Identification of Protein Coding Regions of Rice Genes Using Alternative Spectral Rotation Measure and Linear Discriminant Analysis

Jiao Jin[1,2]*

[1] *Department of Statistics and Financial Mathematics, School of Mathematical Sciences, Beijing Normal University, Beijing 100875;* [2] *Beijing Genomics Institute, Beijing 101300, China.*

An improved method, called Alternative Spectral Rotation (ASR) measure, for predicting protein coding regions in rice DNA has been developed. The method is based on the Spectral Rotation (SR) measure proposed by Kotlar and Lavner, and its accuracy is higher than that of the SR measure and the Spectral Content (SC) measure proposed by Tiwari *et al.* In order to increase the identifying accuracy, we chose three different coding characters, namely the asymmetric, purine, and stop-codon variables as parameters, and an approving result was presented by the method of Linear Discriminant Analysis (LDA).

Key words: Alternative Spectral Rotation measure, DFT, nonparametric fitting, LDA

Introduction

Although improvements in computer gene-finding programs have made it relatively easy to detect genes in uncharacterized genomic DNA sequences, it remains difficult to determine how many exons and introns there are in a given sequence and what are the exact boundaries between them. As we know, gene identification methods may be classified as recognition of protein coding regions and recognition of functional sites of genes. In the past two decades, many new methods for finding distinctive features of protein coding regions have been presented, including the algorithms based on codon usage (*1*), dicodon usage (*2*), 3-base periodicity (*3–5*), and the fifth-order phase Markov chain model (*6*). Although great progress has been made, the situation is still far from being perfect. Undoubtedly, the fifth-order Markov chain model has a better identification accuracy, since this method makes full use of the local statistical characteristics of base distribution in three frames of coding sequences. However, it still has its shortcomings; the parameters determined based on previously discovered sequences cannot be applied to identify genes on different sequences with the same accuracy (*7*). Moreover, it needs a large data set to train the bulky parameters, whose number is nearly five thousand. In recent years, several new algorithms have been proposed, such as MZEF (*8*), GLIMMER (*9*), MORGAN (*10*), GeneMark.hmm (*11*), GENESCAN (*12*), FGENESH (*13*), and so on (*14*, *15*). An up-to-date list of references is maintained by Wentian Li (http://www.nslij-genetics.org/gene/; ref. *16*). And a powerful gene finding program, BGF (Beijing Gene Finder), is proposed by Beijing Genomics Institute (http://bgf.genomics.org.cn/). These algorithms, which use both coding information and splicing signals, perform better than those using only splicing signals (*17*). However, there is still the need of new methods for gene prediction, which utilize features of gene structure that have so far not been incorporated into programs already available (*7*).

In this paper, we propose a new Alternative Spectral Rotation (ASR) measure derived by inverting the Spectral Rotation (SR) measure (*5*). Our method is based on the arguments of the Discrete Fourier Transform (DFT). After the DFT procedure for the four nucletides A, C, G and T, we found that the distributions of arguments C and T seem to have two central values. A cutoff value is decided after the nonparametric fitting and the arguments for all experimental genes are separated into two parts in the cases C and T. So we could select the corresponding central value to rotate clockwise according to the cutoff. This method performs better than the SR measure and the Spectral Content (SC) measure (*3*). In

*** Corresponding author.**
E-mail: jinj@genomics.org.cn

order to increase the identifying accuracy, especially in short exons, we selected three different features of coding regions, namely the asymmetric, purine, and stop-codon variables, which are simple but effective as variables in discriminant. A satisfied prediction result was obtained by the method of Linear Discriminant Analysis (LDA).

Despite the extensive research in the area of gene prediction, current predictors do not provide a complete solution to the problem of gene identification. Short exons are difficult to locate, because discriminative statistical characteristics are less likely to appear in short strands (*5*). The method proposed in this paper is shown to be a potential candidate for locating short genes and exons. We hope that this measure could be incorporated into the gene-finding programs already available and the gene prediction accuracy could be increased.

Databases

We have two data sets used in this paper. One data set with 5,047 sequences was used to train the argument distributions both for coding and noncoding regions. The other consisting of 704 sequences was used for selecting the subsets, which were used to test the identifying accuracy by means of ASR and LDA. The first data set was selected from the KOME full-length rice cDNA. After seeking the best open reading frame (ORF) by dynamic programming, mapping the cDNAs with ORF fixed to BAC sequence in GeneBank, removing redundancy and discarding the sequences that have in-frame stop codons or non-canonical sites, there were 5,047 sequences remained (*19*). The second data set was from GenBank R132. All the rice sequences we chose were marked with "CDS" and "mRNA". After removing redundancy and making full length, there were 704 sequences remained. The two data sets have few redundance, so we chose the first as the training set and the second as the test set.

From the 704 sequences, we extracted all exons and concatenated them to single strands (complementary strand had been changed to forward strand already), thus obtained 704 coding sequences. We also extracted all introns from the 581 multiple-exon genes (there were 123 single genes in the 704 sequences) and got 581 noncoding sequences. The data sets including coding sequences or noncoding fragments were obtained by sliding windows of sizes 90, 120, 180, 240, 300, and 351 bp.

Alternative Spectral Rotation Measure

DFT and SR measure

It is well known that the DFT of a given numeric sequence $x(n)$ of length N is defined by

$$X(k) = DFT\{x(n)\}_{n=0}^{N-1} = \sum_{n=0}^{N-1} x(n)e^{-i\frac{2\pi}{N}nk}, \quad 0 \le k \le N-1 \tag{1}$$

where n is the sequence index (*5*). The DFT itself is another sequence $X(k)$ of the same length N. The sequence $X(k)$ provides a measure of the period at K, which corresponds to a period of N/K samples (*18*).

Because the DNA sequence is a character string, we must assign proper numerical values to each character: A, C, G and T. We assign a binary sequence to each of the four bases (*4*). For example, we have a DNA sequence $x(n) = \{AACGCTAT\cdots\}$, the resulting numeric sequences are

$$x(n) = \{AACGCTAT\cdots\} \rightarrow \begin{cases} u_A(n) = 11000010\cdots \\ u_C(n) = 00101000\cdots \\ u_G(n) = 00010000\cdots \\ u_T(n) = 00000101\cdots \end{cases}$$

Here, $u_b(n)$ (b = A, C, G, or T) is the binary sequence, which takes the value of 1 or 0 at position n, depending on whether or not the character b exists at location n.

So we could define the DFT of the binary sequence $u_b(n)$ of length N as

$$U_b(k) = \sum_{n=0}^{N-1} u_b(n)e^{-i\frac{2\pi}{N}nk}, \quad 0 \le k \le N-1 \tag{2}$$

The total frequency spectrum of the given DNA character string is described as

$$S(k) = \left|U_A(k)\right|^2 + \left|U_C(k)\right|^2 + \left|U_G(k)\right|^2 + \left|U_T(k)\right|^2$$

As we know, the protein coding regions have a feature of 3-base periodicity (*3*), so the total Fourier spectrum of protein coding DNA typically has a peak at frequency $k = N/3$. It is very important for us to get the $(N/3)$th element of the DFT of the binary

sequence $u_b(n)$ of length N associated with base b (b = A, C, G, or T):

$$U_b(\frac{N}{3}) = \sum_{n=0}^{N-1} u_b(n) e^{-i\frac{2\pi}{3}n}$$

Let s be a DNA strand, denote $b[s] = U_b(\frac{N}{3})$. We calculate the values of $arg(A[s])$, $arg(C[s])$, $arg(G[s])$, and $arg(T[s])$ in coding and noncoding regions, where $arg(b[s])$ denotes the argument of $b[s]$. Kotlar and Lavner's analysis of all the experimental genes of *S. cerevisiae* revealed that the distributions of the arguments in all four nucleotides for coding regions were in bell-like curves around a central value, while the corresponding histograms for noncoding regions seemed to be close to uniform (*5*).

Kotlar and Lavner introduced the Spectral Rotation (SR) Measure. Let μ_b be the sample mean of $arg(b[s])$ (b = A, C, G, or T) in coding regions. It is expected that $arg(b[s]) \approx \mu_b$ for a typical coding sequence s. Rotating the vectors $A[s]$, $C[s]$, $G[s]$, and $T[s]$ clockwise by the corresponding argument μ_A, μ_C, μ_G, and μ_T (multiplication by $e^{-i\mu_b}$) respectively will yield four vectors pointing roughly in the same direction. Hence, the vector sum $\sum_b e^{-i\mu_b} b[s]$ will be of large magnitude compared to the case where the vectors point in different directions, as is most likely the case for a noncoding sequence. Considering the shape of the argument distributions, more weight should be given to narrower distributions, so each term can be divided in equation of $\sum_b e^{-i\mu_b} b[s]$ by the corresponding angular deviation, and the SR measure is developed:

$$|V|^2 = \left| \frac{e^{-i\mu_A}}{\sigma_A} A[s] + \frac{e^{-i\mu_C}}{\sigma_C} C[s] + \frac{e^{-i\mu_G}}{\sigma_G} G[s] + \frac{e^{-i\mu_T}}{\sigma_T} T[s] \right|^2 \quad (3)$$

ASR measure

We drew the histograms of $arg(A[s])$, $arg(C[s])$, $arg(G[s])$ and $arg(T[s])$ values in coding and noncoding regions in rice DNA (Figure 1). To get a reliable result, we used the trainning set, from which all exons and introns were extracted and joined as coding and noncoding sequence in each gene.

As Figure 1 shows, for coding regions, the distributions of arguments for A and G are bell-like curves, whereas the histograms of $arg(C[s])$ and $arg(T[s])$ values seem to have two central values, just like two distributions are joined together. For noncoding regions, the distributions seem to be close to uniform. The distributions for coding regions and noncoding regions are very different, which is accordant with the statement of Kotlar and Lavner (*5*). However, as the figure reveals, not all the distributions of the arguments in all four nucleotides taper around a central value as Kotlar and Lavner claimed. Why the histograms of arguments C and T are two-center shapes is a question to be answered, but it is beyond the scope of this paper. In this case, we could also use the SR measure assuming there be only one center value for all four nucleotides. Calculate the sample mean of $arg(b[s])$ (b = A, C, G, or T), and rotate the vectors $b[s]$ clockwise (multiplication by $e^{-i\mu_b}$) respectively. However, a not perfect result would be obtained.

We did the non-parametric fitting for the histograms of arguments C and T (Figure 2). Take $arg(C)$ for example, as the figure shows, we could assume there are two peaks in the histogram. Looking for the lowest value between the two peaks as a cutoff value (-2.689), the arguments for nucleotide C could be separated into two subsets. For each part, a sample mean and a deviation (μ_1, σ_1 in the subset whose value is less than the cutoff value, and μ_2, σ_2 in the other subset) are calculated. Therefore, in the procedure of identifying whether a DNA strand s is coding regions or not, before the vector $C[s]$ is rotated, the parameters μ_C, σ_C could be selected as (μ_1, σ_1) or (μ_2, σ_2) according to whether or not $arg(C[s])$ is less than the cutoff value. The same will be done for the $T[s]$, so an Alternative Spectral Rotation measure is presented.

Result

Table 1 compares the performance of the ASR measure with the SR and SC measures. All measures were tested on coding and noncoding regions from the test data set, and results were obtained by sliding windows of sizes 90, 120, 180, 240, 300, and 351 bp. In order to compare with the SR measure, we also chose the threshold that insured the FP is 10% as Kotlar and Lavner did. As Table 1 shows, the ASR measure performs better than other measures in all window sizes.

Though the ASR measure has made improvements in identification in rice DNA, the accuracy is still far away from being perfect, especially in short fragments. It is somewhat different from the result of Kotlar and Lavner. Maybe it is because of the dis-

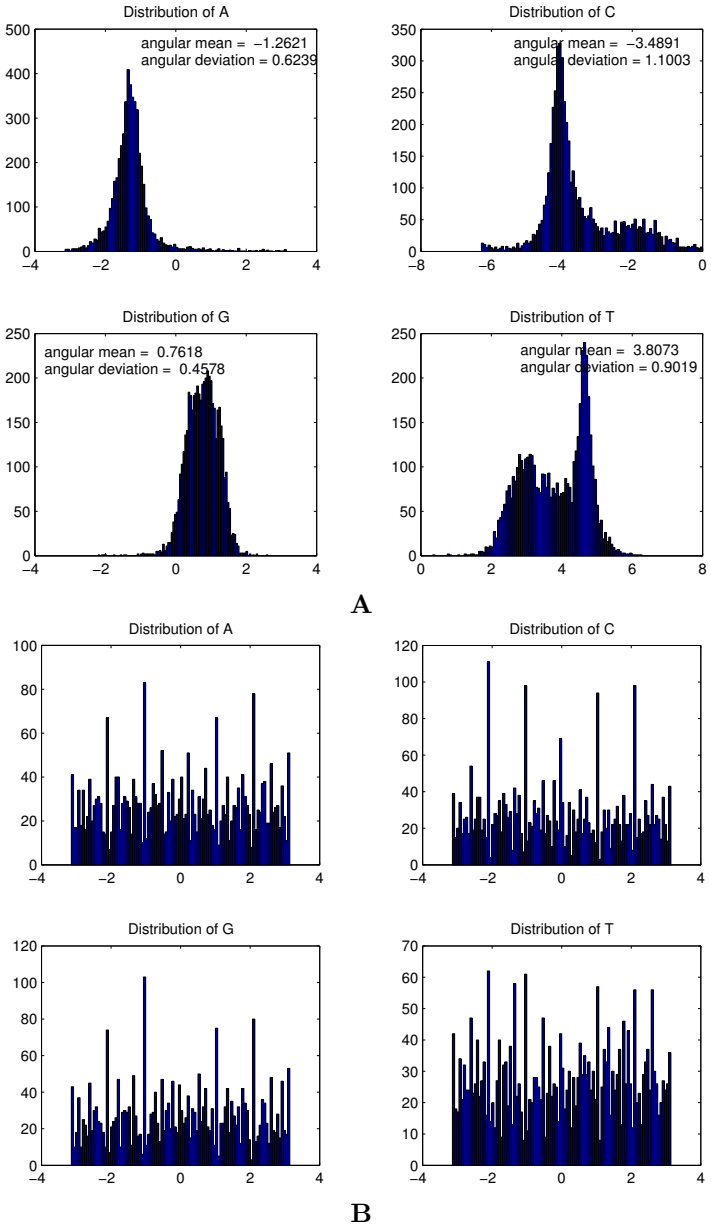

Fig. 1 Argument distributions of A, C, G, T for coding and noncoding regions. **A**. Histograms of $arg(A[s])$, $arg(C[s])$, $arg(G[s])$, and $arg(T[s])$ values for 5,047 coding sequences. **B**. Histograms of $arg(A[s])$, $arg(C[s])$, $arg(G[s])$, and $arg(T[s])$ values for 5,047 noncoding sequences. A 2π shift was applied to part of the data when necessary.

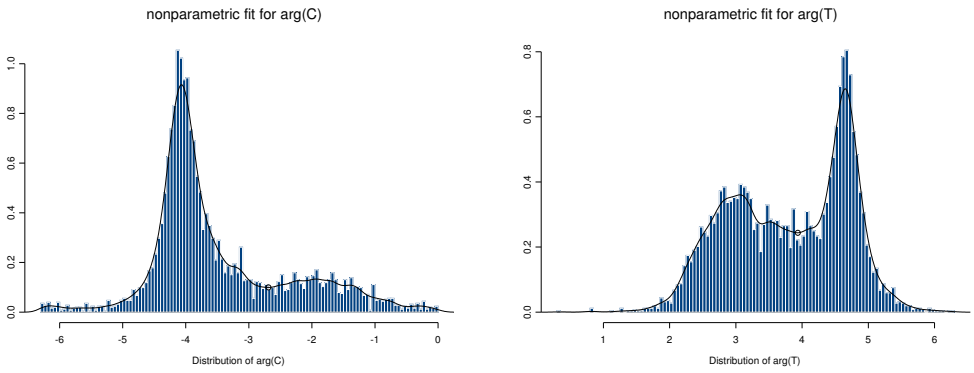

Fig. 2 Nonparametric fit for the histograms of arguments C and T.

Table 1 Performance of Fourier Spectrum Measures Using Different Window Sizes

Measure	Percentage of exons detected for 10% false positive (%)					
	90 bp	120 bp	180 bp	240 bp	300 bp	351 bp
SC	50.33	59.78	73.83	86.65	91.04	93.60
SR	48.88	59.17	71.86	82.66	88.78	92.19
ASR	61.07	71.56	83.86	90.50	94.07	96.01

tinctness of different species. One method also based on DFT was used by Wang *et al* (*16*). Its accuracy of identifying coding regions is apt to show that the methods based on DFT do not have as high performance as Kotlar and Lavner's description.

Linear Discriminant Analysis

Recognition Variables

In order to increase the identification accuracy in rice coding regions, we chose three different variables as discriminant parameters besides the ASR variable, and performed the Linear Discriminant Analysis.

The asymmetric variable

We calculated the distribution of A, C, G, T bases at three codon positions on the test set (Table 2). As Table 2 reveals, the contents of T, G, and A are poor at the first, second and third codon positions, whereas for the noncoding sequences, the contents of A, C, G, and T are nearly a constant no matter which position the nucleotide locates. Considering all the three alternative phases in coding sequences, we assumed that the first inframe codon started at position i (i = 1, 2, or 3) in the sequence, and let $y_1(i)$, $y_2(i)$, $y_3(i)$ represent the contents of T, G, and A at the first, second, and third codon positions, respectively. We denoted R_i as $R_i = \prod_{j=1}^{3} y_j(i)$ (i = 1, 2, or 3) and defined the asymmetric variable as $X1 = \min_i(R_i)$.

Table 2 Contents of A, C, G, T bases at Three Codon Positions

Codon position	A	C	G	T
1st	0.2611	0.2130	0.3559	0.1700
2nd	0.2982	0.2420	0.1862	0.2737
3rd	0.1472	0.3388	0.3071	0.2069

The purine variable

As we know, the predominant bases at the first codon position are purines (nucleotides A and G) and this rule is independent of species. Table 2 could also prove this fact. We defined P_i (i = 1, 2, or 3) as the occurrence frequency of purines in the three phases. The purine variable was defined as $X2 = \max_i(P_i)$.

The stop-codon variable

The stop codon is one of the triplets TAA, TAG, and TGA. As Wang *et al* described, the distribution of the triplets in coding regions is apparently different from those in noncoding regions (*16*). The total number of the triplets contained in all three frames in a sequence was denoted by n. The number of the frames containing the three triplets in a sequence was denoted by K (K = 0, 1, 2, or 3). The stop-codon variable was defined as $X3 = (1 + K^2)n$.

Result

The LDA algorithm was applied by using the three variables mentioned above with the ASR variable. To evaluate the accuracy of prediction, sixfold cross-validation tests were adopted. We selected 1,600 coding and 1,600 noncoding sequences with length of 351 bp randomly from the test set. From these fragments we obtained the data sets by sliding windows of sizes 90, 120, 180, 240, and 300 bp, with the corresponding numbers of the coding and noncoding sequences as 4800, 3200, 1600, 1600, and 1600, respectively. Take the data set with window size 351 bp for example, the database was randomly divided into two parts for three times (400+1200, 800+800, and 1200+400). For each time, Part 1 was taken as a training set and Part 2 as a test set at first, then the procedure was applied by reversing the roles of the two parts. The sensitivity, specificity and accuracy of the algorithm were based on the test set according to the discriminant rules trained from the sequences with different window lengths 90, 120, 180, 240, 300, and 351 bp, respectively (Table 3). We also calculated the prediction results using only one variable each time (Table 4). The procedure was quite like the case of four variables.

The relation between the prediction accuracy of the algorithm and sequence length is shown in Figure 3. As it reveals, we could see that the prediction accuracy of the ASR variable is better than that of the asymmetric and purine variables, while the stop-codon variable performs the best among the four. However, we could see that when sequence length decreases, the accuracy of the stop-codon variable reduces drastically (this phenomenon was also narrated by Wang *et al*; ref. *16*), while the accuracy of ASR reduces relatively slower. Though ASR does not perform better than the stop-codon variable, compared with the asymmetric and purine variables, it is relatively

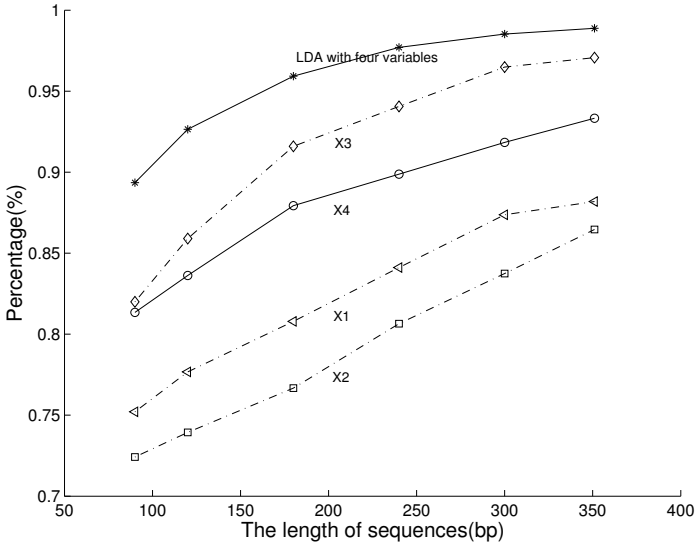

Fig. 3 The relation between the prediction accuracy of the algorithm and sequence length. X1: the asymmetric value; X2: the purine value; X3: the stop-codon value; X4: the ASR value.

Table 3 The Average Prediction Results Using Four Variables

Performance	90 bp	120 bp	180 bp	240 bp	300 bp	351 bp
Sensitivity (training)	90.73	94.54	97.79	98.69	99.35	99.65
Specificity (training)	88.04	90.28	94.35	96.64	97.85	97.97
Accuracy (training)	89.38	92.68	96.07	97.67	98.60	98.81
Sensitivity (test)	90.68	94.49	97.55	98.76	99.32	99.60
Specificity (test)	88.03	90.81	94.31	96.64	97.74	98.15
Accuracy (test)	89.35	92.65	95.93	97.70	98.53	98.88

Table 4 The Average Prediction Accuracy Using One Individual Variable

Variable	90 bp	120 bp	180 bp	240 bp	300 bp	351 bp
asymmetric	75.21	77.67	80.79	84.11	87.37	88.19
purine	72.42	73.84	76.67	80.65	83.74	86.46
stop-codon	82.00	85.90	91.60	94.06	96.49	97.07
ASR	81.34	83.62	87.93	89.88	91.84	93.33

better in recognizing coding sequences, especially in shorter fragments. Meanwhile, the prediction accuracy of coding regions using LDA with the four values increases about 8%–9% compared to the accuracy only using the ASR value in all window lengths.

Discussion

We could predict exons in a gene sequence using a sliding window of 351 bp with the ASR measure. Moreover, the plot of $arg(ASR)$ can be a tool for finding the reading frame (*5*). Figure 4 depicts the graphs of the ASR measure and the $arg(ASR)$ value on gene AB037371.

What's more, we could use the discriminant value obtained by LDA with the four variables to detect exons. As Wang *et al* mentioned, the stop-codon value could help to detect the correct reading frame of coding regions (*16*). Now with the help of $arg(ASR)$ and stop-codon values, we could make our decision that on what phase the exon is. It will make the recognition of coding sequences easier. By defining the prediction score for each gene as:

$$score = \frac{E(V_{coding}) - E(V_{noncoding})}{std(V_{coding}) + std(V_{noncoding})}$$

(V_{coding} and $V_{noncoding}$ are LDA discriminant values that are limited to ASR values), we could give a roughly criterion by which the prediction quality of the whole genes could be scored.

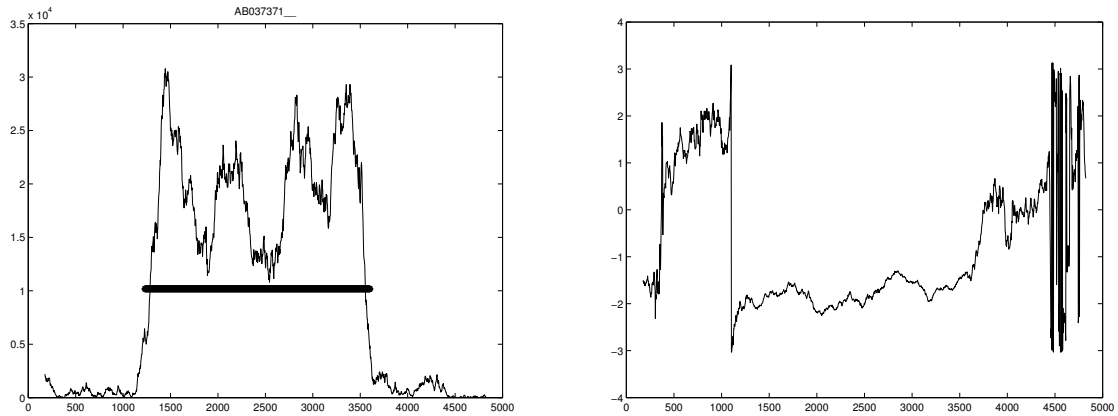

Fig. 4 Graphs of the ASR measure (A) and the $arg(ASR)$ value (B) on the Rice Gene AB037371 using a sliding window of 351 bp.

Acknowledgements

The author is extremely grateful to Dr. Heng Li for his help in organizing the databases used in this paper.

References

1. Staden, R. and McLachlan, A.D. 1982. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res.* 10: 141-156.
2. Farber, R., *et al.* 1992. Determination of eukaryotic protein coding regions using neural networks and information theory. *J. Mol. Biol.* 226: 471-479.
3. Tiwari, S., *et al.* 1997. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput. Appl. Biosci.* 113: 263-270.
4. Anastassiou, D. 2000. Frequency-domain analysis of biomolecular sequences. *Bioinformatics* 16: 1073-1081.
5. Kotlar, D. and Lavner, Y. 2003. Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. *Genome Res.* 13: 1930-1937.
6. Fickett, J.W. and Tung, C.S. 1992. Assessment of protein coding measures. *Nucleic Acids Res.* 20: 6441-6450.
7. Fickett, J.W. 1996. The gene identification problem: an overview for developers. *Comput. Chem.* 20: 103-118.
8. 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.
9. Salzberg, S.L., *et al.* 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 26: 544-548.
10. Salzberg, S.L., *et al.* 1998. A decision tree system for finding genes in DNA. *J. Mol. Biol.* 5: 667-680.
11. Lukashin, A.V. and Borodovsky, M. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* 26: 1107-1115.
12. Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268: 78-94.
13. Salamov, A.A. and Solovyev, V.V. 2000. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* 10: 516-522.
14. Li, W. 1999. Statistical properties of open reading frames in complete genome sequences. *Comput. Chem.* 23: 283-301.
15. Zhang, C.T. and Wang J. 2000. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nucleic Acids Res.* 28: 2804-2814.
16. Wang, Y., *et al.* 2002. Recognizing shorter coding regions of human genes based on the statistics of stop codons. *Biopolymers.* 63: 207-216.
17. Thanaraj, T.A. 2000. Positional characterisation of false positives from computational prediction of human splice sites. *Nucleic Acids Res.* 28: 744-754.
18. Oppenheim, A.V., *et al.* 1999. *Discrete-Time Signal Processing* (2nd edition). Prentice Hall, Upper Saddle River, USA.
19. Li, H., *et al.* Test data sets and evaluation of gene prediction programs on the rice genome. *J. Comput. Sci. Tech.* In press.