

Deriving a novel codon index by combining period-3 and fractal features of DNA sequences

Yan Qi¹, Jianbo Gao², Yinhe Cao^{2,3}, and Wen-wen Tung⁴

¹Department of Biomedical Engineering,

Johns Hopkins University, Baltimore, MD 21205, USA

²Department of Electrical and Computer Engineering,
University of Florida, Gainesville, FL 32611, USA

³BioSieve, 1026 Springfield Drive, Campbell, CA 95008, USA

⁴ National Center for Atmospheric Research

P.O. BOX 3000, Boulder, CO 80307-3000, USA

Email: {yanzju@hotmail.com, gao@ece.ufl.edu,
contact@biosieve.com, wwtung@ucar.edu }

Abstract When a gene finding algorithm incorporates multiple useful sources of information about coding regions, it becomes more successful. It is thus highly desirable to find new and efficient codon indices. Here we propose a novel codon index called the period-3 fractal deviation (*PF**D*). This is obtained by incorporating two incompatible features of DNA sequences, the period-3 feature in coding regions and the fractal feature in both coding and non-coding regions. The former is due to the fact that in coding regions, three nucleotide bases encode an amino acid and that the usage of bases at the three reading frames is highly biased. The fractal feature comes from the fact that the background of a DNA sequence is fairly random. These two features are incompatible because period-3 defines a specific scale of three nucleotide bases while fractal means there are not any specific scales. The *PF**D* is very different for coding and non-coding sequences, and is reading-frame-dependent. When the coding segment starts with the gene-containing reading frame, the period-3 feature causes a large deviation from fractal scaling at the scale of 3 and results in a large *PF**D*. When the segment starts with an incorrect reading frame, the periodicity of 3 does not conflict with fractal behavior and the deviation values are small. Hence, when *PF**D* are separated into three different reading frames, the variation of *PF**D* vs. the nucleotide position shows correctly not only where the coding sequence is, but also the reading-frame. See Fig. 1. The accuracy of the *PF**D* is evaluated by studying all of the 16 yeast chromosomes. It is found that the percentage accuracy is very high and quite independent of the sliding window size. See Table. 1. It is further shown that the *PF**D* is complementary to other codon indices such as Fourier measures of period-3, and that integration of the *PF**D* measure with those indices can significantly improve the accuracy of gene finding algorithms. The method developed here does not require training.

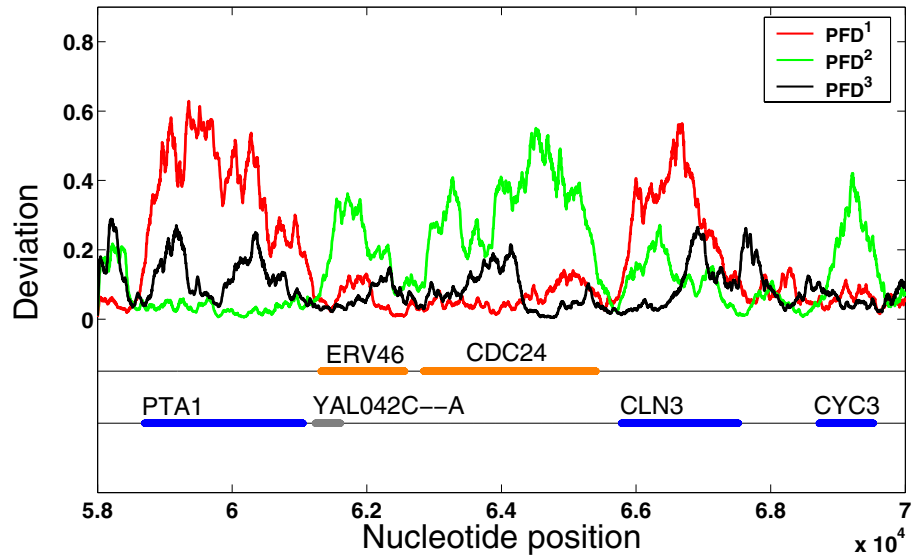


Figure 1: The PFD^i , $i = 1, 2, 3$ curves for a segment of DNA in yeast chromosome I (from nucleotide 58000 to 70000). Colored horizontal bars on the two lines below the deviation curves are the open reading frames on the two strands of the chromosome, (first line: positive strand; second line: reverse strand). The orange and blue bars represent verified ORFs while a gray bar represents a dubious ORF. It is interesting to note that in a coding region, the three reading-frame specific PFD curves separate considerably with one of them being very large. That reading frame is the one that contains the ORF.

| coding (n_1, N_1) | non-coding (n_2, N_2) | Sensitivity/Specificity | | |
|--------------------------|------------------------------|-------------------------|-------|-------|
| | | w=512 | w=128 | w=64 |
| (1, 4125) | (1, 5993) | 80.1% | 83.3% | 81.2% |
| (256, 4067) | (256, 4186) | 83.6% | 84.6% | 81.9% |
| (512, 3756) | (512, 1948) | 87.8% | 86.6% | 83.8% |
| (1026, 2674) | (512, 1948) | 90.4% | 88.4% | 85.8% |
| (1026, 2674) | (1026, 650) | 93.8% | 91.2% | 88.5% |

Table 1: Accuracy of the PFD -based coding-region identification algorithm on different coding/noncoding subsets. The parameters N_1 and N_2 are the number of coding and non-coding sequences with length greater than n_1 and n_2 , respectively. A DNA segment is declared “coding” if the measure is larger than a threshold, and “non-coding” otherwise. Accuracy is defined as the average of sensitivity and specificity. The threshold is set at where sensitivity equals specificity.