Gene Prediction Analysis Based on Spectral Representation of DNA Sequences and Fuzzy Clustering

Daniel Kotlar[(1)] , Zohar Idelson[(2)] and Yizhar Lavner[(1,2)]

[(1)] Tel-Hai Academic College, Upper Galilee, Israel

[(2)] Signal and Image Processing Lab (SIPL), Faculty of Electrical Engineering, Technion IIT, Haifa, Israel

(1) Introduction:

We propose new measures for gene prediction in eukaryotes. The DFT at a frequency of 1/3 is computed for the four binary sequences for A, T, C, and G. The different distributions of the spectral phases in coding vs. non-coding regions are used to construct the *Spectral Rotation* measure for gene finding.

Another proposed measure is based on representing each frame as a point in a three-dimensional complex space. The points corresponding to coding vs. non-coding regions tend to cluster differently. These discriminating features for gene prediction are shown to be potential candidates for locating short genes and exons.

(2) Objectives:

- Presenting new methods for gene prediction, based on the DFT magnitude and phase at a frequency of 1/3, computed for the four binary sequences for A, T, C, and G

- Developing algorithm for gene prediction, using a representation of each genomic sequence as a point in a three-dimensional complex space, where the coordinates of this space are the DFT's of three out of four nucleotides.

(3) Algorithm (I): Spectral Rotation Measure

Calculating the DFT of a DNA sequence*

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

$$U_b\left(\frac{N}{3}\right) = \sum_{n=0}^{N-1} u_b(n) e^{-i\frac{2\pi}{3}n} \qquad b = A,T,C,G$$

S(n) ATCGTACAGCTGCAAAGCATAGATTCGGTCACAGTTG...

$u_A(n)$ 1000010100000111001010100000001010000

$u_T(n)$ 01001000001...

$u_C(n)$ 0010001001...

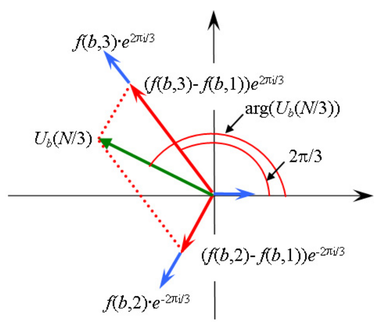
$u_G(n)$ 0001000010001...

$$\frac{1}{N}DFT\left(\frac{N}{3}\right) \rightarrow \frac{1}{N}U_A\left(\frac{N}{3}\right),\ \frac{1}{N}U_T\left(\frac{N}{3}\right),\ \frac{1}{N}U_C\left(\frac{N}{3}\right),\ \frac{1}{N}U_G\left(\frac{N}{3}\right)$$

*Silverman and Linsker 1986; Voss 1992

Fourier Spectra and Position Asymmetry

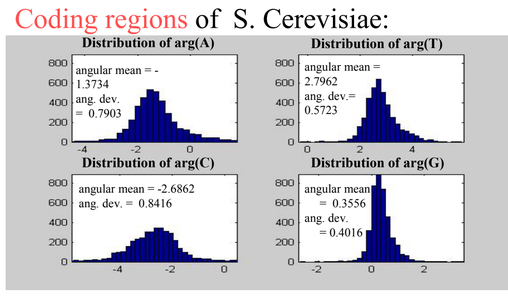
$$U_b\left(\frac{N}{3}\right) = f(b,1)\cdot 1 + f(b,2)\cdot e^{-i\frac{2\pi}{3}} + f(b,3)\cdot e^{i\frac{2\pi}{3}}$$



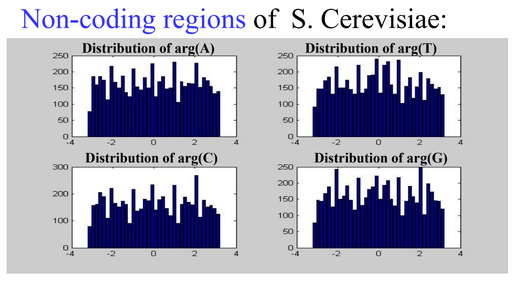
$f(b,i)$ is the frequency of the base b in the codon position i, i=1,2,3.

Distribution of the phase of the DFT at the freq of 1/3

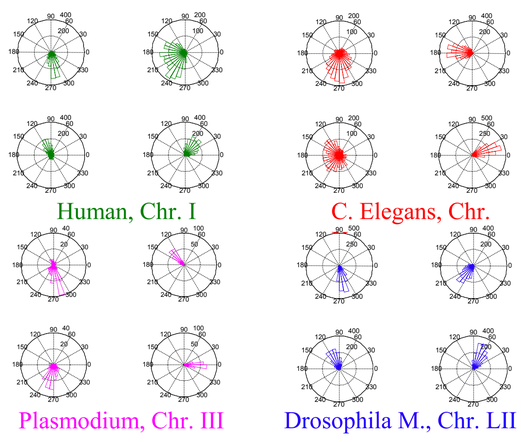
Coding regions of S. Cerevisiae:



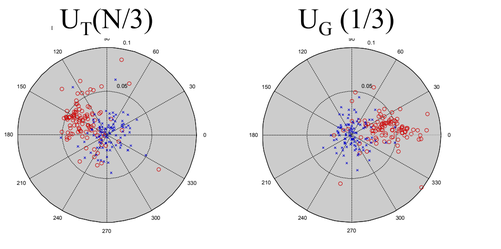
Non-coding regions of S. Cerevisiae:



Coding regions of other organisms



Coding versus non-coding regions



	Non-coding regions	*Coding regions*
Magnitude	**small**	**LARGE**
Phase	**Randomly distributed**	**Narrow distribution**

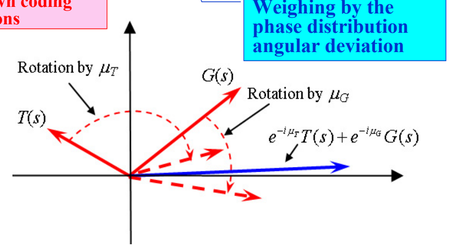
The Spectral Rotation measure

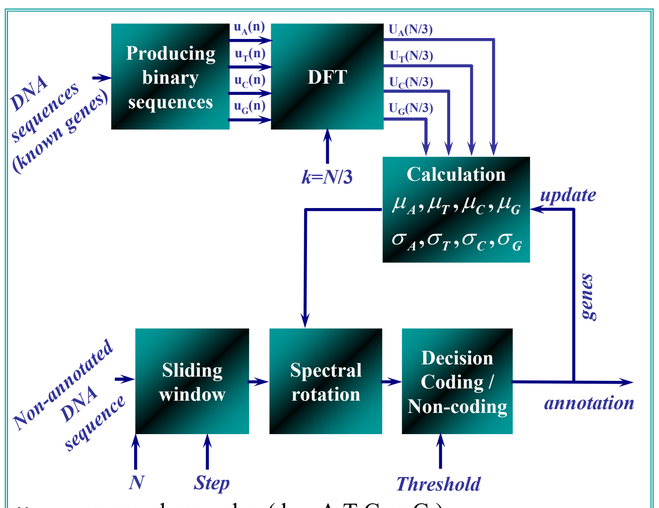
$$|V|^2 = \left|\frac{e^{-i\mu_A}}{\sigma_A}A(s) + \frac{e^{-i\mu_T}}{\sigma_T}T(s) + \frac{e^{-i\mu_C}}{\sigma_C}C(s) + \frac{e^{-i\mu_G}}{\sigma_G}G(s)\right|^2$$

Clockwise rotation by the Fourier spectrum average phase in known coding regions

Fourier spectrum in base T of the tested DNA sequence

Weighing by the phase distribution angular deviation





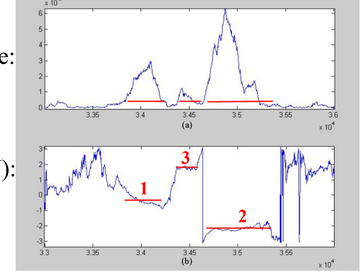
μ_b – average phase value (b = A,T,C or G)

σ_b – the corresponding angular deviation

Reading-Frame Identification

$$\arg(V) \approx \begin{cases} 0 & \text{s in reading frame 1} \\ -\frac{2\pi}{3} & \text{s in reading frame 2} \\ \frac{2\pi}{3} & \text{s in reading frame 3} \end{cases}$$

Example: Gene "SPBC1709.08" of *S. Pombe*:



(3) Algorithm (II): Classification by Clustering

Learning Stage

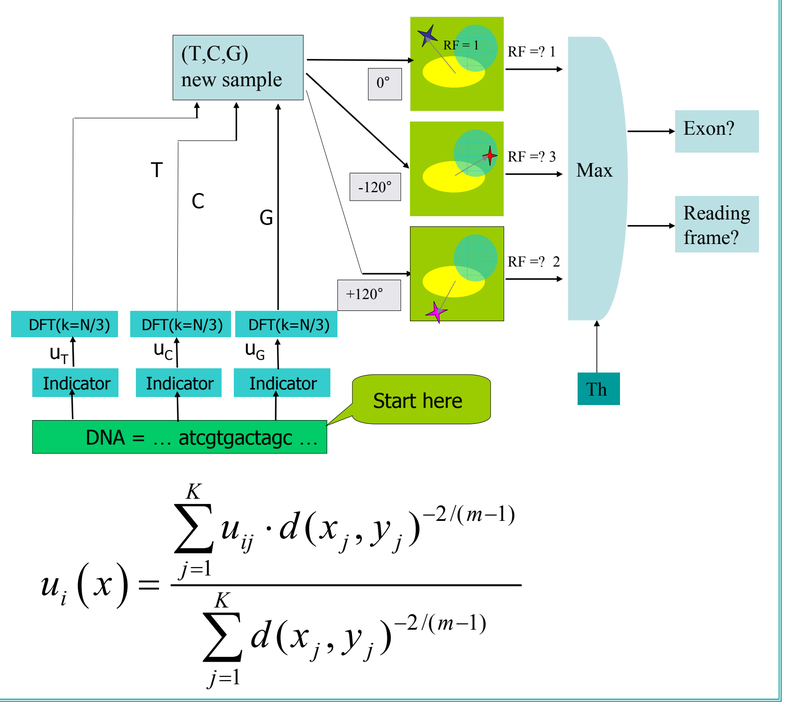
- Computing DFT values for the binary signals of G, C, and T, for each labeled frame (exon or non- exon)
- Representing each frame as a point in $\mathbb{C}^3$
- Clustering using fuzzy K-means (2 clusters for both exons and non-exons, each with its own centroid).
- Computing centroids and membership values:

$$y_i = \frac{\sum_{j=1}^{N}(u_{ij})^m x_j}{\sum_{j=1}^{N}(u_{ij})^m} \qquad u_{ij} = \frac{1}{d(x_j,y_i)^{2/(m-1)}\sum_{l=1}^{K} d(x_j,y_l)^{-2/(m-1)}}$$

Classification Stage

- Two methods for classifying exon/non-exon:
  1. Adding exons and non-exons scores, and max sum wins.
  2. Max centroid score wins.
- 2nd method used
- Scores sums are used for reading frame: max r.f. wins.

- Fuzzy KNN: creating a fuzzy membership function and choosing the one with the highest score.

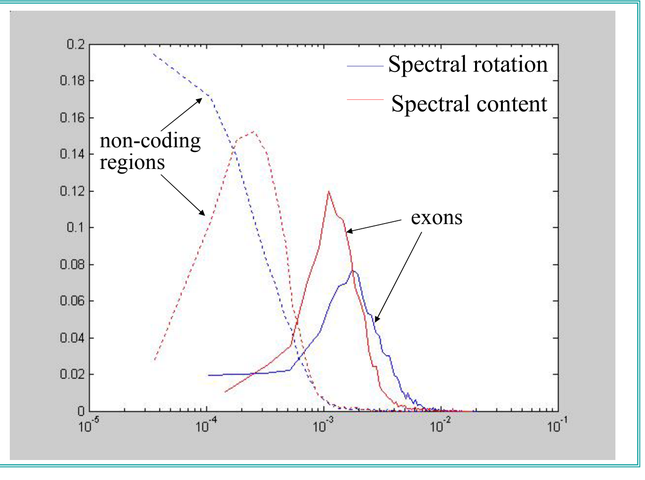


$$u_i(x) = \frac{\sum_{j=1}^{K} u_{ij}\cdot d(x_j,y_j)^{-2/(m-1)}}{\sum_{j=1}^{K} d(x_j,y_j)^{-2/(m-1)}}$$

(4) Results

(I) Spectral Rotation

measure	% of exons detected for 10% false positive		
	120 bp	180 bp	351 bp
Spectral Rotation	**88.0**	**90.8**	**93.0**
Other Fourier measures:			
Optimized Spectral Content (Anastassiou)	**86.0**	**89.3**	**92.2**
Spectral Content (Tiwari et al.)	**79.4**	**86.3**	**90.4**



(II) Classification by Clustering

Total exons	Correct identification	False positive
5376	> 91%	< 10%

Reference: Kotlar D., Lavner Y., Gene prediction by spectral rotation (SR) measure – a new method for identifying protein-coding regions, *Genome Research* 13, 8: 1930-1937 (2003).