Volume 247
Issue 4
21 August 2007
ISSN 0022-5193
ELSEVIER
Journal of Theoretical Biology

Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Journal of Theoretical Biology 247 (2007) 687–694
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence

Changchuan Yin, Stephen S.-T. Yau*

Department of Mathematics, Statistics and Computer Science, The University of Illinois at Chicago, M/C 249, Chicago, IL 60607-7045, USA

Received 22 October 2006; received in revised form 24 March 2007; accepted 26 March 2007
Available online 10 April 2007

Abstract

With the exponential growth of genomic sequences, there is an increasing demand to accurately identify protein coding regions (exons) from genomic sequences. Despite many progresses being made in the identification of protein coding regions by computational methods during the last two decades, the performances and efficiencies of the prediction methods still need to be improved. In addition, it is indispensable to develop different prediction methods since combining different methods may greatly improve the prediction accuracy. A new method to predict protein coding regions is developed in this paper based on the fact that most of exon sequences have a 3-base periodicity, while intron sequences do not have this unique feature. The method computes the 3-base periodicity and the background noise of the stepwise DNA segments of the target DNA sequences using nucleotide distributions in the three codon positions of the DNA sequences. Exon and intron sequences can be identified from trends of the ratio of the 3-base periodicity to the background noise in the DNA sequences. Case studies on genes from different organisms show that this method is an effective approach for exon prediction.

Keywords: Exon; Intron; 3-Base periodicity; Fourier transform

1. Introduction

An important step in genomic annotation is to identify protein coding regions of genomic sequences, which is a challenging problem especially in the study of eukaryote genomes. In an eukaryote genome, protein coding regions (exons) are usually not continuous, but are flanked by noncoding regions (introns). Due to the lack of obvious sequence features between exons and introns, effectively distinguishing protein coding regions from noncoding regions is a challenging problem in bioinformatics.

During the last two decades, a variety of computational algorithms have been developed to predict exons (for reviews, Ficket and Tung, 1992; Fickett, 1996; Zhang, 2002; Mathé et al., 2002). Most of the exon-finding algorithms are based on statistics methods, which usually use training data sets from known exon and intron sequences to compute prediction functions. As examples, GenScan algorithm (Burge and Karlin, 1997) measured distinct statistics features of exons and introns within genomes and employed them in prediction via hidden Markov model (HMM); MZFF method (Zhang, 1997) was developed for predicting protein coding regions using quadratic discriminant analysis of different sequence characters of exons and introns. As combining different gene prediction methods may increase the accuracy of the prediction greatly, development of different effective gene prediction algorithms is one of the fundamental efforts in gene prediction study.

During recent years, signal processing approaches have been attracting significant attentions in genomic DNA research and have become increasingly important to elucidate genome structures because they may identify hidden periodicities and features which cannot be revealed easily by conventional statistics methods. After converting symbol DNA sequences to numerical sequences, signal processing tools, typically, discrete Fourier transform (DFT) or wavelet analysis, can be applied to the numerical vectors to study the frequency domain of the sequences (Anastassiou, 2000; Wang and Johnson, 2002; Kauer and

*Corresponding author. Tel./fax: +1 312 996 3065.
E-mail address: yau@uic.edu (S.S.-T. Yau).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.03.038

Blocker, 2003; Vaidanahan and Yoon, 2004). Using the signal processing methods, a variety of gene prediction algorithms have been developed (Tiwari et al., 1997; Anastassiou, 2000; Kotlar and Lavner, 2003; Jin, 2004; Gao et al., 2005). Tiwari et al. (1997) explored the measure of spectral content (SC) in DNA sequences based on the fact that the 3-base periodicity, identified as a pronounced peak at the frequency $N/3$ of the Fourier power spectrum of the DNA sequences (N is the length of the DNA sequence), is prevalent in most protein coding regions, but does not exist in noncoding regions (Tsonis et al., 1991; Voss, 1992; Chechetkin and Turygin, 1995; Dodin et al., 2000). Anastassiou (2000) presented an optimized SC measure of DNA sequences for gene prediction. Kotlar and Lavner (2003) utilized spectral rotation measure based on the arguments of the DFT to develop a novel gene prediction algorithm, which was later improved by Jin (2004). Gao et al. (2005) combined the 3-base periodicity and the fractal features of DNA sequences to improve gene prediction methods.

Most of the DFT based gene finding algorithms use a short-sequence window approach (Tiwari et al., 1997; Yan et al., 1998; Anastassiou, 2000), in which a fixed-length window is used to slide a DNA sequence to compute the Fourier power spectrum. However, this approach has limitations. A small window frame causes more statistical oscillation, resulting in more prediction errors, whereas a large window frame may miss small exons or introns. The arbitrary choices of window size made the short-sequence window Fourier technique subject to bias. Furthermore, the short-sequence window Fourier transform requires much longer CPU time. It becomes a challenging problem when finding genes for whole genomes as direct computation of Fourier transforms is time consuming.

It was demonstrated that the 3-base periodicity in a DNA sequence is partly caused by the unbalanced nucleotide distributions in the three coding positions in the sequence (Fickett, 1982; Ficket and Tung, 1992; Tiwari et al., 1997; Yin and Yau, 2005). In an exon sequence, the nucleotide distribution in the three codon positions is unbalanced, while in an intron sequence, the nucleotides distribute uniformly in the three codon positions. The reason of the unbalanced distribution is that proteins prefer special amino acid compositions and thus nucleotide usage in a coding region is highly biased (Ficket and Tung, 1992; Tiwari et al., 1997; Yin and Yau, 2005). This paper presents an extension of the current gene prediction algorithms (Tiwari et al., 1997; Anastassiou, 2000), called EPND method (exon prediction via nucleotide distributions), which is based on the peak at the frequency of $N/3$ of the DFT and the frequencies of the nucleotides in the three codon positions (position asymmetry measure) within known genes. The algorithm is tested for identifying exons/introns within known genes from several organisms in this paper. Case studies indicate that the method described in this paper is an effective protein coding region prediction method in terms of accuracy and efficiency.

2. Methods and algorithms

2.1. Fourier spectrum analysis of DNA sequences

A symbolic DNA sequence, denoted as, $x(0), x(1), \ldots, x(N-1)$, is first converted to four binary indicator sequences, $u_A(n), u_T(n), u_C(n)$, and $u_G(n)$, which indicate the presence or absence of four nucleotides, A, T, C, and G, at the nth position, respectively (Voss, 1992; Tiwari et al., 1997; Anastassiou, 2000). For instance, the indicator sequence, $u_A(n) = 0001010111\ldots$, indicates that the nucleotide A is in the positions 4, 6, 8, 9, and 10 of the DNA sequence.

The DFT converts a signal in the signal domain to a set of new values in the frequency domain. For a signal of length N, $f(n), n = 0, 1, \ldots, N-1$, its DFT is defined as follows:

$$F(k) = \sum_{n=0}^{N-1} f(n) e^{-i\frac{2\pi nk}{N}} \tag{2.1}$$

where $i = \sqrt{-1}$. The DFT power spectrum of a signal at the frequency k is defined as:

$$PS(k) = |F(k)|^2, \quad k = 0, 1, 2, \ldots, N, \tag{2.2}$$

where $F(k)$ is the kth DFT coefficient.

The DFT power spectrum of a DNA sequence is the sum of the power spectrum of its four binary indicator sequences (Silverman and Linsker, 1986; Tiwari et al., 1997; Anastassiou, 2000):

$$PS(k) = PS_A(k) + PS_T(k) + PS_C(k) + PS_G(k) \tag{2.3}$$

where $PS_A(k), PS_T(k), PS_C(k)$ and $PS_G(k)$ are the Fourier power spectrum of the four indicator sequences $u_A(n), u_T(n), u_C(n)$ and $u_G(n)$, respectively. Due to the symmetry property of the DFT spectrum of real number signals, the figures in this paper only show half of the Fourier spectrum of DNA sequences.

2.2. Computing the 3-base periodicity and background noise from nucleotide distributions of a DNA sequence

The asymmetry in the nucleotide distributions in the three codon positions and its connection to the DFT peak in $N/3$ at coding regions were addressed by Ficket (Fickett, 1982; Ficket and Tung, 1992). The 3-base periodicity magnitude and background noise can be directly computed from the nucleotide distributions (Ficket and Tung, 1992; Yin and Yau, 2005). Let F_{x1}, F_{x2}, F_{x3} be the occurrence frequencies of the nucleotide $\mathbf{x} \in \{A, T, C, G\}$ in the first, the second and the third codon positions, respectively. Then the 3-base periodicity magnitude can be computed as follows:

$$PS(N/3) = \sum_{x=A,T,C,G} [F_{x1}^2 + F_{x2}^2 + F_{x3}^2 - (F_{x1} * F_{x2} + F_{x1} * F_{x3} + F_{x2} * F_{x3})]. \tag{2.4}$$

The background noise of a DNA sequence of length N, represented as the average power spectrum E over all the frequencies, is determined mainly by the length of the DNA sequence (Yin and Yau, 2005). Thus, the ratio of the 3-base periodicity signal to the background noise of a DNA sequence, denoted as $SN(N)$, is defined as follows:

$$SN(N) = \frac{PS(N/3)}{N}. \tag{2.5}$$

$SN(N)$ can be interpreted as strength of the 3-base periodicity per nucleotide. Based on the computational simulation of computer generated sequences and verified with 12 exons/introns from *Yeast* and *C. elegans*, it was shown that $SN(N)$ is equal to or larger than 2 for most exon sequences (Yin and Yau, 2005, also refer to Fig. 3), while it is less than 2 for most intron sequences. The threshold value of the signal-to-noise is set to 2 in the gene finding algorithm in this paper.

2.3. Algorithm for exon prediction by nucleotide distribution (EPND)

For a DNA sequence of length N, let D_k denote the DNA walk sequence of length k, i.e., D_k is the sub-region of the DNA sequence ranging from the beginning to the position k. To find exon regions and intron regions within the given DNA sequence, the EPND algorithm is developed as follows, and the flow chart of the algorithm below is shown in Fig. 1.

1. Set $k = 1$.

2. Compute nucleotide distributions of D_k in the three codon positions of F_{xi} ($\mathbf{x} \in \{A, T, C, G\}$, $i \in \{1, 2, 3\}$). The nucleotide distribution of a DNA walk sequence of length k can be obtained recursively from the DNA walk sequence of length $k-1$ with the occurring frequencies of the nucleotides on the position k.

3. Compute the magnitude of the 3-base periodicity $PS(k/3)$ in D_k based on the formula (2.4).

4. Compute the ratio of 3-base periodicity to background noise, $SN(k) = PS(k/3)/k$, within the DNA sequence D_k.

5. Increase k by 1 and repeat step 2 to step 4, until $k = N$.

6. Compute the slope of SN at each position on the SN plot as follows: since most of the exon or intron sequences in a genome are longer than 50 base pairs, the slope at the ith position is computed as $(SN(i) - SN(i-50))/50$, where i is from 51 to N.

7. Set the nucleotide at each position to exon or intron region as follows: if the slope at the position is larger than 0 and SN is larger than or equal to 2, set the nucleotide at the position as exon nucleotide; otherwise, set it as intron nucleotide.

8. Reduce local noise. If a DNA region less than 50 base pairs is identified as an intron from step 7, and is flanked by two exon regions, this region is often a false negative, and is reset as exon region; similarly, if a DNA region less than 50 base pairs is identified as an exon from step 7, and is flanked by two intron regions, this region is often a false positive, and is reset as an intron region.

2.4. Improving prediction accuracy using different starting points

For a long DNA sequence that may contain more than two exons (or two introns), such as exon–intron–exon, the accumulated signal-to-noise ratio of the last exon will become low especially when the intron in between is long, which may affect the accuracy of the prediction. It would improve the algorithm if we divide a DNA sequence into different sub regions. In addition, to reduce false exons and false introns, we apply the algorithm at different arbitrary starting points so that each nucleotide may be tested multiple times. The following algorithm is developed to improve exon prediction accuracy when using EPND method:

1. If a DNA sequence is longer than 2000 base pairs (bp), divide it to sub-sequences of 2000 base pairs.

2. For each 2000 base pairs sub-sequence, set $P_1 = 1$, $P_2 = 401, P_3 = 801, P_4 = 1201, P_5 = 1601$ and $P_6 = 2000$ be the six even-spaced points.

3. Identify exon or intron nucleotides using EPND method on the sub-sequence between point P_i and P_6 where $i = 1, 2, 3, 4, 5$. So each nucleotide after points P_3 is tested at least three times using EPND method from different start points. A nucleotide is identified as an exon nucleotide when it is predicated in an exon region in the majority of the tests.

2.5. Database and measures for performance evaluation

The data set used for the evaluation of the performance of the EPND method is Xpro (Gopalan et al., 2004), which contains the eukaryotic protein coding DNA sequences from GeneBank release 139. The data set was downloaded from the Xpro web site as flat files (Xpro version: v.1.2, 2004, http://origin.bic.nus.edu.sg/xpro). One file, *exonse—intron*—139.*gz*, contains protein coding regions (exons), and the other file, *intronseq—intron*—139.*gz* contains non-protein coding regions (introns). Both files consist of DNA sequences and the corresponding header information which indicates gene locus, organism that the genes belong to, intron lengths and their corresponding positions within the genes. Based on the intron positions in the header sections, introns are conjugated with the corresponding exons to form a full original gene structure beginning with a start codon and ending with a stop codon. The full length genes are used in this study to test algorithm performance.

The performance of the EPND algorithm is measured in terms of sensitivity, specificity and accuracy, which are defined in the literature as follows (Burset and Guigo, 1996). The sensitivity, $S_n = TP/(TP + FN)$, and the specificity, $S_p = TN/(TN + FP)$, where TP is the true positive, which is the length of nucleotides of correctly

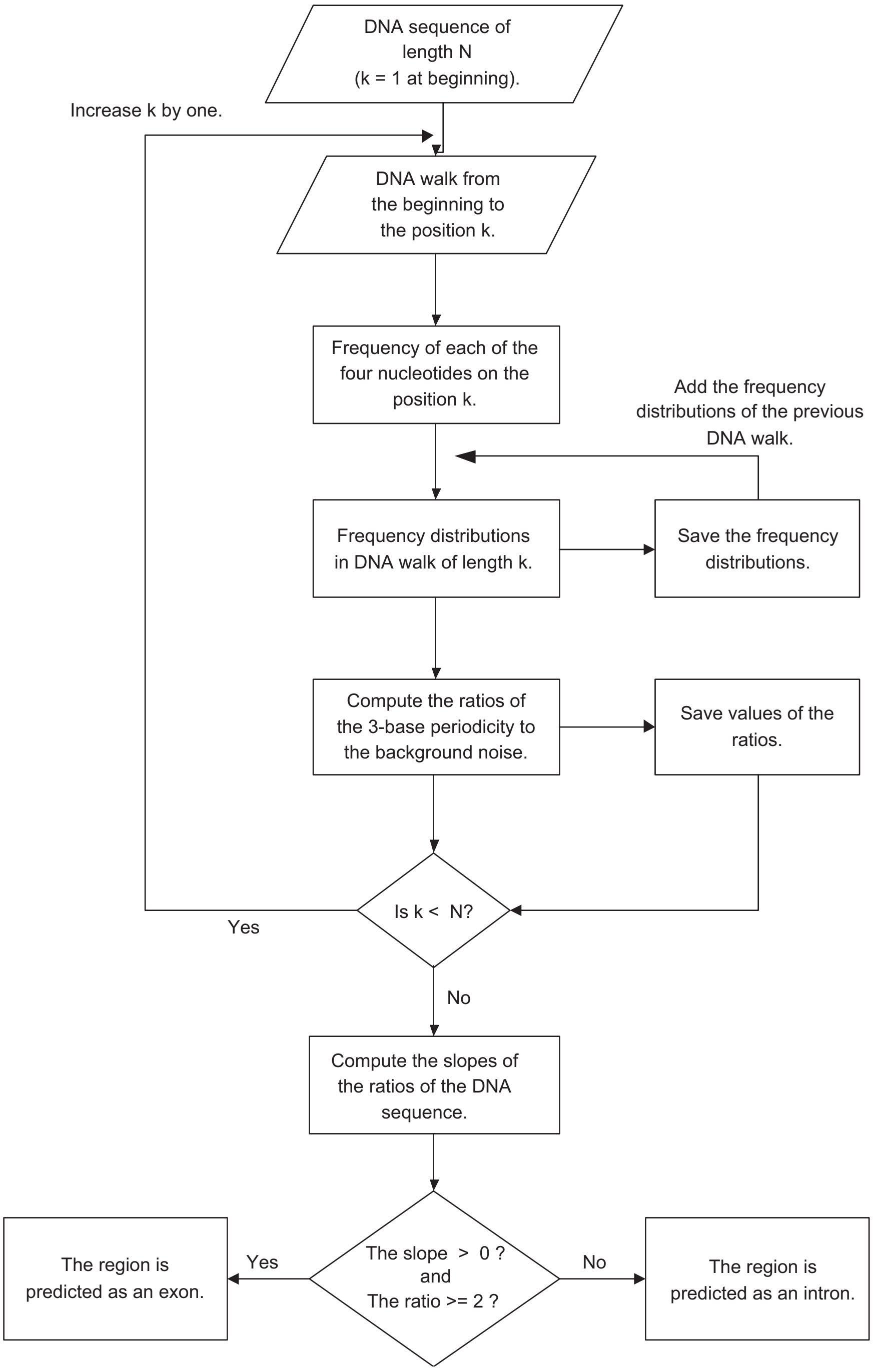
DNA sequence of length N (k = 1 at beginning).
Increase k by one.
DNA walk from the beginning to the position k.
Frequency of each of the four nucleotides on the position k.
Add the frequency distributions of the previous DNA walk.
Frequency distributions in DNA walk of length k.
Save the frequency distributions.
Compute the ratios of the 3-base periodicity to the background noise.
Save values of the ratios.
Is k < N?
Yes
No
Compute the slopes of the ratios of the DNA sequence.
The slope > 0 ? and The ratio >= 2 ?
Yes
No
The region is predicted as an exon.
The region is predicted as an intron.

Fig. 1. Flow chart of the EPND exon finding algorithm.

predicted exons; *TN* is the true negative, which is the length of nucleotides of correctly predicted introns; *FN* is the false negative, which is the length of nucleotides of wrongly predicted introns; and *FP* is the false positive, which is the length of nucleotides of wrongly predicted exons. In other words, S_n is the proportion of coding sequences that have been correctly predicted as coding, and S_p is the proportion of noncoding sequences that have been correctly predicted as noncoding. The accuracy *AC* is defined as the average of S_n and S_p.

3. Results and discussions

3.1. Features of the signal-to-noise ratios of DNA walks from exons and introns

The 3-base periodicity uniquely exists in most exon sequences, but it is not in the majority of intron sequences. Thus, there is a pronounced peak in the Fourier spectrum of an exon sequence. As an example to illustrate the 3-base periodicity, Fig. 2 is the Fourier spectrum of an exon

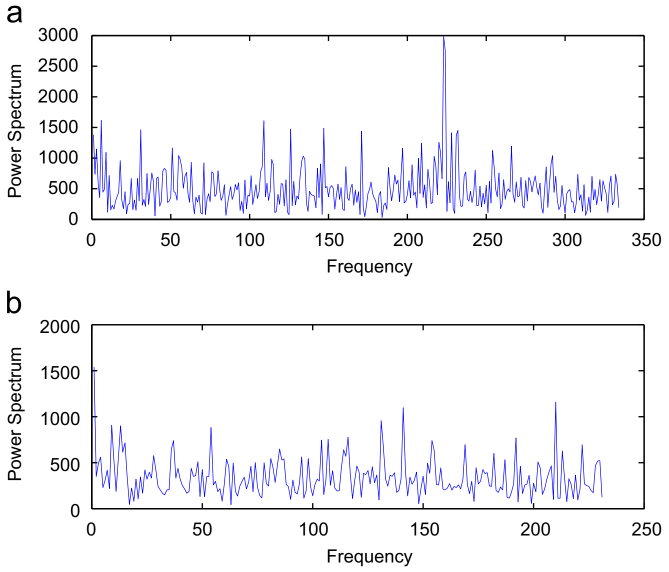

Fig. 2. DFT power spectrum of an exon (a) and an intron (b) from the gene of AAM70818.2 of *Drosophila melanogaster* (*fruit fly*).

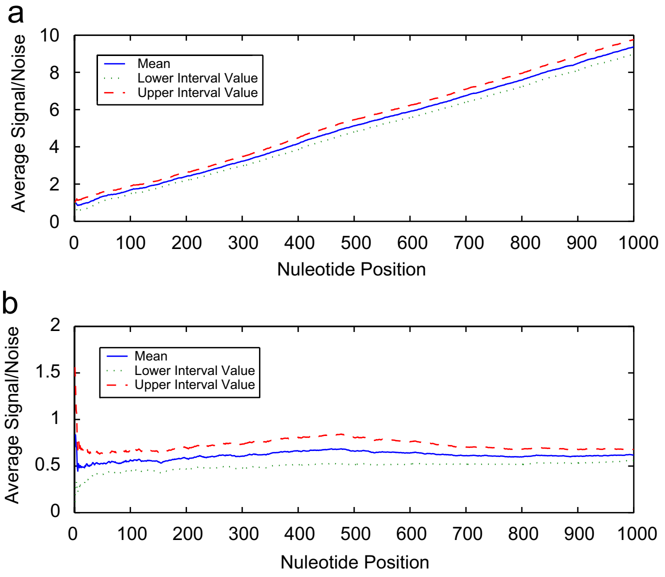

Fig. 3. Plots of the average signal-to-noise ratios of the DNA walks of 1000 bp DNA fragments of 258 exons and 216 introns from human genome. (a) The average signal-to-noise ratios of 258 exons (the middle plot) and the 95% confidence intervals of the average ratios (the lower and upper plots). (b) The average signal-to-noise ratios of 216 introns (the middle plot) and the 95% confidence intervals of the average ratios (the lower and upper plots).

sequence and an intron sequence from the gene AAM70818.2 of *Drosophila melanogaster* (*fruit fly*) (the exon sequence is located from 1 to 670, and the intron from 671 to 1135), there is a distinct peak at the frequency $N/3$ ($N = 670$) in the Fourier power spectrum in the exon DNA sequences. But the peak does not exist in the spectrum for the intron DNA sequences.

The strength of the 3-base periodicity can be measured by computing the Fourier power spectrum at the frequency $N/3$. However, the computational complexity of Fourier transforms is expensive when a DNA sequence becomes large. To measure the strength of the 3-base periodicity of a DNA sequence, EPND method uses the nucleotide distributions in the three codon positions (Yin and Yau, 2005). To measure the background noise of the Fourier spectrum, EPND method uses the length of the DNA sequence. To associate the signal-to-noise ratio $SN(k)$ to the gene structures, $SN(k)$ values from the stepwise DNA sequences are plotted versus the nucleotide positions k. Fig. 3(a) is the plot of the averages $SN(k)$ of DNA walks of 1000 base pairs fragment of 258 exon sequences from human genome, and Fig. 3(b) is the plot of the averages $SN(k)$ of DNA walks of 1000 base pairs fragments of 216 intron sequences from human genome. The results in Fig. 3 indicate that for exon regions, with the increase of the length of the DNA walk sequences, the signal-to-noise ratios of the DNA walks from most exon sequences are increased, which show strengthened 3-base periodicity signals. On the other hand, for the intron regions, due to the absence of the 3-base periodicity, the signal-to-noise ratios are randomly fluctuated around some low values with the increase of DNA walks. It also shows that the $SN(k)$ values for exons are larger than 1, while those for introns are less than 1. In addition, we also notice that short exons may have signal-to-noise ratios less than 1 and short introns may have large signal-to-noise ratios. The 95% confidence intervals of the mean SN values of exons and introns are plotted in the figure. For instance, the average SN value and the 95% confidence intervals of the mean for 500 bp exon sequences are 5.1223 and 4.7960, 5.4486, respectively. The confidence intervals on the mean of SN of exons and introns have narrow ranges, indicating that the average SN values obtained from this study have reasonable accuracy. As for an example, an exon and an intron from gene 1J942 of *C.elegans* (the exon sequence is located from position 114 to position 374, and the intron is located from position 845 to position 1027) are chosen for the test. Fig. 4(a) is the plot for the exon and Fig. 4(b) is the plot for the intron. It indicates that an exon displays an upward trend in the plot of $SN(k)$ ratio versus position, while an intron displays a flat trend. Thus, for a given DNA sequence, putative exons and introns can be identified from the plot of signal-to-noise ratios versus the nucleotide lengths.

3.2. Identification of exon and intron regions by the EPND method

To test the feasibility of the EPND algorithm in protein coding region prediction, the intron fragments and the exon fragments of test genes from different organisms are selected. The test data sets used in the performance evaluation are the full length gene sequences containing both introns and exons, recovered by joining intron and

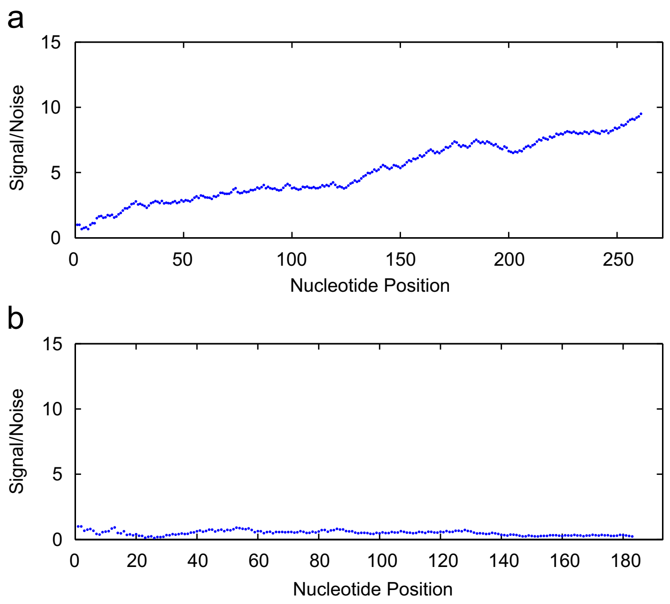

Fig. 4. Plots of the signal-to-noise ratios of the DNA walk sequences of an exon and an intron by EPND method. (a) Exon, (b) Intron.

exon sequences based on the intron positions in the database files. The exon prediction tests are based on the demarcated genes in which the introns and exons are not known. The full length gene sequences are applied to EPND program to validate the program. As a typical example, Fig. 5(a) is the plot of $SN(k)$ versus the nucleotide positions k of a test gene with a structure of exon-intron-exon (the gene locus is at AAB26989.1 of *Drosophila melanogaster* (*fruit fly*)). Fig. 5(b) is the slope plot from SN. It indicates that most of the slopes of the exon regions are positive, whereas the slopes of introns are negative. Fig. 5(c) is the expected gene structures that are verified by biological experiments. Fig. 5(d) is the predicted gene structures without applying the improvement algorithm. The values of S_n, S_p and AC of the test in Fig. 5(d) are 0.8684, 0.4372 and 0.6528, respectively. Fig. 5(e) is the predicted gene structure after applying the improvement algorithm. The values of S_n, S_p and AC of the test in Fig. 5(e) are 0.9450, 0.7556 and 0.8503, respectively. The average accuracy (AC) of exon prediction is improved for 19.75% in this test case. Generally, compared with the original EPND method, the improvement algorithm may increase the average prediction accuracy for 1% based on the test results on 643 human genes. The result shows that the majority of exon and intron sequences are effectively identified though short sequences are identified as false positive or false negative. There are approximate changing points between exons and introns in the plot, which indicate that this method can identify the approximate regions of the exon and intron fragments. However, as there are also changes in the slope within the intron, it is difficult to identify the accurate boundaries between exon and intron regions. This issue is under further investigation.

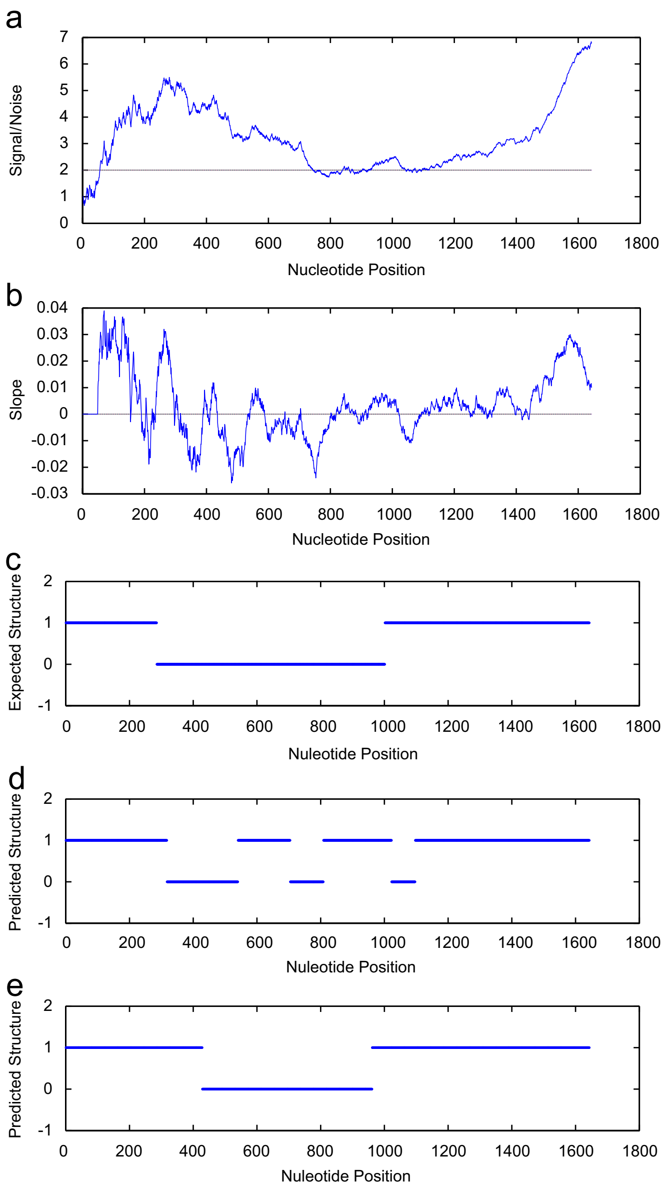

Fig. 5. Gene structure prediction by EPND method. The gene locus is at AAB26989.1 of *Drosophila melanogaster* (*fruit fly*). (a) The signal-to-noise ratios, SN, of the DNA walks from this sequence calculated by EPND method. (b) Plot of the slopes of every two points at a distance of 50 base pair from the SN plot. (c) The expected gene structure that is verified by biological experiments. Exon regions are marked as 1, and intron regions are marked as 0. (d) The predicted gene structure by EPND method. (e) The predicted gene structure by EPND method with overlapping improvement.

3.3. *Performance evaluation of the EPND algorithm*

The detailed results on gene sets of known structures of several genomes including *H. sapiens* (*human*), *Drosophila* (*Fruit fly*) and *Arabidopsis*, are shown in Table 1. The table indicates that accuracy of the EPND program at the nucleotide level can be 0.8149 when a set of full length genes from the *Drosophila*(*Fruit fly*) genome is tested. The accuracy of this method is better than or comparable to

Table 1
Summary of the performance evaluation of EPND program

Organism	S_n	S_p	AC	Genes
H. sapiens (*human*)	0.6526	0.9613	0.8070	643
Drosophila (*Fruit fly*)	0.7158	0.9140	0.8149	1101
Arabidopsis	0.5398	0.9393	0.7400	2771

other gene prediction programs (Burset and Guigo, 1996; Rogic et al., 2001).

Compared with two currently available algorithms, SC measure (Tiwari et al., 1997; Yan et al., 1998; Anastassiou, 2000) and rotational SC measure (Kotlar and Lavner, 2003; Jin, 2004) of the DFT, the method presented in this paper is an improvement of the two algorithms for applying the measures to exon prediction. The method described in this paper has the following three features: (1) the important feature of our method is to use extendable windows to measure the 3-base periodicity and compute the slopes of the trends of the signal-to-noise ratios of DNA walks in 50 base pair window distance, which reduces the bias when fixed length windows are used. (2) This method explores the signal-to-noise ratio as an alternative measure to distinguish exon and intron. Both signal and noise of the spectral content are not computed directly from the DFT, where they are computed from the count of frequencies of four nucleotides in the three coding positions, which reduces the bias when fixed window lengths are used. The computation of magnitude of the 3-base periodicity is based on nucleotide distributions on the three coding positions. The computation of the nucleotide distributions on the DNA walk sequences uses a recursive approach in which computation of nucleotide distributions on the DNA sequence of length k uses the results of the nucleotide distributions on the $k-1$ length DNA segment. In terms of computational complexity, the algorithm has a linear computation time proportional to the length of the DNA sequence, which is very efficient. (3) This method only requires small training data set to obtain the signal-to-noise threshold value, whereas other statistics gene finding methods require large training sets from which different statistics parameters are derived. Thus, this method is of importance to predict genes especially when the information on the known genes in a genome is limited.

The algorithm described here, while offering a level of predictive accuracy that is comparable to other methods, has limitations that need to be addressed. The algorithm may not easily identify a short exons as short exons may yield weak Fourier spectrum signals and the signal-to-noise ratios have many statistical fluctuations. In addition, if a short intron is located between a long exon and a short exon, the signal-to-noise ratio of the intron segment may still be positive, and the intron may be predicted as an exon (false positive). The improvement of the limitations is under further investigation.

4. Conclusion

In this paper, an improved method to predict exon and intron locations within genes has been proposed based on the nucleotide distribution in the three codon positions. The extensible windows approach is employed in the method to avoid the bias caused by short time Fourier transform method, improving the performance of the computation significantly. Case studies indicate that the method described in this paper is an effective method to predict protein coding regions in terms of accuracy and efficiency.

Acknowledgments

We wish to thank Professor Jiangsong Wang with the Department of Mathematics of Nanjing University, China, for his valuable discussion on this paper. We also appreciate Dr. Rong He (Lucy) with the Department of Pharmacology of University of Illinois at Chicago, for providing part of the DNA sequences. We thank all the anonymous referees for their very valuable suggestions, and Bridget Flynn for editing our English.

References

Anastassiou, D., 2000. Frequency-domain analysis of biomolecular sequences. Bioinformatics 16, 1073–1081.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78–94.

Burset, M., Guigo, R., 1996. Evaluation of gene structure prediction programs. Genomics 34, 353–367.

Chechetkin, V.R., Turygin, A.Y., 1995. Size-dependence of three-periodicity and long-range correlations in DNA sequences. Phys. Lett. A 199, 75–80.

Dodin, G., Vandergheynst, P., Levoir, P., Cordier, C., Marcourt, L., 2000. Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. J. Theor. Biol. 206, 323–326.

Ficket, J.W., Tung, C.S., 1992. Assessment of protein coding measures. Nucleic Acids Res. 20, 6441–6450.

Fickett, J.W., 1982. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. 10, 5303–5318.

Fickett, J.W., 1996. The gene identification problem: an overview for developers. Comput. Chem. 20, 103–118.

Gao, J., Qi, Y., Cao, Y., Tung, W.W., 2005. Protein coding sequence identification by simultaneously characterizing the periodic and random features of DNA sequences. J. Biomed. Biotechnol. 2, 139–146.

Gopalan, V., Tan, T.W., Lee, B.T., Ranganathan, S., 2004. Xpro: database of eukaryotic protein-encoding genes. Nucleic Acids Res. 32, D59–D63.

Jin, J., 2004. Identification of protein coding regions of rice genes using alternative spectral rotation measure and linear discriminant analysis. Genomics, Proteomics & Bioinformatics 2, 167–173.

Kauer, G., Blocker, H., 2003. Applying signal theory to the analysis of biomolecules. Bioinformatics 19, 2016–2021.

Kotlar, D., Lavner, Y., 2003. Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. Genome Res. 13, 1930–1937.

Mathé, C., Sagot, M.-F., Schiex, T., Rouzé, P., 2002. Current methods of gene prediction, their strength and weaknesses. Nucleic Acids Res. 30, 4103–4117.

Rogic, S., Mackworth, A.K., Ouellette, F.B.F., 2001. Evaluation of gene-finding programs on Mammalian sequences. Genome Res. 11, 817–832.

Silverman, B.D., Linsker, R., 1986. A measure of DNA periodicity. J. Theor. Biol. 118, 295–300.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattachrya, S., Ramaswamy, R., 1997. Prediction of probable genes by Fourier analysis of genomic sequences. CABIOS 113, 263–270.

Tsonis, A.A., Elsner, J.B., Tsonis, P.A., 1991. Periodicity in DNA coding sequences: implications in gene evolution. J. Theor. Biol. 151, 323–331.

Vaidanahan, P.P., Yoon, B.J., 2004. The role of signal-processing concepts in genomics and proteomics. J. Franklin Inst. 1, 1–27.

Voss, R., 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. Phys. Rev. Lett. 68, 3805–3808.

Wang, W., Johnson, D.H., 2002. Computing linear transforms of symbolic signals. IEEE Trans. Signal Process. 50, 628–634.

Yan, M., Lin, Z.S., Zhang, C.T., 1998. A new Fourier transform approach for protein coding measure based on the format of Z curves. Bioinformatics 14, 685–690.

Yin, C., Yau, S.S.-T., 2005. A Fourier characteristic of coding sequences: origins and a non-Fourier approximation. J. Comput. Biol. 9, 1153–1165.

Zhang, M.Q., 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc. Natl Acad. Sci. USA 94, 565–568.

Zhang, M.Q., 2002. Computational prediction of eukaryotic protein-coding genes. Nature (Genetics) 3, 698–710.