Published online 10 December 2008

Nucleic Acids Research, 2009, Vol. 37, No. 2 ***591–601***
doi:10.1093/nar/gkn917

Identification of protein-coding sequences using the hybridization of 18S rRNA and mRNA during translation

Chuanhua Xing[1,*], Donald L. Bitzer[2], Winser E. Alexander[3], Mladen A. Vouk[4] and Anne-Marie Stomp[5]

[1]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7911, [2]Department of Computer Science, North Carolina State University, Raleigh, NC 27695-8206, [3]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7911, [4]Department of Computer Science, North Carolina State University, Raleigh, NC 27695-8206 and [5]Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC 27695-8002, USA

Received September 26, 2008; Revised October 29, 2008; Accepted October 30, 2008

ABSTRACT

We introduce a new approach in this article to distinguish protein-coding sequences from non-coding sequences utilizing a period-3, free energy signal that arises from the interactions of the 3′-terminal nucleotides of the 18S rRNA with mRNA. We extracted the special features of the amplitude and the phase of the period-3 signal in protein-coding regions, which is not found in non-coding regions, and used them to distinguish protein-coding sequences from non-coding sequences. We tested on all the experimental genes from *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. The identification was consistent with the corresponding information from GenBank, and produced better performance compared to existing methods that use a period-3 signal. The primary tests on some fly, mouse and human genes suggests that our method is applicable to higher eukaryotic genes. The tests on pseudogenes indicated that most pseudogenes have no period-3 signal. Some exploration of the 3′-tail of 18S rRNA and pattern analysis of protein-coding sequences supported further our assumption that the 3′-tail of 18S rRNA has a role of synchronization throughout translation elongation process. This, in turn, can be utilized for the identification of protein-coding sequences.

INTRODUCTION

The development of computational methods for the identification of protein-coding sequences is one of the primary research issues in computational biology. Most computational methods for the identification of protein-coding regions are based on various measures that find the differences between coding regions and non-coding regions [as reviewed in (1–3)]. The period-3 signals in coding regions have been used as a measure to identify protein-coding genes [4–9]. Some recent research efforts including those by Tiwari *et al.* (9), Anastassiou (4) and Kotlar *et al.* (7) involved the use of the discrete fourier transform (DFT) to study the period-3 signal in coding regions for the identification of protein-coding genes. Tiwari *et al.* (9) used the magnitude of a period-3 signal to construct a *spectral content* measure for the identification. Anastassiou (4) improved on the former measure by proposing the *optimized spectral content* measure that is based on an optimization technique. Kotlar *et al.* (7) incorporated a phase component to maximize a magnitude optimization for discrimination between coding regions and random DNA sequences (equivalent to non-coding regions). Gao *et al.* (10) used a 'deviation' of the period-3 component from the sequence's fractal 'background' to distinguish protein-coding sequences from non-coding sequences. One suggested explanation for the difference between coding regions and non-coding regions, that may also explain the period-3 signal in coding regions, is that there appears to be a relationship between tRNA abundance and codon bias (one measure for the difference between coding regions and non-coding regions) in the coding regions (11–13). However, most algorithms are based on the statistical analysis of the characters of the DNA sequences without investigating the underlying biological mechanisms.

The investigations of the role of the 3′-end of 16S rRNA in prokaryotes during the translation processes (14–17) have led us to investigate the role of the 3′-end of 18S

*To whom correspondence should be addressed. Tel: +1 919 684 0621; Fax: +1 919 684 0900; Email: cx6@duke.edu

rRNA during the translation process in eukaryotes. The various regions of 18S rRNA have a base-pairing interaction with mRNA during the translation initiation (18–20) and the translation elongation (21–26). Demeshkina *et al.* (23) proposed that the nucleotides of 18S rRNA surrounding mRNA codons at the human ribosomal A, P and E sites are the most strongly conserved regions of the small subunit RNA structure that correspond to nucleotides at four positions of bacterial 16S rRNA. Furthermore, Weiss *et al.* (27) indicated that the 3′-end of 16S rRNA scans the mRNA and is very close to the decoding sites of A, P and E sites during elongation. The interaction of the 3′- end of 16S rRNA–mRNA has been postulated to have a role of synchronization with the correct reading frame during the translation elongation, which has been used for the identification of coding regions [see reviews in (28,29)]. The above observations therefore led to our hypothesis that a prokaryotic-like interaction of the base-pairings between the 3′-end of 18S rRNA and mRNA plays a synchronization role with the correct reading frame during the translation elongation process (28,29).

We introduce a new approach to distinguish protein-coding sequences from non-coding sequences utilizing the interaction of the 3′-terminal nucleotides of the 18S rRNA with mRNA in this article. We discovered a period-3 signal in protein-coding regions by calculating the variable free energy of hybridization of the 3′-terminal nucleotides of the 18S rRNA with the mRNA as it moves through progressive alignments during elongation (28,29). However, the period-3 signal is buried under strong background noise so that sequence identification becomes difficult. Although cumulating over every three nucleotides (28) is able to effectively emphasize the signal while deemphasizing the noise (28,29), the experiments did not explicitly designate how to distinguish protein-coding sequences from non-coding sequences using the features of the period-3 signal. We therefore propose a novel and effective approach to identify protein-coding and non-coding sequences in this article by extracting the features from the phase and the amplitude of the period-3 signals, utilizing the assumption that the interactions of the 3′- end of 18S rRNA and mRNA plays a synchronization role with the correct reading frame during the translation elongation process.

METHODS AND MATERIALS

Previous studies describe proposed measures for gene prediction based on either the phase or the amplitude of the spectral content at the normalized frequency of 1/3 or at other frequencies (4,9,7). We propose to use both the features of the phase and the amplitude of the period-3 signal in the coding regions, which is not in the non-coding regions, to distinguish the protein-coding sequences from the non-coding sequences in this article. We describe our approach in detail in this section. We first construct a free energy sequence by computing the free energy scores between the 3′-end of 18S rRNA and a DNA sequence. We next extract a period-3 signal using a cumulative sinusoidal wave method. We then describe the approach to identify protein-coding and non-coding sequences using the different features of the phase and the amplitude of the period-3 signal. The experimental data are described at the end of this section.

Calculation of free energy sequence

We calculated the free energies for the base-pairings between the 3′- end of 18S rRNA and a DNA sequence. We moved the 3′-end of 18S rRNA, 3′-ATTACTAG-5, downstream (in the 3′-direction) along the DNA sequence one nucleotide at a time from the begining to the end, generating a series of alignments. For each alignment, a free energy score was calculated using a dynamic programming algorithm (30,31). We then calculated a free energy sequence, $E = [e_0, e_1, \ldots, e_L]$, for all the alignments over the whole gene sequence, where L is the length of the sequence.

A number of researchers have used free energy as a metric for studying the interactions of sequences. Starmer *et al.* (32) reviewed and compared with RNAhybrid and RNAcofold (33,34). Our free energy calculation differs from those by Xia *et al.* (35) and Starmer *et al.* (32) in that it considers bulge loops and asymmetric internal loops, but it dose not consider the penalties for terminal AU pairs [Please refer to Michael Zuker and Patrick Stiegler, (36) for the definitions].

Period-3 signal from cumulative sinusoidal wave

Our analysis revealed a period-3 signal from the free energy sequence of a DNA protein-coding sequence, and we used a method, called the cumulative sinusoidal wave, to extract the phase and the amplitude of a period-3 signal (28,29) as is described below. We first summed the free energy modulo 3 over the first 3k nucleotides (k codons). We then fitted a sinusoidal wave to the accumulated free energy values. The cumulative amplitude A_k and cumulative phase θ_k (we call them amplitude and phase for short from now on in this article) were obtained from the cumulative binding energy sequence, and were used to study the period-3 signal in the coding regions, where k can be any number from 1 to the last codon of the gene. The mathematical expressions are listed below.

$$X_k = \sum_{k=1}^{\lfloor L/3 \rfloor} e_{3k-2},$$

$$Y_k = \sum_{k=1}^{\lfloor L/3 \rfloor} e_{3k-1},$$

$$Z_k = \sum_{k=1}^{\lfloor L/3 \rfloor} e_{3k},$$

where L is the length of a sequence in nucleotides. We then subtracted the mean from the values of X_k, Y_k and Z_k. The removal of the mean is based on the mathematical fact that we can always find a sinusoid wave for any three equally spaced points provided the sum of these three points equals to zero. Then we formed a

sinusoidal function using the data with the mean removed. Thus, we have

$$DC = \frac{X_k + Y_k + Z_k}{3},$$
$$x_k = X_k - DC = A_k \sin(\theta_k),$$
$$y_k = Y_k - DC = A_k \sin(\theta_k + \frac{2\pi}{3}),$$
$$z_k = Z_k - DC = A_k \sin(\theta_k + \frac{4\pi}{3}).$$

We can obtain a solution by solving the equations for A_k and θ_k.

$$\theta_k = \arctan \frac{\sqrt{3}x_k}{x_k + 2y_k},$$
$$A_k = \frac{x_k}{\sin(\theta_k)}.$$

The resulting amplitude for the free energy signal from the coding regions tends to be linearly increasing, and the phase tends to be constant. The cumulative calculation compensates for the weak signal to noise ratio (SNR) due to performing the computations for each individual codon (28,29). Therefore, whether a sequence has the features of linearly increasing amplitude and constant phase as discussed provides useful information to determine whether a sequence is a protein-coding sequence or a non-coding sequence. We will discuss this further in the following section.

Sequence identification

We describe the detailed method for sequences identification using the features of the amplitude and the phase of the period-3 signal in this section. We observed the features of the linearly increasing amplitude and the constant phase from ~91 out of 106 protein-coding sequences. However, we did not observe the same patterns in the non-coding sequences (28). We illustrate such features as in Figure 1a;. A protein-coding gene *YRF1-3* has the well-behaved features of the linearly increasing amplitude and the constant phase. In contrast, a long enough non-coding sequence, however, nests around the origin (difficult to distinct by line style). Although we observed such features in previous work, it is still unsolved as to how to identify a protein-coding sequence or a non-coding sequence by evaluating the features of the amplitude and the phase. We next describe the method for extracting the features from the amplitude and the phase in this section, and then use these features to identify protein-coding and non-coding sequences.

(i) Extraction of features for phase θ. (a) Figure 1a shows a polar plot of phase θ versus amplitude A for a protein-coding sequence. We can derive the phase from Figure 1a, and plot it versus position as shown in the top half of Figure 1b. The phase has a large variation in the beginning due to relatively smaller SNR. It converges to almost a constant with very small variations as it moves to higher codon positions due to the noise cancelation from cumulation calculation (See Period-3 signal from cumulative sinusoidal wave section). However, the variations in phase of the non-coding sequences continue to be large even for higher codon positions. The phase versus position plot for a non-coding sequence is also illustrated for comparison in the bottom half of Figure 1b. We observed a similar trend in other protein-coding and non-coding sequences. We therefore defined two variables, terminal phase and phase variation, to define the different features of phase between protein-coding and non-coding sequences. Terminal phase is the phase of the cumulative sinusoidal signal at the end of a free energy sequence. We can use terminal phase to represent the general phase feature (or averaged phase feature) of an entire sequence, because it results from the cumulation of a whole DNA sequence (See Period-3 signal from cumulative sinusoidal wave section). Phase variation measures variation of phase of the cumulative sinusoidal signal at the different positions for a sequence (or the relation of phase variation with the length of sequences).

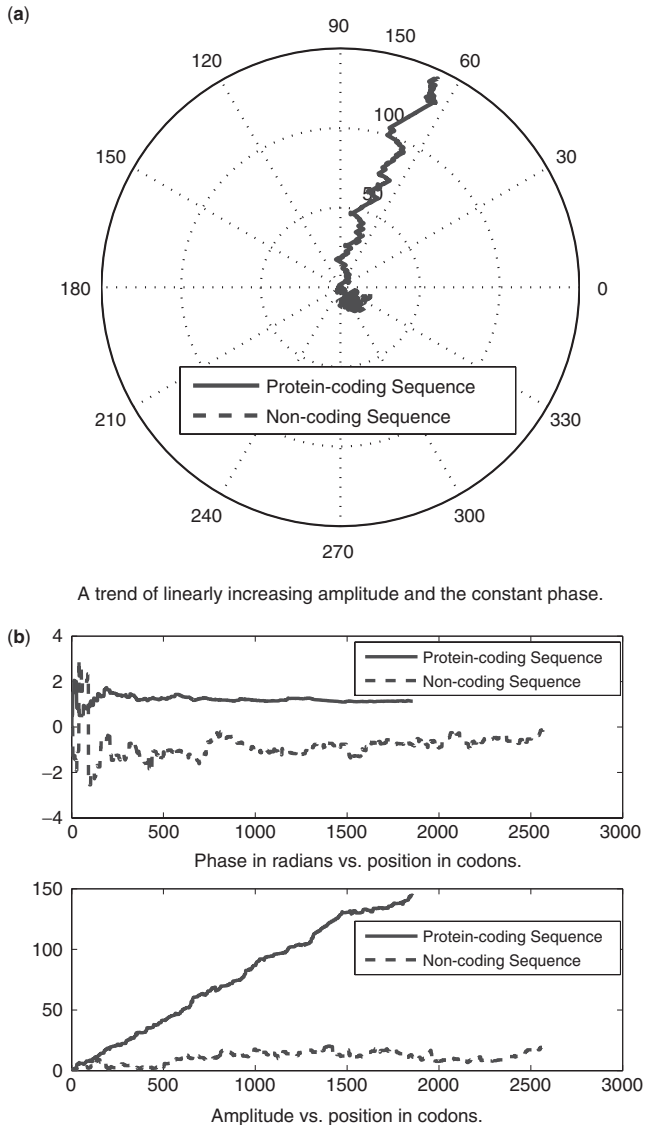

Figure 1. The comparison of the polar plots for protein-coding gene *YRF1-3* and a randomly selected non-coding sequence.

(b) We used the boundaries of 95% and 99.9% confidence intervals (CIs) (37) to study the behavior of the terminal phases for protein-coding sequences. We used the popular statistical measurement 95% CI to study the feature of a centrally distributed dataset. The use of 99.9%, instead of 100%, was to avoid the possible influence of rarely biased data from a given dataset. We considered the sequences with terminal phases within the boundary of 95% as highly possible to be protein-coding sequences, the sequences with terminal phases outside of the boundary of 99.9% as non-protein sequences, and the sequences with terminal phases between these two boundaries as uncertain sequences.

(c) Although the overall trend of phase tends to be constant, there is some phase variations at the different positions for a protein-coding sequence. Phase variations for a protein-coding sequence appear to be larger at lower positions, and tend to be smaller at higher positions. However, the phase for a non-coding sequence appears to be random, and so phase variations appear to be random and large as well. We can observe such trends the top half of Figure 1b. We therefore can define the boundaries of the phase variations for protein-coding sequences for eliminating non-coding sequences. We can analyze phase variations by dividing the range of either phase variations or positions into small slots, and then determine the boundary of the ensemble behavior of the sample data in each slot. We chose the division of phase variations to avoid the possibility of having too small sample size over each slot. We evaluated angle variation using three measures. The distance to the mean D measures the absolute distance of a phase to the mean. The first-order phase difference $\Delta\theta$ measures the absolute value of phase difference between the adjacent positions. The second-order phase difference $\Delta\Delta\theta$ measures the absolute value of $\Delta\theta$ difference, which corresponds to direction changes of adjacent phases. All three measures should converge to zero, as the phase moves to higher positions, which will be consistent with the observation the top half of Figure 1b. For each slot, we can find the boundaries of three measures by finding the positions for their corresponding upper bound one-sided 95% CIs.

(ii) Extraction of features for amplitude A. We can compute the amplitude of the cumulative sinusoidal signal (call it amplitude for short) for a protein-coding sequence, and plot it versus positions as shown in the bottom half of Figure 1b. We can observe that the amplitude is approximately linearly increasing as it moves to higher codon positions. However, this is not the case for a non-coding sequence, as illustrated in the bottom half of Figure 1b. We therefore defined amplitude rate to measure the features of the amplitude for protein-coding sequences, and used the different features of amplitude rate to distinguish protein-coding sequences from non-coding sequences. We captured the position-independent amplitude feature by the amplitude difference as

$$\frac{A(i,k+win)-A(i,k)}{win},$$

where $A(i,k)$ is the amplitude for gene i at codon position k and win is the window size. The amplitude difference is averaged further over the sample size N and the sequence length. The average over sequence length is set for observing the contribution of amplitude rate at the different positions of a sequence. The final expression then is given below.

$$A_{Rate}=\frac{1}{Len}\frac{1}{N}\sum_{k=1}^{Len}\sum_{i=1}^{N}\frac{A(i,k+win)-A(i,k)}{win}, \qquad \mathbf{1}$$

where k is position number, N is the number of sequences in the sample set and $k = 1, 2, \ldots Len$, where Len is the maximum possible length investigated, and $win = 1$, 10 and 27. The different window sizes, win = 1, 10 and 27, were set for observing the relation of amplitude rate with the different window size win. The maximum window size $win = 27$ was set as one of the test window sizes because the minimum length of an intron required for the proper biological processing is around 80 nt ($80/3 \simeq 27$) (38). We expect a higher amplitude rate for protein-coding sequences than non-coding sequences based on their different amplitude features. We can use the usual data analysis principle of maximizing the correct prediction while minimizing the false prediction to decide wether an amplitude rate is for a protein-coding sequence or a non-coding sequence.

(iii) Summary of the features to identify sequences. We can identify sequences by summarizing the features for the phase and the amplitude, the boundaries of 95% CI and 99.9% CI for terminal phase, one-sided 95% CI for three measures of phase variation and the difference of amplitude rates for the protein-coding and non-coding sequences. We can consider a sequence as a non-coding sequence, when either the terminal phase is outside of the wider bound 99.9% CI, all three measures of phase variation are outside of their one-sided 95% CI boundaries, its amplitude rate is too low to be a protein-coding sequence, or all the non-positive conditions for terminal angle, angle variation and amplitude rate hold at the same time. Non-positive conditions for sequences include the cases when terminal phases are between the boundaries of 95% CI and 99.9% CI, some but not all of phase variation measures, D, $\Delta\theta$ and $\Delta\Delta\theta$ are outside of their 95% CI boundaries, and amplitude rate can be either a protein-coding sequence or a non-coding sequence. Otherwise, we consider a sequence as a protein-coding sequence.

Experimental data

The protein-coding and non-coding sequences we used for tests include three parts. (i) We obtained the *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* sequences from GenBank [ftp://ftp.ncbi.nih.gov/genomes/Fungi/Saccharomyces_cerevisiae/ (25 July 2006)]. The protein-coding sequences are open reading frames (ORFs) starting with a start codon 'ATG' and ending with a stop codon either 'TAA', 'TAG' or 'TGA'. The selected sequences are summarized in Table 1. We selected the experimental ORFs as the

Table 1. Datasets for experiments

S. cerevisiae			*S. pombe*		Fly		Mouse			Human		
sc-Cod	*sc-Non*	*sc-Pseu*	*sp-Cod*	*sp-Non*	*fl-Cod*	*fl-Non*	*mo-Cod*	*mo-Non*	*mo-Pseu*	*hu-Cod*	*hu-Non*	*hu-Pseu*
4670	5664	182	591	1997	4000	4000	4000	4000	4401	4000	4000	3576

The first row is specie names, the second row is dataset names and the third row is the number of sequences in each dataset.

protein-coding sequences, and picked the introns and some intergenetic regions as non-coding sequences. We put 4670 experimental ORFs for *S. cerevisiae* into dataset *sc-Cod*, and 5664 non-coding sequences into dataset *sc-Non* which includes 228 introns and 5436 inter-genetic sequences (longer than 50 nt). Similarly, we put 591 experimental protein-coding sequences into dataset *sp-Cod*, and 1997 non-coding sequences into dataset *sp-Non* which includes 1121 introns and 876 inter-genetic sequences (longer than 50 nt). (ii) We randomly selected 4000 protein-coding and 4000 non-coding sequences (longer than 50 nt) from each of three species, fly (*Drosophila melanogaster*), mouse (*Mus musculus*), and human (*Homo sapiens*) from UCSU genome (http://genome.ucsu.edu. Choose 'Genes and Gene Prediction Tracks group', and then choose 'flyBaseGene table' for fly and choose 'xenoRefGene table' for mouse and human). (iii). We obtained 182 pseudo_ORFs from *S. cerevisiae* (http://pseudogenes.org/), 3576 randomly selected pseudo_ORFs from human (http://genome.uiowa.edu/pseudogenes/) and 4401 pseudo_ORFs from mouse (http://pseudogenes.org/) for test. All of these sequences are summarized in Table 1.

RESULTS

We randomly divided 4670 experimental ORFs from dataset *sc-Coding* into two subsets, 2000 ORFs and 2670 ORFs, to serve as the protein-coding training and test sets. Similarly, we randomly selected 2000 and 2670 non-coding sequences from dataset *sc-NonCoding* to form the non-coding training and test sets. Therefore, the training set consisted of 2000 protein-coding sequences (ORFs) and 2000 non-coding sequences. The test set consisted of 2670 protein-coding sequences (ORFs) and 2670 non-coding sequences. Repeating the above procedure another four times, we obtained five training sets and their corresponding five test sets.

Features of phase and amplitude

We used the data from the training sets to observe the different features of the phase and the amplitude between the protein-coding sequences and the non-coding sequences. We then used the extracted features to distinguish the protein-coding sequences from the non-coding sequences, using the approach described in Sequence identification of Methods and material section.

Distribution of terminal phases for sequences. We calculated the phases and then the terminal phases for sequences in the training set using the method in Period-3 signal from cumulative sinusoidal wave section.

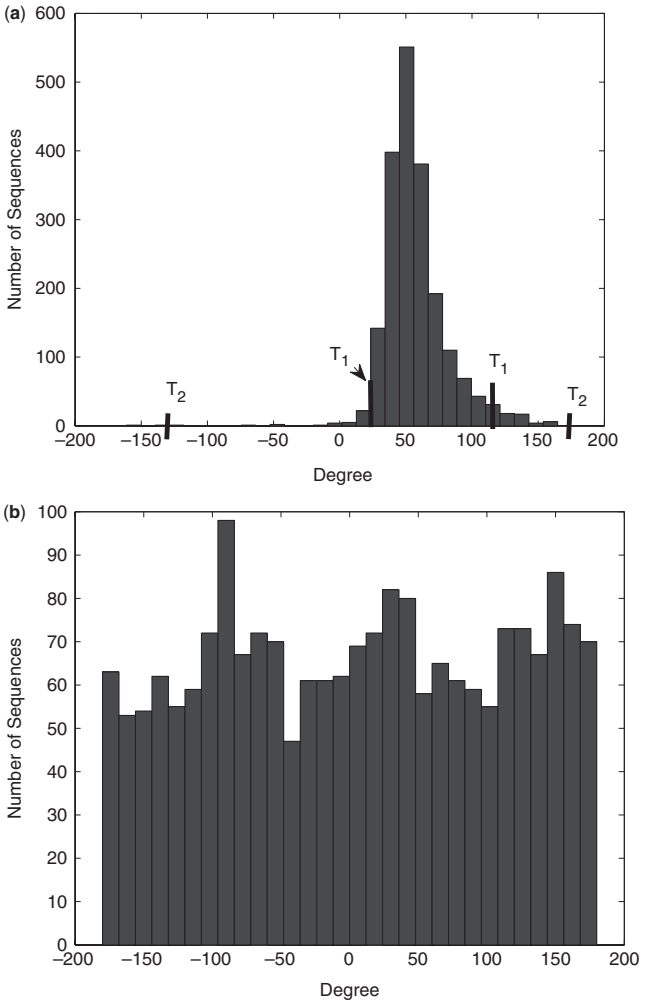

Figure 2. The comparison of the histograms for the terminal phases of the protein-coding and non-protein-coding sequences. (**a**) The histogram of the terminal phases for the protein-coding sequences, where T_1 and T_2 mark the boundaries of 95% CI and 99.9% CI. (**b**) The histogram of the terminal phases for the non-coding sequences.

The histogram of the terminal phases for the protein-coding sequences from the training set is given in Figure 2a. As observed, the terminal phases are distributed around a central value. However, the terminal phases for the non-coding sequences from the training set are randomly distributed over the whole range, as indicated in Figure 2b, where some slightly focused regions in Figure 2b may result from some DNAs with close sequence structure. Similar plots were observed as given in Figure 2 when we repeated the same procedure using

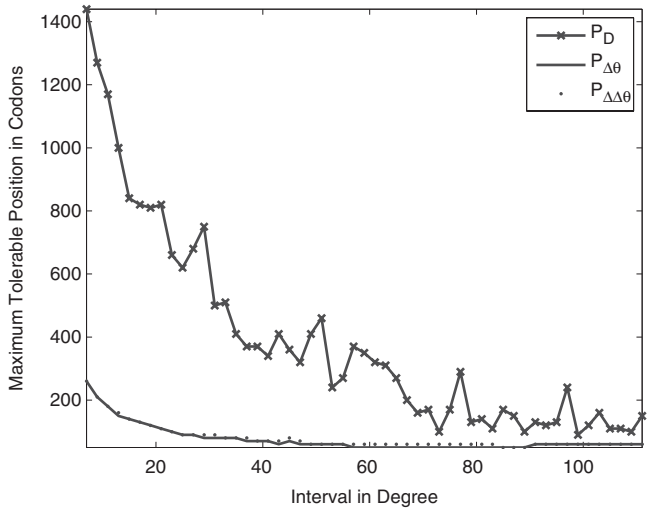

Figure 3. The position boundaries versus the phase variations for three measures of phase variation.

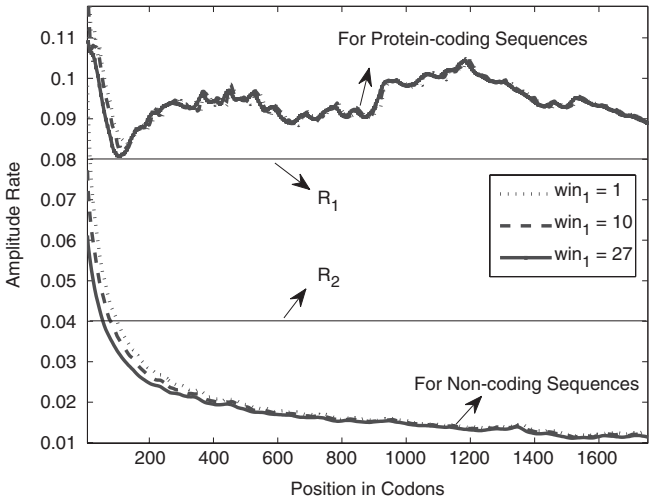

Figure 4. Two groups of amplitude rate plots for the protein-coding and non-coding sequences.

other randomly selected training and test sets. This observation is consistent with the general findings of Kotlar *et al.* (7). The main difference is that the phases for Kotlar *et al.* (7) are for A, T, C and G individually, whereas the phases for our approach are from the combinational effect of A, T, C and G in the free energy signal of the interactions of two sequences.

Relation of phase variation with sequence position. Phase variation is an another variable we created to investigate the relation of phase variation with position for sequence identification. We obtained three position boundaries for D, $\Delta\theta$ and $\Delta\Delta\theta$ versus their phase variations, as given in Figure 3, for the protein-coding sequences from the training set. From Figure 3, the position boundaries decrease when the phase variation increases for all three lines ($P_{\Delta\theta}$, line for $\Delta\theta$, and $P_{\Delta\Delta\theta}$, line for $\Delta\Delta\theta$, are very close). This verifies our anticipation that phase variation for protein-coding sequences depends on sequence position. Phase variation is bigger at lower positions (or short sequences), and becomes smaller when it moves to higher positions (or long sequences). However, the phase variation for non-coding sequences appears to be randomly large at all positions.

Amplitude Rate. Two groups of plots for amplitude rate A_{Rate}, with three different window sizes for each group, are plotted in Figure 4 using Equation (1). We plotted three A_{Rate} for three window sizes *win* versus the position, which are indicated on the top half of Figure 4, using the protein-coding sequences. Similarly, we plotted another three A_{Rate} for three window sizes *win* versus position, which are indicated on the bottom half of Figure 4, using the non-coding sequences. Obviously, there is a distinctive gap between A_{Rate} for the protein-coding sequences and A_{Rate} for the non-coding sequences. The plots between different window sizes, however, do not show too much difference in A_{Rate}. We used the two values R_1 and R_2, the bottom of the amplitude rates for the protein-coding sequences and the average of the amplitude rates for the non-coding sequences, for better maximizing the correct prediction while minimizing the false prediction of sequences identification using the method described in Results section. The values of R_1 and R_2 are 0.08 and 0.04 ,respectively, for the given data in Figure 4. The amplitude rate for our method is equivalent to the amplitude effects in Tiwari *et al.* (9) and Anastassiou (4).

Sequence identification

We observed the differences of features above, from terminal phase, phase variation and amplitude rate, between the protein-coding and non-coding sequences from the training sets. We used [T_1, T_2], as indicated in Figure 2a, to record 95% CI and 99.9% CI of terminal phase, [P_D, $P_{\Delta\theta}$, $P_{\Delta\Delta\theta}$] to record one-sided 95% CIs of three measures, D, $\Delta\theta$ and $\Delta\Delta\theta$, for phase variation, and [R_1, R_2], as indicated in Figure 4, to record the different amplitude rates for the protein-coding and non-coding sequences. We then used these evaluated features to identify 2670 sequences in the test set using the approach described in Sequence identification of Methods and material section. We repeated this procedure another four times. The identification results of five experiments were averaged and observed by the different minimum length of sequences for more explicit and detailed analysis. We selected a series of minimum sequence lengths [17 50 85 171 200 342 440 500 800], where 85, 171, and 342 were selected for comparison with the results by Gao *et al.* (10).

The averaged experimental results for *S. cerevisiae* are listed in Table 2. N_1 and N_2 are the average number of coding and non-coding sequences with minimum length greater than n for *S. cerevisiae* over the five experiments. Sensitivity (Se.) was defined as the proportion of segments in the protein-coding genes set correctly labeled as 'coding'; specificity (Sp.) was defined as the proportion of segments in the non-coding sequences set correctly

Table 2. Accuracy of the synchronization based coding-region identification algorithm on different coding/non-coding subsets

n codons	*S. cerevisiae*								*S. pombe*			
	Our work				Gao *et al.* [10]				for Our work			
	N_1	Se. (%)	N_2	Sp. (%)	N_3	N_4	Se./Sp. (%)	Se./Sp. (%)	N_5	Se. (%)	N_6	Sp. (%)
17	2670	92.38	2670	73.39	–	–	–	–	591	92.89	1997	75.79
50	2646	92.55	2396	78.27	–	–	–	–	587	93.19	1782	83.09
85†	2579	93.15	1902	84.92	4067	4186	85.7	86.7	583	93.57	1601	87.73
171†	2325	94.15	877	94.41	3756	1948	89.89	89.4	510	95.29	1168	92.00
200	2219	94.61	722	95.57	–	–	–	–	483	95.61	1021	92.81
342†	1663	95.53	314	97.89	2674	650	95.4	94.4	349	96.28	554	96.93
440	1263	95.74	215	98.79	–	–	–	–	273	96.34	380	97.89
500	1054	96.14	178	99.1	–	–	–	–	225	97.11	306	99.02
800	430	97.35	83	99.32	–	–	–	–	78	98.72	121	100

†These are the results from Gao *et al.* (10).

labeled as 'non-coding'. Accuracy was defined as the average of the sensitivity and specificity.

The sensitivity and the specificity for *S. cerevisiae* in Table 2 indicate that we obtained very significant results for our experiments. We can observe that the sensitivity for *S. cerevisiae* is rather high, 92.38%, even when the minimum length *n* is 17 codons. As the minimum length *n* increases, the sensitivity is even higher. The specificity tends to be smaller in the beginning as 73.39% when the minimum length *n* is 17 codons, but increases quickly as *n* increases and even larger than the sensitivity when the minimum length *n* is around 171 codons. One suggested explanation for the lower identification rates is that the period-3 signal appears not to be so strong in the beginning of the ORFs for *S. cerevisiae* (29). The significant results suggested the feasibility of utilizing the biological interactions between the 3′-tail of 18S rRNA and mRNA and its resulting synchronization signal to identify protein-coding and non-coding sequences. Alternatively, it supported our hypothesis that the interactions between the 3′-tail of 18S rRNA and mRNA may maintain the right reading frame and produces a synchronization signal throughout the translation elongation process.

Our comparison with the results by Gao *et al.* (10) are illustrated in Table 2. N_3 and N_4 are the numbers of coding and non-coding sequences with minimum length greater than *n* for window size 512 and 218 nt for Gao *et al.* (10) using *S. cerevisiae*. The datasets used by Gao *et al.* include 4125 verified ORFs (fully coding regions or exons) and 5993 non-coding segments (fully non-coding regions or introns) (sequences dated 1 October 2003). We instead used 4670 verified ORFs and 5664 non-coding segments dated 25 July 2006, with 545 more verified ORFs. Gao *et al.* (10) gave two groups of results with the sliding window size $w = 512$ and 128 nt. For comparison, the results for Gao *et al.* (10) are marked by '†' in Table 2. From Table 2, we can observe that both our sensitivity and specificity showed higher identification accuracy than the ones for Gao *et al.* (10) in all three rows (with '†') except the specificity when the minimum length of sequences *n* is 85 codons. A larger portion of short non-coding sequences in our dataset could be one of the reasons for the lower specificity. Even so, our accuracy, the average of the sensitivity and the specificity, is still higher than that of Gao *et al.* (10) when *n* is 85 codons.

Kotlar *et al.* (7) compared the performance of four measures that are based on the Fourier transform at a frequency of 1/3 of the DNA characters for the identification of the protein-coding and non-coding sequences. Two measures from Kotlar *et al.* (7) have the better performance than the other two from Anastassiou (4) and Tiwari *et al.* (9) according to the report from Kotlar *et al.* (7). Kotlar *et al.* (7) used chromosome 16 of *S. cerevisiae* for training, and their performance was tested on the remaining 15 chromosomes. The best result from the work of Kotlar *et al.* (7) is 93.0% with sequence length 351 bp, in which their results evaluation is equivalent to sensitivity in our tests. The sensitivity of our tests, however, becomes higher than 93.0% when *n* is 85 codons and larger, which suggests that our tests give better or competitive performance if the influence of the different dates and the division of datasets are ignorable.

We tested *S. pombe* using the same procedure as *S. cerevisiae* except that we used the training results from *S. cerevisiae*. There are only 591 experimental ORFs for *S. pombe* found in GenBank (http://www.ncbi.nlm.nih.gov), which is not large enough to build a stable training set. We therefore used the training from *S. cerevisiae* to test the sequences in *S. pombe*, assuming that they have the similar features of the synchronization signal in protein-coding regions. We tested the sequences for *S. pombe* five times using the training from *S. cerevisiae*. The test sets of five experiments for *S. pombe* were identical and included the 591 experimental protein-coding sequences (ORFs) and 1997 non-coding sequences. We therefore obtained the five test results for *S. pombe*.

The results for *S. pombe* are shown in Table 2, where N_5 and N_6 are the numbers of the coding and non-coding sequences with length greater than *n*. We can observe a similar trend but slightly better results than *S. cerevisiae* in Table 2. The successful results for *S. pombe* suggest that we can use the genes from *S. cerevisiae* as the training for the tests for *S. pombe* due to the similar features of

the synchronization signals between them. In general, we need to be careful when using the training from one organism to test genes for another organism (7). Since many genes for *S. pombe* include multiple exons, the test results suggest the feasibility of applying our method to the other higher eukaryotic genes with more than two exons.

We primarily tested some higher eukaryotic genes as given in Table 1. The 3′-tail of 18S rRNA sequences for these three species is the same tail as the one for *S. cerevisiae*, '3′-ATTACTAG-5′'. Research indicated that the variations for 18S rRNAs of eukaryotic genes are less, and the 3′-tail of 18S rRNA is highly conserved among species (39). The same identification procedure used for *S. cerevisiae* was applied, and the identification results are given in Table 1 of Supplementary Data. The sensitivity (the percent of protein-coding sequences recognized as protein-coding sequences) for all three species are lower than the ones for *S. cerevisiae* and *S. pombe*, but the specificity (the percent of non-coding sequences recognized as non-coding sequences) for them tend to be higher even when the minimum length of sequences in test dataset *n* is as small as 17 codons. The complex biological mechanisms, including the 3′-tail of 18S rRNA and other possible involved biological mechanisms, cooperate to perform the exact identification of protein-coding regions. Therefore, many features can contribute to the different identification results between yeasts and higher eukaryotic genes. These features can be analyzed from both the rRNA side and protein-coding sequences side, although each of them may contribute differently. The differences of the biological mechanisms other than the 3′-tail of 18S rRNA during translation process may contribute to the differences. Some differences may also be resulted from the different structure of DNA sequences, such as GC contents and the distribution of nucleotides (more discussion is given in next section). Another possible reason includes the different lengths of the sequences. The sequences are longer, a more stable prediction can be conducted due to the cumulation calculation described in Period-3 signal from cumulative sinusoidal wave section. The standards chosen for sequence identification, as described in Sequence identification of Methods and material section, can also contribute to the lower sensitivities and the higher specificities. There are also possible spaces to improve the algorithm for increasing the identification accuracy. Other possible reason includes some possible un-verified genes extracted from UCSC genome (http://genome.ucsc.edu). UCSC genome used some computational skills to select the genes with higher qualities. However, we selected the experimental genes for yeast tests, which are verified and expected to have a higher reliability. While our approach tends to reject any sequence that does not satisfy our evaluations as a protein-coding sequence. However, the accuracy, the average of the sensitivity and the specificity, for these species are still very high, compared with the results in Table 2. More explanation about the period-3, free energy signal and its application to higher eukaryotic genes is needed in the future work.

We also tested our approach on some pseudo_ORFs to observe whether our approach is able to distinct protein-coding sequences from pseudogenes. Pseudogenes are defunct relatives of known genes that have lost their protein-coding ability or are otherwise no longer expressed in the cell. However, they may still have some gene-like features such as promoters, CpG islands and splice sites. What we concern here is whether period-3 signal still exists in pseudo_ORFs. We tested pseudo_ORFs for *S. cerevisiae*, mouse and human as listed in Table 1. Only 49 out of 182 yeast pseudo_ORFs, 1895 out 3576 human pseudo_ORFs and 1562 out of 4401 mouse pseudo_ORFs are considered to have the period-3 signal and therefore recognized as protein-coding sequences. These results make sense in that some of pseudogenes keep some features of protein-coding sequences, although they are not functional to be translated into proteins. For example, when such dis-functionality is caused by stop codon, the period-3 signal before the new stop codon is kept so that a period-3 signal can be detected. Another example is the pseudogenes without proper promoters. Because our approach does not examine whether the promoters before the start codon are intact or not, such pseudo_ORFs may still be recognized as protein-coding sequences. However, most of pseudogenes are recognized as non-protein-coding sequences. Therefore, the period-3 signal is absent in most pseudogenes.

Explanations of period-3 signal

The experiments showed that our approach performs well for the identification of protein-coding sequences. A question is then why the period-3 signal exists and therefore can be used for sequence identification. Research revealed that there appears to be a relationship between the tRNA abundance and codon bias in the coding regions (11–13). However, no essential explanation has been explored for the period-3 signal. In this section, we launch some discussion as to why the revealed period-3 signal from the interactions of the 3′-tail of 18S rRNA and mRNA can work as an indicator of protein-coding sequences?

We tested random sequences, instead of the 3′-tail of 18S rRNA, to observe whether the period-3 signal arises only from the hybridization of the 3′-tail of 18S rRNA and mRNA. We used a genetic algorithm (GA) to search the optimal sequence since there will be totally $4^8 = 65536$ candidate sequences for the exhaustive search. We selected 100 protein-coding sequences with strong period-3 signal for test, based on a previous study that the ensemble signal of a few protein-coding sequences can indicate a strong dominate period-3 signal (29). We calculated the free energy signal between a candidate sequence and a protein-coding sequence, and then obtained SNR from the ensemble free energy signal of 100 free energy sequences [the same procedure can be found from (28)]. The results indicated that an arbitrary random sequence instead of the 3′-tail of 18S rRNA will not produce a period-3 signal. However, it is possible to obtain a dominate period-3 signal if the chosen sequence is closely similar to the 3′ tail of 18S rRNA. The resulting consensus

sequence from the top 30 sequences is '3′-ATTACTAN-5′' with '3′-ACTA-5′' in positions 4, 5, 6 and 7 completely conserved. Slight variations on other positions, especially on positions 1, 2 and 3 suggest that slight variations, such as site mutations, may be tolerable. 'N' on position 8 suggests that the nucleotide on that position may not be important for interacting on mRNAs for synchronization (please see the detailed discussion and description in Supplementary Material). We then can conclude that the 3′-tail of 18S rRNA, instead of a random sequence, may play a role in keeping synchronization throughout the translation elongation process, and therefore contribute to the period-3, free energy signal by interacting on protein-coding sequences.

We also analyzed protein-coding sequences (i) from the difference of GC contents between protein-coding and non-coding sequences and (ii) by randomizing protein-coding sequences completely and partially. GC contents is the percent of G plus C in sequences, and is found to have the distinct values between protein-coding and non-coding sequences. Therefore it has been used for the identification of protein-coding sequences [as reviewed in (3)]. The question is whether the different values of GC content has a relation with the period-3 signal? We tested GC contents in *S. cerevisiae*, and GC contents in true positive, false negative, true negative and false positive data sets are 0.3971, 0.3959, 0.3505 and 0.3532, respectively. We can see that GC contents in protein-coding sequences is higher than in non-coding sequences. However, no obvious difference between the correctly identified gene sequences and falsely identified gene sequences. That is, 0.3971 compares with 0.3959, and 0.3505 compares with 0.3532. We did not find such differences for the tested genes from fly, mouse and human either. Therefore, although GC contents is one of the different features between protein-coding sequences and non-coding sequences, it may not contribute to the period-3 signal pattern directly. Some patterns inside protein-coding regions instead may be important to generate the period-3 signal.

We explored further by randomizing the protein-coding sequences from five species, *S. cerevisiae*, *S. pombe*, fly, mouse and human. We started to permutate the nucleotides at either the first, the second or the third positions within either the first, the second or the third positions, then permutate the nucleotides within any two of the first, the second and the third positions, and finally permutate the nucleotides randomly at all positions. We did the above tests within their own sequences and within all the sequences. We then observed whether the period-3 signal exists for the sequences that are permutated using each of above schemes, using the method described in (29). The results indicated that when we permutate the third positions of codons within the third positions, regardless within their sequences or whole dataset wide, we are able to observe the existence of the period-3 signal. Therefore, the nucleotides at the third positions of codons are important to the period-3 signal. However, the strength of the period-3 signal changes when the different permutations are executed. It seems that the nucleotides re-order themselves to search the optimal organization for obtaining period-3 signal. More tests are needed to understand the rules of nucleotides within protein-coding sequences.

DISCUSSION

We proposed a method of identifying the protein-coding sequences using period-3, free energy signal that arises from the interactions of the 3′-tail of 18S rRNA and mRNAs. We assumed that 18S rRNA has a synchronization role with the right reading frame throughout the translation elongation process. We analyzed the different features of the amplitude and the phase of the free energy signal from the protein-coding and non-coding sequences. We then used these features to distinguish the protein-coding sequences from non-coding sequences in the test sets.

Our method performs well for the identification of the protein-coding genes and non-coding sequences. The sensitivity was >92% when the minimum length starts with 17 codons, and it increases as n increases. The specificity increases faster than the sensitivity and becomes larger than sensitivity when n is greater than 171 codons, although it starts with a lower value. These results have a relatively small standard deviation (SD) with 0.5%. The results supported our assumption that the 18S rRNA may play a synchronization role with the right reading frame throughout the translation process.

Our method shows a better or competitive performance in comparison with several methods that used period-3 signal. Our experiments indicated higher identification accuracy than the results by Gao *et al*. (10). The accuracy of our identification for *S. cerevisiae* increased above 93.15% with the SD <0.5% when n became larger than 85 codons. This is larger than the highest accuracy, 93% with sequence length 351 bp, from the report of Kotlar *et al*. (7). We therefore can conclude that we may have developed an approach that has a better performance for the identification of sequences than several other methods reported in the literature if the influence of different dates and division of datasets are negligible.

All experimental protein-coding genes in *S. cerevisiae* and *S. pombe* were utilized for our experiments. The competitive accuracies for the identification of the protein-coding and non-coding sequences for *S. pombe* with *S. cerevisiae* indicate that there have similar features for the synchronization signal in the protein-coding regions between these two species. The primary tests on some randomly selected genes of fly, mouse and human suggest that our method is applicable to predict genes for higher eukaryotic genes. The test on some pseudogenes further supported our claim that our approach is functional in distinguishing protein-coding sequences from pseudogenes.

Some explanations for the existence of the period-3 signal were explored. The optimal ribosomal sequence search suggests that the 3′-tail of 18S rRNA, instead of a random sequence, plays a role in synchronization during translation elongation process. Some pattern analysis of protein-coding sequences indicates that the nucleotides on

the third positions are important in keeping the right pattern for generating the period-3 signal. More analysis and further exploration of the pattern analysis of protein-coding sequences and its relation with the period-3, free energy signal is needed in the future, especially for higher eukaryotic genes.

SUPPLEMENTARY DATA

Supplementary Data is available at NAR Online.

ACKNOWLEDGEMENTS

We are greatful to two referees for their beneficial comments for improving the article.

FUNDING

Funding for open access charge: Bioinformatics Lab, Computer Science, North Carolina State University.

Conflict of interest statement. None declared.

REFERENCES

1. Claverie,J.-M. (1997) Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet*., **6**, 1735–1744.
2. Fickett,J.W. (1996) The gene identification problem: an overview for developers. *Comput. Chem*., **20**, 103–118.
3. Mathé,C. Schiex,M.-F. and Rouzé,P. (2002) Current methods of gene prediction, their strength and weaknesses. *Nucleic Acids Res*., **30**, 4103–4117.
4. Anastassiou,D. (2000) Frequency-domain analysis of biomolecular sequences. *Bioinformatics*, **16**, 1073–1081.
5. Chechetkin,V.R. and Turygin,A.Y. (1995) Size-dependence of 3-periodicity and long-range correlations in DNA sequences. *Physics Lett. A*., **199**, 75–80.
6. Fickett,J.W. and Tung,C.S. (1992) Assessment of protein coding measures. *Nucleic Acids Res*., **20**, 6441–6450.
7. Kotlar,D. and Lavner,Y. (2003) Gene prediction by spectral rotation measure: a new method for identifying proteincoding regions. *Genome Res*., **13**, 1930–1937.
8. Silverman,B.D. and Linsker,R. (1986) A measure of DNA periodicity. *J. Theor. Biol*., **118**, 295–300.
9. Tiwari,S., Ramachandran,S., Bhattacharya,A., Bhattacharya,S. and Ramaswamy,R. (1997) Prediction of probable genes by fourier analysis of genomic sequences. *Comput. Appl. Biosci*., **13**, 263–270.
10. Gao,J., Qi,Y., Cao,Y. and Tung,W.-W. (2005) Protein coding sequence identification by simultaneously characterizing the periodic and random features of DNA sequences. *J. Biomed. Biotechnol*., **2**, 129–146.
11. Ikemura,T. (1981) Correlation between the abundance of escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J. Mol. Biol*., **151**, 389–409.
12. Karlin,S., Mrazek,J. and Campbell,A.M. (1998) Codon usages in different gene classes of the escherichia coli genome. *Mol. Microbiol*., **29**, 1341–1355.
13. Sharp,P.M. and Li,W.H. (1987) The codon adaptation indexa measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*., **15**, 1281–1295.
14. Lee,K., Holland-Staley,C. and Cunningham,P. (1996) Genetic analysis of shine-dalgarno interaction: selection of alternative functional mRNA-rRNA combinations. *RNA*, **2**, 1270–1285.
15. Ponnala,L., Stomp,A., Bitzer,D. and Vouk,A. (2006) Analysis of free energy signals arising from nucleotide hybridization between rRNA and mRNA sequences during translation in eubacteria, *EURASIP J. Bioinform. Syst. Biol*., 1–9.
16. Rosnick,D.I. (2001) Free energy periodicity and memory model for E.coli codings. *Dissertation*. NC State University, Raleigh, USA.
17. Schurr,T., Nadir,E. and Margalit,H. (1993) Identification and characterization of Eshcherichia coli ribosomal binding sites by free energy computation. *Nucleic Acids Res*., **21**, 4019–4023.
18. Hagenbuchle,O., Santer,M., Steitz,J.A. and Mans,R.J. (1978) Conservation of the primary structure at the 3′ end of 18S rRNA from eucaryotic cells. *Cell*, **13**, 551–563.
19. Sargan,D.R., Gregory,S.P. and Butterworth,P.H. (1982) A possible novel interaction between the 3′-end of 18S ribosomal RNA and the 5′-leader sequence of many eukaryotic messenger RNAs. *FEBS Lett*., **147**, 133–136.
20. Chappell,S.A., Dresios,J., Edelman,G.M. and Mauro,V.P. (2006) Ribosomal shunting mediated by a translational enhancer element that base pairs to 18S rRNA. *PNAS*, **103**, 9488–9493.
21. Halic,M., Becker,T., Pool,M.R., Spahn,C.M.T., Grassucci,R.A., Frank,J. and Beckmann,R. (2004) Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature*, **427**, 808814.
22. Kiparisov,S.V., Sergiev,P.V., Bogdanov,A.A. and Dontsova,O.A. (2006) Structural changes in the ribosome during the elongation cycle. *Mol. Biol*., **40**, 675–687.
23. Demeshkina,N., Repkova,M., Ven'yaminova,A., Graifer,D., and Karpova,G. (2000) Nucleotides of 18S rRNA surrounding mRNA codons at the human ribosomal A, P, and E sites: a crosslinking study with mRNA analogs carrying an aryl azide group at either the uracil or the guanine residue. *RNA*, **6**, 1727–1736.
24. Matveeva,O.V. and Shabalina,S.A. (1993) Intermolecular mRNA–rRNA hybridization and the distribution of potential interaction regions in murine 18S rRNA. *Nucleic Acids Res*., **21**, 1007–1011.
25. Shenvi,C.L., Dong,K.C., Friedman,E.M., Hanson,J.A. and Cate,J.H. D. (2005) Accessibility of 18S rRNA in human 40S subunits and 80S ribosomes at physiological magnesium ion concentrations–implications for the study of ribosome dynamics. *RNA*, **11**, 1898–1908.
26. Tranque,P., Hu,M.C., Edelman,G.M. and Mauro,V.P. (1998) rRNA complementarity within mRNAs: a possible basis for mRNA-ribosome interactions and translational control. *Proc. Natl Acad. Sci. USA*., **95**, 12238–12243.
27. Weiss,R.B., Dunn,D.M., Dahlberg,A.E., Atkins,J.F. and Gesteland,R.F. (1988) Reading frame switch caused by base-pair formation between the 3′ end of 16S rRNA and the mRNA during elongation of protein synthesis in Escherichia coli. *EMBO J*., **7**, 1503–1507.
28. Xing,C., Bitzer,D., Alexander,W., Stomp,A.-M. and Vouk,M. (2006) Free energy analysis on the coding region of the individual gene of Saccharomyces cerevisiae. In 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, New York, NY, USA, Aug. 30–Sept. 3, 2006, pp. 4225–4228.
29. Xing,C., Mishra,M., Vu,S.K., Alexander,W., Bitzer,D. and Vouk,M. (2004) Free energy based analysis of the coding region of Saccharomyces cerevisiae. NC Symposium on Biotechnology & Bioinformatics, Research Triangle Park, North Carolina, USA, October 12–15, 2004, pp. 25–27.
30. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.
31. Thanaraj,T.A. and Pandit,M.W. (1989) An additional ribosome-binding site on mRNA of highly expressed genes and a bifunctional site on the colicin fragment of 16S rRNA from Escherichia coli: important determinants of the of the efficiency of translation-initiation. *Nucleic Acids Res*., **17**, 2973–2985.
32. Starmer,J., Stomp,A., Vouk,M. and Bitzer,D. (2006) Predicting shine-dalgarno sequence locations exposes genome annotation errors. *PLoS Comput. Biol*., **2**, e57.
33. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem*., **125**, 167–188.
34. Bernhart,S.H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2006) Partition Function and Base Pairing Probabilities of RNA Heterodimers, *Algorithms Mol. Biol*., **1**, 3.

35. Xia,T., SantaLucia,J.Jr, Burkard,M.E., Kierzek,R., Schroeder,S.J., Jiao,X., Cox,C. and Turner,D.H. (1998) Thermodynamic parameters for an expanded nearest neighbor model for formation of RNA duplexes with watson-crick base pairs. *Biochemistry*, **37**, 14719–14735.
36. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of lare RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*., **9**, 133–148.
37. Casella,G. and Berger,C. (2002) *Statistical inference*, (chapter 9), 2nd edn. Pacific Grove, CA, Duxbury.
38. Green,M.R. (1986) Pre-mRNA Splicing. *Ann. Rev. Genet*., **20**, 671–708.
39. McCallum,F.S., and Maden,B.E. (1985) Human 18 S ribosomal RNA sequence inferred from DNA sequence. Variations in 18 S sequences and secondary modification patterns between vertebrates. *Biochem. J*., **232**, 725–733.