denotes a complex conjugate.

$$P_x(\omega) = \frac{T\sigma_e^2}{|1 + \sum_{k=1}^{M} a_k e^{ijk\omega T}|^2} \qquad \longrightarrow \qquad (4)$$

where ω is the frequency, T is the sampling interval, σ_e^2 is the variance of the prediction error, M is the order of pre and a_k are the LP coefficients.

$$x(n) = [x_{AT}(n)\, x_{AC}(n)\, x_{AG}(n)\, x_{TC}(n)\, x_{TG}(n)\, x_{CG}(n)]^T \qquad \longrightarrow \qquad (5)$$

The power spectrum as constructed using (4) can be given by

$$\begin{aligned} |P(f)|^2 &= w_{AT}|P_{AT}(f)|^2 + w_{AC}|P_{AC}(f)|^2 + w_{AG}|P_{AG}(f)|^2 \\ &\quad + w_{TC}|P_{TC}(f)|^2 + w_{TG}|P_{TG}(f)|^2 + w_{CG}|P_{CG}(f)|^2 \end{aligned} \qquad \longrightarrow \qquad (6)$$

where *f* is the frequency index. $P_{DB}(f)$ are the power spectra of $x_{DB}(n)$ respectively, where DB∈{AT,AC,AG,TC,TG,(w_{DB} is the weight contribution from the six sequences. The weighted summation can be obtained by computing the ei decomposition of the six DB-curves and taking the largest eigenvalue.

$$P_{multi} = \frac{1}{W} \sum_{w=1}^{W} P_{ratio}^{w} \qquad \longrightarrow \qquad (7)$$

where *W* is the number of different window sizes or the number of scales.

Tables:

	Multiple-size moving window	Window size 60	Window size 90	Window size 180	Window size 360
FBLP-SVD (DB-curve)	0.8387	0.7888	0.8160	0.8374	0.8298
FBLP-SVD representation)	0.8154	0.7729	0.7934	0.8213	0.8159
DFT (DB-curve)	0.8258	0.7754	0.8023	0.8265	0.8319

Table 1: Comparison of AUC of multiple-size moving window and fixed window analysis of the proposed algorithm using DB-curve and binary representation, and DFT using the DB-curve.

Nucleotides	Biological meaning		
Adenine (A)	Purine	Weak hydrogen bond	Amino-type
Cytosine (C)	Pyrimidine	Strong hydrogen bond	Amino-type
Guanine (G)	Purine	Strong hydrogen bond	Keto-type
Thymine (T)	Pyrimidine	Weak hydrogen bond	Keto-type

Table 2: Biological meaning of the different nucleotides is given.