Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Free Energy Analysis on the Coding Region of the Individual Genes of *Saccharomyces cerevisiae*

Chuanhua Xing[1]*, Donald L. Bitzer[2], Winser E. Alexander[1], Anne-Marie. Stomp[3], and Mladen A. Vouk[2]

***Abstract*— Two methods, power spectrum density analysis (PSD) and synchronization signal approximation, were investigated to determine if underlying periodic, free energy signals could be detected for the individual genes in this paper. These signals could be revealed assuming Watson-Crick type hybridization between the eight, 3'-terminal nucleotides of the 18S rRNA and pre- and mature-mRNA sequences in Saccharomyces cerevisiae in a manner similar to that used to analyze coding region sequences in procaryotic genes. Using PSD, a periodic signal could only be detected in 35 of 106 genes tested; using the synchronization signal approximation, 91 of 106 genes showed linearly increasing magnitude and phase, characteristics consistent with the presence of an underlying periodic signal with an assumed frequency of one-third. The majority of introns did not show magnitude and phase behavior consistent with an underlying non-periodic signal. The periodicity property for the free energy on the protein-coding regions can contribute to finding the approximate boundaries of the exons (protein-coding regions) and the introns, which provides a foundation for future studies in identifying the exact positions of the splice sites, especially for the higher eukaryotic genes.**

I. Introduction

Periodic signals on coding region using the free energy have been studied on some procaryotic and eukaryotic genes. Rosnick et al. [3], [4] used the E. coli genome and computed the free binding energies between the 16S rRNA of E. coli K-12 and mRNA with an ensemble of around 2000 genes. Rosnick [4] and Mishra [2] detected the coding region by the periodic signals which are present in the coding region of the average signal, and they do not occur in the non protein-coding region. In eukaryotic genes, the question is whether the free binding energy calculation between the 3 end of 18S rRNA and mRNA can still work as a good indicator to detect the periodical signal in the protein-coding region as it works for prokaryotic genes? Xing et al. [8] investigated the free-energy ensemble behavior for 135 genes for Saccharomyces cerevisiae. The results showed the dominant one-third frequency component for the PSD of the free energies for the protein-coding regions, and primary tests on non-protein-coding regions showed no significant signal.

We further investigate the free energy behavior for the individual gene in this paper. A prokaryotic-like interaction between the 3' end of the 18S rRNA and the region of the initiation codon of mRNA was suggested by Hagenbuchle et al., [10] and Sargan et al., [11]. As proposed in Xing et al. [8], we assume that the 3'-terminal 18S rRNA tail sequence, 3'-attactag-5', has interactions with the mRNA sequence during translation initiation and elongation. This corresponds to the 3'-terminal 18S rRNA tail sequence, 3'-attactag-5', which interacts with the 5' untranslated region and the coding region sequentially. The periodical signal was found in coding region (exon after start codon); the free energy calculation on pre-mRNA or intron was built with the aim to compare when the gene has or has no intron(s). Thus, the goal for this paper is to use two methods, power spectrum density analysis (PSD) and synchronization signal approximation which based on free energy, to determine if underlying periodic, free energy signals could be detected for the individual genes, and analyze the differences between exons and introns. Most of current research of the splice sites analysis focus on the local gene sequences around the splice sites. While the periodic property in the protein-coding region can extend the work to the gene-wide sequence by finding the approximate region of the splice sites (the boundary of the exons and introns), then locating the splice sites gene-wide, even genomic-wide, becomes possible.

II. Method and Material

We used two approaches to determine the presence of a periodic signal in the coding region of genes: a) power spectrum density analysis and b) cumulative synchronization signal analysis. Both approaches start with the generation of a free-energy vector. This vector is created by utilizing an algorithm that aligns the 8, 3'-terminal nucleotides of the 18S rRNA tail of yeast ribosomal RNA with mRNA sequences. Once this vector is created, it is analyzed using PSD or cumulative synchronization signal as described below.

Free Energy Calculation: The free energy was calculated as given in [1], [4], [8], [9], and utilized by us in previous investigations [9]. We calculated the free binding energy between the 3' tail end sequence of 18S rRNA, 3'-attactag-5', and its underlying sequence, and observed the binding energy at the coding region to investigate the periodic signal and its function to distinguish exons from introns. The 3' tail end sequence of 18S rRNA, 3'-attactag-5', was moved along the gene one nucleotide at a time from the start codon to the end generating a series of alignments. The optimal secondary structure and its corresponding optimal binding energy are determined for each alignment using a dynamic programming algorithm [6]. Thus, a series of optimal binding

*Corresponding author: cxing@ncsu.edu
[1]Chuanhua Xing and Winser E. Alexander are from the Department of Electrical and Computer Engineering, NC State University, Raleigh, NC 27695, USA
[2]Donald L. Bitzer and Mladen A. Vouk are from the Department of Computer Science, NC State University, Raleigh, NC 27695, USA
[3]Anne-Marie Stomp is from the Department of Forestry and Environmental Resources, NC State University, Raleigh, NC 27695, USA

energies or a free energy vector for all the corresponding alignments was computed for the whole sequence.

Synchronization Signal Approximation: Our method here involves analyzing the free binding energy vector by a sinusoidal signal with a normalized frequency of one-third. We accumulate the value per codon (one codon includes three nucleotides) and compute the amplitude and phase by solving the equations for a sinusoidal signal. We choose three consecutive numbers in the free energy vector to represent three points on a sinusoidal signal in one period, then move over three points next. The three accumulated scores over the first k codons are from Xing et al., 2004 [8] and given as the following.

$$A_k = \sum_{k=1}^{\lceil n/3 \rceil - 1} e_{3k-2};$$

$$B_k = \sum_{k=1}^{\lceil n/3 \rceil - 1} e_{3k-1};\ C_k = \sum_{k=1}^{\lceil n/3 \rceil - 1} e_{3k}.$$

The non-varying DC term is removed from A_k, B_k and C_k as follows

$$DC = \frac{A_k + B_k + C_k}{3};$$

$$a_k = A_k - DC;\ b_k = B_k - DC;\ c_k = C_k - DC.$$

Then the amplitude and phase can be obtained by solving two of the following three equations.

$$a_k = M_k sin(\phi_k);\ b_k = M_k sin(\phi_k + \frac{2\pi}{3});$$

$$c_k = M_k sin(\phi_k + \frac{4\pi}{3}).$$

The amplitude M_k and phase ϕ_k are the cumulative representations of the free binding energy starting from the start codon 1 to the current codon k, where k can be any number from 1 to the last codon of the gene. The cumulative calculation compensates for the noise due to the computations for a single codon.

Material: One hundred and six verified genes of Saccharomyces cerevisiae [1] were chosen from GenBank [2]. Each gene has a two-exon, one GU-AG intron structure, with annotation derived from experimental data on the plus strand.

III. PSD Analysis

Fig. 1 gives the PSD of the free energy vector for an individual mature mRNA. In this particular example, a dominant one-third frequency component can be observed. However, not every individual mRNA has a dominant one-third frequency component in the PSD of the free energy vector. Fig 2 shows the PSD of the free energy vector for another example individual mature mRNA. We can see that this PSD does not have a dominant one-third frequency component. In contrast with the results for exons, the PSD of the free energy vector for an intron has no essential difference with Fig 2 that there is no apparent dominant one-third frequency component.

[1]http://www4.ncsu.edu/∼cxing/YeastGenesList.txt

[2]http://www.ncbi.nih.gov

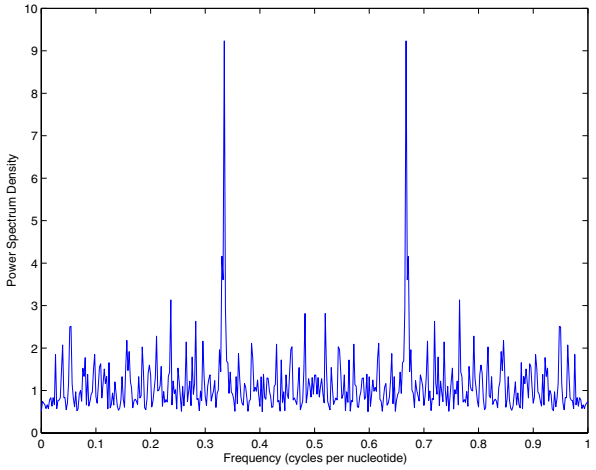

Fig. 1. Power Spectrum Density for an Individual mRNA with the Dominant one-third Frequency Component

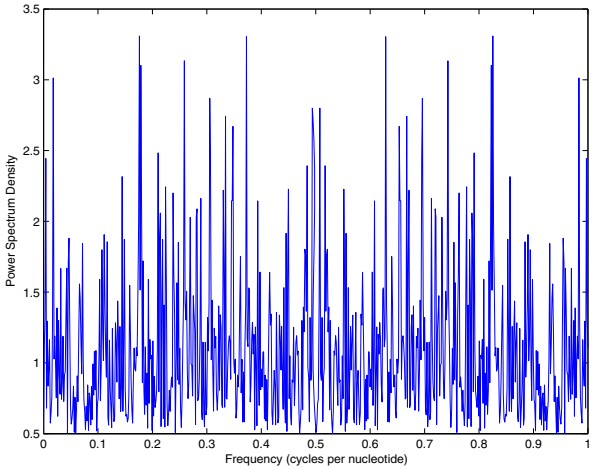

Fig. 2. Power Spectrum Density for an Individual mRNA with the Weak one-third Frequency Component

Repeating the same procedure on 106 mRNAs showed that 35 out 106 mRNAs had the dominant one-third frequency component. We then analyzed the SNRs on 106 mRNAs by using the following equations.

$$SNR = 10 * \log_{10}(Aver\_PSD_s / Aver\_PSD_n),$$

where,

$$Aver\_PSD_s = \frac{1}{Win} \sum_{i=1}^{Win} PSD_s(i),$$

$$Aver\_PSD_n = \frac{1}{n - Win}[\sum_{i=1}^{n} PSD(i) - \sum_{i=1}^{Win} PSD_s(i)].$$

PSD_s represents the PSD for signal; PSD_n represents the PSD for noise; *Win* is the window size around one-third frequency component; *n* is the length of the sequence. Usually *Win* is set 3 if *n* is even number of 3, and set 4 if not.

The analysis of SNR (in dB) distribution with the length of mRNAs is then given in Figure 3. Observation from Figure 3 shows that Saccharomyces cerevisiae needs a gene length

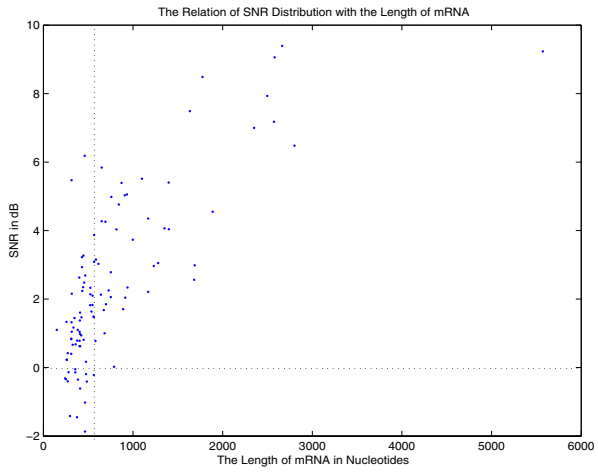

Fig. 3. The relation of SNR distribution with the length of exons

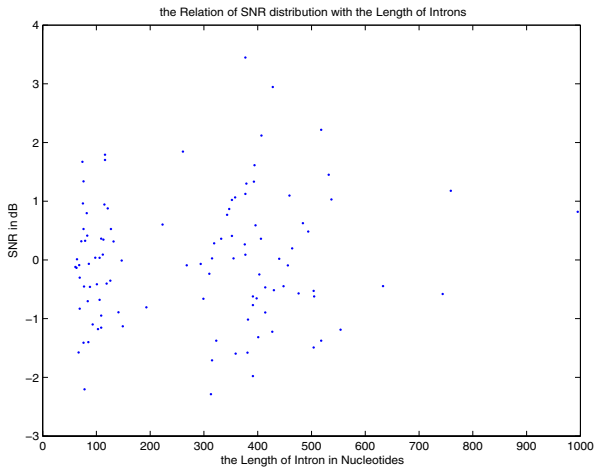

Fig. 4. The relation of SNR distribution with the length of introns

around 500 nucleotides so that a positive SNR greater than 0 dB can be stably obtained, while a gene with the length 788 nucleotides still has a SNR close 0 dB. The length of mRNA need increases to 900 nucleotide so that a SNR greater than 2 dB cab be stably obtained.

Meanwhile, none of 106 introns has shown the dominant one-third frequency component from its PSD test. The length of the introns varied from 60 to more than 1000 nucleotides. The relation of SNR distribution with the length of introns is analyzed in Figure 4. We observed that the SNR did not improve above 3.45 dB as the length of intron increased beyond 1000 nucleotides.

Contrasting the results for exons (mRNAs) with the ones for introns, we can see that the SNRs for mRNAs tend to be larger. The range for mRNAs is [-1.8679 9.3931] with mean 2.4357 dB, and the range for introns is [-2.2873 3.4472] with mean -0.0046 dB. 50% mRNAs, contrasting with 98% introns, have the SNRs $<$ 2 dB, while 14.15% mRNAs, contrasting with 51.89% introns, have the SNRs $<$ 0 dB. However, the mRNAs need to be long enough to stably obtain the SNR. This should be at least 500 nucleotides to obtain a SNR greater than 0 dB, and at least 900 nucleotides to obtain a SNR greater than 2 dB. Therefore, two main conclusions can be draw from PSD analysis. (1). The dominant one-third frequency component can be observed at 35 out 106 mRNAs, while none of 106 introns showed the dominant one-third frequency component by PSD analysis. (2). Comparing the SNR analysis for mRNAs (combined exons) and introns, we know that SNR for mRNAs is much more stronger than the one for introns. However, at least 500 nucleotides is needed to obtain a SNR $>$ 0 dB, and at least 900 nucleotides is needed to obtain a SNR $>$ 2 dB. The longer length of 900 nucleotides is required to observe the dominant one-third frequency component using PSD analysis. Therefore, it is desirable to find a better approach to explore the use of the periodicity to distinguish exons and introns.

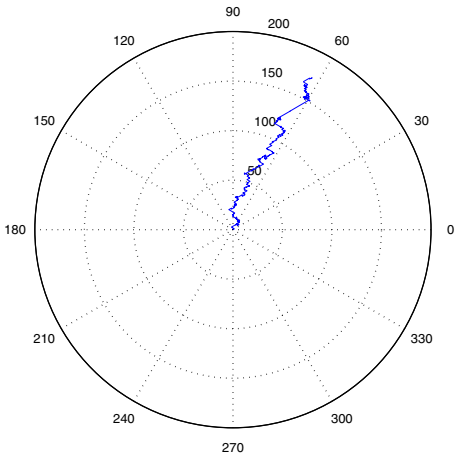

Fig. 5. Polar Plot for an Individual mRNA

IV. Synchronization Signal Analysis

The potential for very noisy signals limits the ability of the PSD analysis to resolve a periodic signal in the coding regions of genes. What is needed is a method that will enhance the signal while damping the noise. The approach utilizes a cumulative calculation to cancel out the noise while strengthen the signal. Therefore, we used the synchronization signal approximation method, described in Section II, to analyze the genes further to distinguish the exons from the introns. According to this method, if a gene has a 3-nucleotide periodical signal, then a linearly increasing amplitude M_k and a constant phase ϕ_k should be observed.

An example polar plot of the amplitude and phase for a mRNA is given in Fig 5. A linearly increasing amplitude M_k and a constant phase ϕ_k is present. As the assumption is made in the synchronization signal method that the dominant frequency for the signal is one-third, this result is consistent with the existence of a periodic signal of frequency one-third. We applied this method to 106 mRNA sequences and determined that approximately 91 out of the 106 mRNAs have the linearly increasing amplitude and constant phase

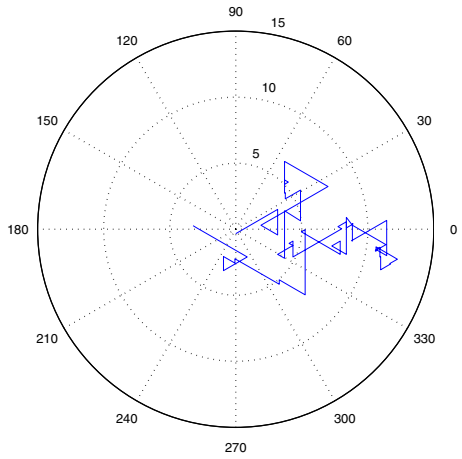

Fig. 6. Polar Plot for an Individual Intron

property. The behavior of the remaining approximately 15 of 106 genes does not show the characteristic of linearly increasing amplitude and constant phase.

We used the method to determine if the one-third frequency component could be detected in intron sequences. A typical polar plot for an example intron is given as Fig 6. In this case, there is no linear increase in amplitude or constant phase, consistent with the interpretation that an underlying periodic signal of frequency one-third is not present in this intron. This is true for approximately 97 introns when we repeat the same test. However, the sample test shows that approximately 19 introns show the trend of the linearly increasing amplitude and the constant phase. One possible explanation is that some introns could have some functional sequences, like micro-RNAs. The free energy investigation of this part is unknown. Other possible explanation is that splicing does not always occur at one site; alternative splice sites exist and, when used, result in sequences normally assigned to the intron, to be included in the exon. Therefore, more study is needed to find out the mechanism of the identifying the boundaries of the exons and the introns .

V. Summary and Contribution

We studied the periodical signal, from the free binding energies between the 3'-attactag-5' and mRNA, for individual genes. Our goal was to determine if the periodic signal, if it exists, could be used to differentiate the approximate regions of intron and exon. The results derived from our experiments are summarized below.

1) From the PSD analysis, the dominant one-third frequency component was found for 35 out 106 mRNAs (the combined exons), but not all. No significant one-third frequency component was found from 106 introns. SNR analysis showed the improvement for exons contrasting with introns. However, SNR analysis of the PSD requires at least 500 nucleotides for SNRs greater than 0 dB and at least 900 nucleotides for SNRs greater than 2 dB. This showed that the PSD approach does not provide sufficient resolving power to reveal periodic signals in short coding sequences
2) The synchronization signal approximation was more successful, because it can compensate for the noise due to the cumulative calculation as the algorithm discussed in section II. With this second approach, approximately 91 of 106 protein coding regions showed linearly increasing magnitude and constant phase, consistent with the presence of an underlying periodic signal of frequency one-third. However approximately 19 of the 106 introns showed the trend of linearly increasing amplitude and constant phase. We did not explore the possible explanation beyond that these sequences could result from alternative splice sites or micro-RNA sequences.

The periodic signal study on the individual genes will contribute to the further analysis of the splice sites. If robust methods can be developed to resolve the periodic signal of coding regions, i.e. exons, the presence or absence of this signal can be used to differentiate the approximate boundaries of exons and introns. Our synchronization signal method shows potential in that regard. Additional and more analysis could then be used to pinpoint the nucleotide boundary within that approximate sequence. The work in this paper will be more significant for the higher eukaryotic genes especially the ones with extra long introns in thousands of nucleotides, and the similarity of the biological interaction processes between species hint the feasibility of applying this method to higher eukaryotic genes, for example, human genes.

References

[1] B. Lewin, Genes VI, *Oxford University Press*, pp.101-104, 1997.
[2] M. Mishra, S. K. Vu, D. L. Bitzer, M. A. Vouk, and A Stomp, Coding sequence detection and free energy periodicity in prokaryotes, *Genomic Signal Processing and Statistics 2004 Conference*.
[3] D. I. Rosnick, D. L. Bitzer, M. A. Vouk and E. E. May, Free energy periodicity in E.coli, *in The First Joint BMES/EMBS Conf.*, vol. 2, pp. 1216, Oct. 1999.
[4] D. I. Rosnick, D. L. Bitzer, M. A. Vouk and E. E. May, Free energy periodicity in E.Coli coding, *in Proceedings of 22nd Annual EMBS Intl. Conf.*, vol. 4, pp. 2470-2473, July. 2000.
[5] A. A. Salamov, et al., Accessing protein coding region integrity in cDNA sequencing projects, *Bioinformatics*, vol. 14, pp 384–390, 1998.
[6] TF Smith, MS Waterman, Identification of common molecular subsequences, *J. Mol. Biol*, vol., 147, pp. 195–197, 1981.
[7] V. V. Solovyev et al., Identification of human gene structure using linear discriminate functions and dynamic programming, *in Proceedings of the 3rd International Conference on Intelligent Systems for Molecular Biology*, pp. 367–375, 1995.
[8] Xing, C., Mishra, M., Vu, S.K., Alexander, W.E., Bitzer, D.L., and Vouk, M.A. Free Energy Based Analysis of the Coding Region of Saccharomyces. *IEEE, the Technology for Life on Biotechnology and Bioinformatics Conference*, October 12–15, 2004, Raleigh.
[9] Xing, C., Bitzer, D., Alexander, W., and Stomp, A. Thermodynamics Exploration to Identify Donor Sites for Yeast. *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics*, May 2006.
[10] Hagenbuchle, O., Santer, M., Steitz, J.A., Mans, R.J. Conservation of the primary structure at the 3' end of 18S rRNA from eucaryotic cells. *Cell*, 13:551-563, 1978.
[11] Sargan, D.R., Gregory, S.P., Butterworth, P.H., A possible novel interaction between the 3'-end of 18S ribosomal RNA and the 59-leader sequence of many eukaryotic messenger RNAs. *FEBS Lett*, 147:133-136, 1982.