Gene Prediction by Spectral Rotation (SR) Measure: A New Method for Identifying Protein-Coding Regions.

Daniel Kotlar and Yizhar Lavner

Department of Computer Science, Tel-Hai Academic College, Upper Galilee, 12210, Israel.

All correspondence to be sent to:
Dr. Yizhar Lavner
Department of Computer Science,
Tel-Hai Academic College,
Upper Galilee, 12210, Israel.
e-mail: yizhar_1@kyiftah.org.il
Phone: 972-4-6952965
Fax: 972-4-6952899

Running title: Gene prediction spectral analysis.

Keywords: Gene prediction, DNA spectral analysis, spectral rotation, position asymmetry, codon usage.
ABSTRACT

A new measure for gene prediction in eukaryots is presented. The measure is based on the DFT phase at a frequency of 1/3, computed for the four binary sequences for A, T, C, and G. Analysis of all the experimental genes of *S. Cerevisiae*, revealed distribution of the phase in a bell-like curve around a central value, in all four nucleotides, while the distribution of the phase in the non-coding regions was found to be close to uniform. Similar findings were obtained for other organisms. Several measures based on the phase property are proposed. The measures are computed by clockwise rotation of the vectors, obtained by DFT for each analysis frame, by an angle equal to the corresponding central value. In protein coding regions, this rotation is assumed to closely align all vectors in the complex plane, thereby amplifying the magnitude of the vector sum. In non-coding regions this operation does not significantly change this magnitude. Computing the measures with one chromosome, and applying them on sequences of others reveals improved performance compared with other algorithms that use the 1/3 frequency feature, especially in short exons. The phase property is also used to find the reading frame of the sequence.
INTRODUCTION

Gene prediction analysis, and specifically, the computational methods for finding the location of protein-coding regions in uncharacterized genomic DNA sequences, is one of the central issues in bioinformatics (Fickett 1996; Salzberg et al. 1998). For a given DNA sequence of an organism, in which the genes and other functional structures are not already known, it is very important to have an accurate and reliable tool for automatic annotation of the sequence: the number and location of genes, the location of exons and introns (in eukaryots), and their exact boundaries (Claverie 1997). Therefore, along with standard molecular methods, many new methods for finding distinctive features of protein-coding regions have been proposed in the past two decades (see reviews by Fickett 1996; Claverie 1997; Mathé et al. 2002). These methods are based on different measures for discriminating between protein-coding regions and non-coding regions. Some of the measures are based on statistical regularities in genes or exons, which are not present in introns and intergenic sections, such as, for example, differences in codon usage (Staden and McLachlan 1982), hexamer counts (Claverie et al. 1986; Farber et al. 1992; Fickett and Tung 1992), codon position asymmetry (Fickett 1982), different periodicities (Fickett 1982; Silverman and Linsker 1986; Chechetkin and Turygin 1995; Tiwari et al. 1997; Herzel et al. 1998; Trifonov 1998; Herzel et al. 1999; Anastassiou 2000), autocorrelations, nucleotide frequencies (Shulman et al. 1981; Fickett 1982; Borodovsky et al. 1986, 1994), entropy measures (Almagor 1985), and many others. Other measures are based on signals of the gene expression machinery (reviewed in Mathé et al. 2002). Sophisticated algorithms for gene prediction based on both types of measures have been proposed. These algorithms use, for instance, artificial neural networks (Lapedes et al. 1990; Uberbacher and Mural 1991; Farber et al. 1992; Xu et al. 1994; Snyder and Stormo 1995), Hidden Markov Models (Krogh et al. 1994; Baldi and Brunak 1998), and linguistic methods (Searles 1992; Dong and Searls 1994; Mantegna et al. 1994).

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, since discriminative statistical characteristics are less likely to appear in short strands. Furthermore, some genes do not possess the characteristic features that identify most genes, and hence it is not possible to track them using gene predictors that rely on these features.

In this paper a new discriminating feature for gene prediction is proposed. This measure is based on the arguments of the Discrete Fourier Transform (DFT), and is shown to be a potential candidate for locating short genes and exons. The paper is organized as follows: In section 2, frequency analysis of DNA sequence is presented. Section 3 details the Fourier analysis at a frequency of 1/3, and discusses the relationship between spectral arguments and the position frequencies. Section 4 introduces the distribution of the arguments of Fourier spectra. The applications for gene prediction are described in sections 5 and 6.
METHODS

Periodicities in DNA sequences and DFT analysis

The importance of measuring different periodicities for a given DNA, in order to determine the locations of protein-coding regions has already been addressed by Fickett (1992, 1996), and these periodicities have been used as discriminant features in several studies of gene prediction (Silverman and Linsker 1986; Fickett 1992, 1996; Chechetkin and Turygin 1995; Tiwari et al. 1997; Herzel et al. 1999; Anastassiou 2000). The Discrete Fourier Transform (DFT) is a powerful tool for studying periodicities.

The DFT of a given numeric sequence $x(n)$ of length $N$ is defined by

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

(Oppenheim and Schafer 1999), where $n$ is the sequence index, and $k$ corresponds to a period of $N/k$ samples, or discrete frequency of $\frac{2\pi k}{N}$.

Since the DNA sequence is a character string, numerical values must be assigned to each character: A, T, C, and G. One possible way of performing this conversion is to assign a binary sequence to each of the four bases (Voss 1992). This binary sequence will take a value of 1 or 0 at location $n$ of the sequence, depending on the existence or absence of that base. Thus we have four binary sequences, one for each base, denoted by $u_A(n)$, $u_T(n)$, $u_C(n)$, and $u_G(n)$, respectively (Anastassiou 2000; Voss 1992).

Applying the DFT to each of these sequences produces four spectral representations, denoted as $U_A(k)$, $U_T(k)$, $U_C(k)$, and $U_G(k)$, respectively. That is, for a base $b$ (b=A,T,C, or G), the DFT of the binary sequence $u_b(n)$ of length $N$ is

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

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

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

(Silverman and Linsker 1986; Tiwari et al. 1997)

A distinctive feature of protein-coding regions in DNA is the existence of short-range correlations in the nucleotide arrangement, especially a 1/3-periodicity (Fickett 1982), arising from the fact that coding DNA consists of triplets (codons). As a consequence, the total Fourier spectrum of protein coding DNA (equation (2.3)) typically has a peak at the frequency $k = N/3$, while the total Fourier spectrum of non-coding DNA generally does not have any significant peaks (Tsonis et al. 1991; Voss 1992; Chechetkin and Turygin 1995). Tiwari et al. (1997) used the measure in equation (2.3) with $k = N/3$, known as the
Spectral Content measure, to construct a gene predictor. It can be shown that this measure is the same (up to a $3/2$ multiplicative factor) as the sum of the four Position Asymmetry measures (Fickett and Tung 1992), namely

\begin{equation}
S \left( \frac{N}{3} \right) = \frac{3}{2} \left( \text{asymm}(A) + \text{asymm}(T) + \text{asymm}(C) + \text{asymm}(G) \right)
\end{equation}

Where, for a base $b$ (b=A,T,C, or G) $\text{asymm}(b) = \sum_{i=1}^{3} \left( f(b,i) - \mu(b) \right)^2$, $\mu(b) = \frac{1}{3} \sum_{i=1}^{3} f(b,i)$, and $f(b,i)$ is the frequency of $b$ in the codon position $i$, $i=1,2,3$.

Anastassiou (2000) introduced the Optimized Spectral Content measure

\begin{equation}
|W|^2 = aU_A \left( \frac{N}{3} \right) + tU_T \left( \frac{N}{3} \right) + cU_C \left( \frac{N}{3} \right) + gU_G \left( \frac{N}{3} \right)
\end{equation}

In this measure, the coefficients $a, t, c,$ and $g$ are calculated using an optimization technique applied to the known genes of a given organism. The measure in equation (2.5) shows significant improvement over the measure presented by Tiwari et al. (1997) in predicting genes in $S. Cerevisiae$ (Anastassiou 2000).

In the following sections we show how signal processing measures for gene prediction can be improved by considering a new feature of protein-coding DNA regions, which can be measured by DFT, namely, the arguments of the Fourier spectra at $N/3$. We show that the arguments of the Fourier spectra in coding regions are narrowly distributed around corresponding central values.

**Relationship between spectral arguments and position frequencies**

As evident from equation (2.4), the peak of the total Fourier spectrum at $k = N/3$ in protein-coding DNA sequences, is directly related to the asymmetric distribution of each of the four bases among the three codon positions. This asymmetry is strongly related to the codon usage of the particular organism. For any given organism, most genes have similar codon usage; therefore, in most protein coding-regions, the ratios between each pair of the counters $\left\{ f(b,i) \right\}_{i=1}^{3}$ for each base $b$ can be expected to be close to some constant values. These ratios determine the value of $\arg \left( U_b \left( \frac{N}{3} \right) \right)$. To demonstrate, let $s$ be a DNA sequence. The $\frac{N}{3}$-th element of the DFT of the binary sequence $u_b(n)$ of length $N$ associated with the base $b$ (b=A, T, C, or G) is obtained by substituting $k = N/3$ in equation (2.2):

\begin{equation}
U_b \left( \frac{N}{3} \right) = \sum_{n=0}^{N-1} u_b(n) e^{-\frac{2\pi i n}{3}}
\end{equation}

Since $u_b(n) = 0$ or 1, there are three distinct possible nonzero terms in the sum in (3.1), namely $1$, $e^{-\frac{2\pi i}{3}}$, and $e^{\frac{2\pi i}{3}}$, and equation (3.1) takes the form:
Since, \(1 + e^{-\frac{2\pi}{3}} + e^{\frac{2\pi}{3}} = 0\) equation (3.2) can be expressed as follows

\[
U_b \left( \frac{N}{3} \right) = \left( f(b,1) - f_{\min} \right) \cdot 1 + \left( f(b,2) - f_{\min} \right) \cdot e^{-\frac{2\pi}{3}} + \left( f(b,3) - f_{\min} \right) \cdot e^{\frac{2\pi}{3}}
\]

where \(f_{\min} = \min \{f(b,i)\}_{i=1}^{3}\). If all \(f(b,i)\), \(i=1,2,3\), are equal, then \(U_b \left( \frac{N}{3} \right) = 0\). Figure 1 illustrates the case where \(f_{\min} = f(b,1)\):

![Figure 1: Computing \(U_b \left( \frac{N}{3} \right)\) in the case \(f_{\min} = f(b,1)\)](image)

A simple trigonometric computation yields:

\[
\arg \left( U_b \left( \frac{N}{3} \right) \right) = \arccot \left( \frac{2 \left( \frac{f_1}{f_{\min}} - 1 \right)}{\sqrt{3} \left( \frac{f_2}{f_{\min}} - 1 \right)} \right) + \varphi
\]

where \(f_1\) and \(f_2\) are the other two counters, numbered in counter-clockwise direction from the vector corresponding to \(f_{\min}\), and \(\varphi\) is the argument of the vector corresponding to \(f_1\) \((0, 2\pi/3, \text{or } -2\pi/3)\).

It can be seen that \(\arg \left( U_b \left( \frac{N}{3} \right) \right)\) will shift by \(-2\pi/3\) or \(2\pi/3\) for reading frames 2 and 3, respectively.

Since it was assumed above that the ratios between each pair of the counters \(\{f(b,i)\}_{i=1}^{3}\) in most coding regions are close to some constant values, the same holds for \(\arg \left( U_b \left( \frac{N}{3} \right) \right)\), where \(b\) is one of the bases A, T, C, or G.

## RESULTS

The distribution of the arguments of the Fourier spectra

Let \(s\) be a DNA strand, and for each base \(b\), denote \(b(s) = U_b \left( \frac{N}{3} \right)\). We calculated the values of \(\arg(A(s))\), \(\arg(T(s))\), \(\arg(C(s))\), and \(\arg(G(s))\) in coding and non-coding regions, for different organisms. The histograms describing these distributions for all experimental genes in the 16
chromosomes of *S. Cerevisiae*\(^1\) (GenBank accession numbers NC001133 - NC001148, at [http://www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) are shown in Figure 2.\(^2\) As the figure reveals, in all four nucleotides the distributions of the arguments taper around a central value, with the distributions of arg\((G(s))\) and arg\((T(s))\) being much narrower than the other two. Similar results, both in shape and statistics, were obtained for each of the 16 chromosomes of *S. Cerevisiae*.

The corresponding histograms for non-coding regions\(^3\) in the 16 chromosomes appear in Figure 3. The distributions for non-coding regions seem to be close to uniform, and very different from the distributions that were obtained for coding regions. A similar pattern was observed for each separate chromosome.

---

\(^1\) Multiple-exon genes were concatenated to single strands

\(^2\) Since the arguments are originally in principal values (between -\(\pi\) and \(\pi\)), a \(2\pi\) shift was applied to part of the data so that the histograms are plotted around the angular mean.

\(^3\) Intergenic spacers and introns in all genes (experimental and not experimental)
To make sure that the former results are not unique for *S. Cerevisiae*, the same analysis was performed on other organisms. The resulting histograms 4a-4c show the argument distribution for *S. Cerevisiae*, *S. Pombe*, and *Guillardia theta*, respectively. It is readily evident that the three histograms greatly resemble each other, although the exact statistical values differ somewhat. In particular, the central value of arg(G), in all three organisms, is located somewhere in the vicinity of 0. This means that the base G appears in the first codon position a lot more than in the second and third. This is consistent with the findings of Trifinov (1987).

In the following section we show how the difference between coding and non-coding regions in terms of argument distribution can be applied to gene prediction.
Figure 4a: Argument distribution for all experimental genes in all chromosomes of $S$.

Figure 4b: Argument distribution for all genes in chromosomes 2 and 3 of $S$.

Figure 4c: Argument distribution for all genes in chromosome 1 of Guillardia theta
Rotational measures for gene prediction

Several measures were constructed using the argument distribution described above. These measures are based on the notion of spectral rotation and alignment.

Assume we have an organism for which \( \arg(A(s)), \arg(T(s)), \arg(C(s)), \text{ and } \arg(G(s)) \) are distributed in a similar manner to that observed in the organisms described above (i.e., bell-shaped in coding regions, and close to uniform in non-coding regions). Let \( \mu_A, \mu_T, \mu_C, \text{ and } \mu_G \) be the approximated average values, in coding regions, of \( \arg(A(s)), \arg(T(s)), \arg(C(s)), \text{ and } \arg(G(s)) \), respectively. Since for a typical coding sequence \( s \) it is expected that \( \arg(A(s)) = \mu_A, \arg(T(s)) = \mu_T, \arg(C(s)) = \mu_C, \text{ and } \arg(G(s)) = \mu_G \), rotating the vectors \( A(s), T(s), C(s), \text{ and } G(s) \) clockwise, each by the corresponding argument \( \mu_A, \mu_T, \mu_C, \text{ and } \mu_G \) (multiplication by \( e^{-i\mu_A}, e^{-i\mu_T}, e^{-i\mu_C}, \text{ and } e^{-i\mu_G} \) respectively) will yield four vectors pointing roughly in the same direction: towards the positive real axis for reading frame 1, and at \(-2\pi/3\) and \(2\pi/3\) for reading frames 2, and 3, respectively. Hence the vector sum

\[
(5.1) \quad e^{-i\mu_A}A(s) + e^{-i\mu_T}T(s) + e^{-i\mu_C}C(s) + e^{-i\mu_G}G(s)
\]

will be of large magnitude compared to the case where the vectors point in different directions, as is most likely the case for a non-coding sequence. Figures 5a and 5b illustrate this idea for the sum of two vectors \( e^{-i\mu_T}T(s) + e^{-i\mu_G}G(s) \).

Dividing each term in (5.1) by the corresponding angular deviation \( \sigma_A, \sigma_T, \sigma_C, \text{ and } \sigma_G \) of \( A(s), T(s), C(s), \text{ and } G(s) \) respectively) will give more weight to narrower distributions, yielding the measure

---

**Figure 5a:** Rotation and alignment of the vectors \( G(s) \) and \( T(s) \), when \( \arg(T(s)) = \mu_T \) and \( \arg(G(s)) = \mu_G \)

**Figure 5b:** Rotation and alignment of the vectors \( G(s) \) and \( T(s) \), when \( \arg(T(s)) \) and \( \arg(G(s)) \) are any random values
which we call a Spectral Rotation (or SR) measure. This resembles the Optimized Spectral Content measure of Anastassiou (2000) given in equation (1.3).

Table 1 compares the performance of four measures: two introduced here, namely the SR measure and the G Rotation measure (described below), and two known measures based on Fourier spectrum: The Spectral Content measure (Tiwari et al. 1997), and Optimized Spectral Content measure (Anastassiou, 2000). All measures were tested on all experimental exons, and non-coding strands (intergenic spacers and introns) from the first 15 chromosomes of S. Cerevisiae,6 of a length greater than 50 bp. The results were obtained by sliding windows of sizes 90, 120, 180, 240, 300, and 351 bp, with gaps of size 30, 40, 60, 60, 75, and 99 bp, respectively. The threshold in each case was chosen so that the percentage of introns falsely detected as exons (false positives) is 10%. As the table indicates, the SR measure shows better performance, especially in smaller analysis frames.

<table>
<thead>
<tr>
<th>Measure</th>
<th>% of exons detected for 10% false positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90 bp</td>
</tr>
<tr>
<td>Spectral Rotation</td>
<td></td>
</tr>
<tr>
<td>Optimized Spectral Content (Anastassiou)</td>
<td>84.5</td>
</tr>
<tr>
<td>Spectral Content (Tiwari et al.)</td>
<td>76.0</td>
</tr>
<tr>
<td>G Rotation</td>
<td>83.3</td>
</tr>
</tbody>
</table>

Table 1: Performance of Fourier spectrum measures on all experimental exons and all non-coding strands of length greater than 50 bp, in S. Cerevisiae, using different window sizes.

Figure 6 compares the probability density functions of the Spectral Content measure and the SR measure. The values of the Spectral Content measure were scaled so that intersection points of the two curves of each color are vertically aligned. The better performance of the SR measure is illustrated by the fact that the distance between the bold curves is greater than the distance between the fine curves.

---

6 Since the measures were calculated using chromosome 16 of S. Cerevisiae, their performance was tested on the remaining 15 chromosomes.

7 An analysis frame of 351 bp is used by Anastassiou (2000) and by Tiwari et. al. (1997). The choice of this size of frame is explained in the latter.
Detection of short exons may be rendered more effective by using one statistical parameter that is narrowly distributed when calculated over short strands. As shown in Figure 7, the distribution of \( \arg(G(s)) \) in the genes of \textit{S. Cerevisiae} is narrow when calculated over coding strands of length 120bp.

The following measure uses only \( \arg(G(s)) \). Since only the vector \( G(s) \) is rotated in this measure, we need a fixed reference vector to maximize the vector sum. Suppose \( R \) is a real number. If \( s \) belongs to a
coding region, and is in reading frame 1, the vector \( G(s) \), rotated clockwise by the approximated average argument \( \mu_G \) will most likely be directed towards the positive real axis, that is, in the same direction as the vector \( R \). On the other hand, if \( s \) does not belong to a coding region, the rotated vector may point in any direction. The vector sum of \( R \) and the rotated \( G(s) \) will discriminate best between coding and non-coding regions, when \( R \) is of the same order of magnitude as \( |G(s)| \). Thus, we choose \( |G(s)| \) as the reference vector. In order to identify genes in all three reading frames, we define the \( G \) Rotation measure as

\[
\|V_G\|^2 = \left| e^{-i\mu_G} G(s) + |G(s)| \right|^2
\]

where \( \mu_G \) is chosen from the set \( \{ \mu_G, \mu_G + \frac{2\pi}{3}, \mu_G - \frac{2\pi}{3} \} \), so that the value of the measure in equation (5.3) is maximal. Table 1 compares the performance of the \( G \) Rotation measure with that of other measures, on the experimental genes and exons of \( S. \) Cerevisiae.

When using a measure calculated from data of one organism to predict genes in another organism, it may be preferable to use a subset of the vectors \( A(s), T(s), C(s), \) and \( G(s) \). For example, the vectors \( T(s) \) and \( G(s) \), which have narrowly distributed arguments, can be aligned to yield the \( TG \)-Rotation measure \( \|V_{TG}\|^2 = \left| e^{-i\mu_T} T(s) + \frac{e^{-i\mu_G}}{\sigma_T} G(s) \right|^2 \). This measure is determined by \( |\mu_G - \mu_T| \). Hence, where this value happens to be similar in two organisms (e.g., \( S. \) Cerevisiae and \( S. \) Pombe – see Figures 4a and 4b), it is possible to predict genes in one organism by using the parameters of another.

Figure (8a) shows the curve of the \( TG \)-Rotation measure, constructed from data in chromosome 16 of \( S. \) Cerevisiae, on a typical split gene of \( S. \) Pombe (gene \( SPBC582.08 \) in chromosome II). For comparison, Figure (8b) shows the graph of the Codon Usage measure on the same gene. The horizontal lines represent the actual location of the three exons. Note the short intron between the second and third exons.

In general, one should be wary about using data from one organism to predict genes in another organism, since the respective central argument values in different organisms may not be similar. For example, note that in Figures 4a-4c the central values of \( \text{arg}(A) \) and \( \text{arg}(C) \) are very different in the three organisms.

Table 2 summarizes the data on the gene.

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start base</th>
<th>End base</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>352249</td>
<td>352711</td>
<td>463</td>
</tr>
<tr>
<td>2</td>
<td>352903</td>
<td>353702</td>
<td>800</td>
</tr>
<tr>
<td>3</td>
<td>353756</td>
<td>354010</td>
<td>255</td>
</tr>
</tbody>
</table>

Table 2: Gene \( SPBC582.08 \) in chromosome II of \( S. \) Pombe
Complementary sequences and reading frame identification

Genes on the complementary strand can be detected using the following transformation from Anastassiou (2000):

If \( V = aA(s) + iT(s) + cC(s) + gG(s) \) then the predictor for the complementary strand is \( \tilde{V} = \tilde{a}A(s) + \tilde{i}T(s) + \tilde{c}C(s) + \tilde{g}G(s) \), where \( \tilde{a} = e^{-\frac{2\pi}{3}}a' \), \( \tilde{i} = e^{-\frac{2\pi}{3}}i' \), \( \tilde{c} = e^{\frac{2\pi}{3}}c \), and \( \tilde{g} = e^{\frac{2\pi}{3}}g' \), and \( a', i', c', \) and \( g' \) are the complex conjugates of \( a, i, c, \) and \( g \) respectively.

An non-annotated strand is examined using both measures \( |V|^2 \) and \( |\tilde{V}|^2 \). A detected gene will be considered complementary if \( |\tilde{V}|^2 > |V|^2 \).

Figure 8: Graphs of gene prediction applied on the gene “SPBC582.08” in chromosome II of S. Pombe, using a sliding window of 180 bp: (a) TG-Rotation measure; (b) Codon Usage measure. The horizontal segments represent the actual location of the three exons. To get the actual base location in the chromosome add 300,000 bp to the numbers on the horizontal axis.
As mentioned in section 3, \( \arg \left( U_b \left( N/3 \right) \right) \) will shift by \(-2\pi/3\) or \(2\pi/3\), relative to its value for reading frame 1, if the actual reading frames are 2 and 3, respectively (see Anastassiou 2000). As explained in the previous section, the rotational measures identify exons, regardless of their reading frame. To identify the reading frame, for the SR measure, we look at \( \arg(V) \). For a coding sequence in reading frame 1, the rotated vectors will be aligned close to the positive real axis, and thus \( \arg(V) \) should be close to zero. For reading frames 2 and 3, \( \arg(V) \) will be in the vicinity of \(-2\pi/3\) and \(2\pi/3\), respectively. This is illustrated by the following two examples.

Figure 9a depicts the graph of the SR measure on the gene \( SPBC1685.08 \) in chromosome 2 of \( S.\ Pombe \) (GenBank accession number NC003423). The measure’s parameters were calculated from the genes in the smaller chromosome 3 (GenBank accession number NC003421). Figure 9b depicts the graph of \( \arg(V) \). The graphs were obtained by calculating the measure with a sliding window of 351 bp, using a step of 3 bp. The gene has three exons, in reading frames 1, 3, and 2 respectively. Table 3 summarizes the data on the gene. Note the short intron between the second and third exons.

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start base</th>
<th>End base</th>
<th>Length</th>
<th>Reading frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>433915</td>
<td>434252</td>
<td>338</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>434423</td>
<td>434626</td>
<td>204</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>434671</td>
<td>435403</td>
<td>733</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3: Gene \( SPBC1685.08 \) in chromosome II of \( S.\ Pombe \)
Figure 9 illustrates how the curve of $\text{arg}(V)$ can be used to identify the exact boundaries of an exon. It is expected that along an exon the value of $\text{arg}(V)$ will remain in the vicinity of one of the values 0, $-2\pi/3$ or $2\pi/3$, while outside the exon, the value of $\text{arg}(V)$ will change to some “random” value.

Figure 10a depicts the graphs of the SR measure on the gene $SPBC1709.08$ in chromosome 2 of $S.\, Pomba$ (GenBank accession number NC003423). The measure’s parameters were calculated as in the previous example. Figure 10b depicts the graph of $\text{arg}(V)$. In this example, the gene has one exon, between nucleotides 1033037 and 1037362, in reading frame 2. The fact that the curve of $\text{arg}(V)$ is constant at around the value of $-2\pi/3$ along the whole gene indicates that the gene consists of one exon, and not of multiple exons, as might be incorrectly deduced by looking only at the curve of Figure 10a. This procedure can therefore assist in differentiating between multiple exons and single exons.
DISCUSSION

In this paper a new method for gene prediction is proposed, based on several measures of protein coding regions. The measures are derived from a regularity of the spectral phase within coding regions. In this study we found that the phase of the DFT at a frequency of 1/3 is distributed with a bell-like-shaped curve around a central value in coding regions, while in non-coding regions, the distribution was uniform-like. This behavior was shown to exist in all chromosomes of *S. Cerevisiae*, and also in two other organisms, namely, *S. Pombe* and *Guillardia theta*. This regularity was used for the construction of measures for discriminating between coding and non-coding regions in a given non-annotated DNA sequence. The measures are constructed by clockwise rotation of the vectors, which are the values obtained by DFT analysis for the four binary sequences of each nucleotide, with the corresponding central values. After such rotation, the four vectors in coding regions tend to be aligned close to each other, while the arrangement of vectors in non-coding regions is random. Earlier studies have proposed measures for gene prediction based on Fourier transform at a frequency of 1/3 or at other frequencies (Trifonov and Sussman 1980; Fickett 1982; Silverman and Linsker 1986; Fickett and Tung 1992; Tiwari et al. 1997; Anastassiou 2000). In most of these studies, the information was derived from the magnitude of the DFT, while the information of the phase component was not explicitly used. Tiwari et al. (1997) used the magnitude to construct the so-called *Spectral Content* measure (see equation (2.3)). Anastassiou (2000) improved on the former measure by proposing the *Optimized Spectral Content* measure (see equation (2.5)), which is based on an optimization technique. In this measure, the Fourier component for

![Graph of the SR measure on the gene “SPBC1709.08” in chromosome II of *S. Pombe*, using a sliding window of 351 bp. (a) The measure; (b) arg(V). The horizontal segment represents the actual location of the gene. To get the actual base location in the chromosome, add 1,000,000 bp to the numbers on the horizontal axis.](image-url)
each nucleotide was multiplied by a coefficient in order to maximize an optimization criterion for
discrimination between coding regions and random DNA sequence. However, this technique was not
justified analytically in order to explain why it yields better performance than the measure of Tiwari et al.
(1997) and its optimization criterion, which discerns between introns (and intergenic spacers) and exons,
is based on random DNA. Since introns and intergenic spacers might reveal non-random characteristics,
it is assumed that better results could be achieved if introns or sequences from intergenic spacers were
used in the optimization. However, in the construction of the measures proposed in the present work,
there is no need for random DNA or for introns, since the rotation parameters are the central values of the
spectral phases in coding regions.

The attempt to use parameters derived from one organism to recognize genes in another organism is
based on an implicit assumption of the universality of genes, at least with regard to the structure that
elicits the above spectral features. However, as this study (and also previous ones) show, the peak at a
frequency of 1/3 is attributed to position asymmetry of the nucleotide within the 3 possible locations in
the codon. This asymmetry was shown to be the result of codon usage (Tsonis et al., 1991) and codon
bias. Since different organisms exhibit different codon usage, it is expected that such prediction will not
be optimal for use in organisms with different codon usage. Using the Spectral Rotation measure
presented here, better performance was achieved than in both studies mentioned above (of Tiwari et al.
1997 and Anastassiou 2000) (Table 1). As mentioned, the measures proposed in the current study yielded
improved results even in short analysis frames (120 bp and even 90 bp). This was notably true for the
measure based only on G. Assuming a narrow distribution of \( \arg(G) \), as is the case for the organisms
studied, the relative simplicity of computing the DFT for only one nucleotide makes the \( G \) Rotation
measure a fitting candidate to serve for identification of short genes and exons. Indeed, in this work it
was shown that this measure outperforms other known measures (Table 1). For other organisms, if the
existing gene data enables identification of the base \( b \) (\( b = A, T, C, \) or \( G \)) for which \( \arg(b) \) is most
narrowly distributed, it is possible to construct a \( b \) Rotation measure accordingly.

Considering the argument distributions obtained in this study, it was predicted that wherever an analysis
frame slides within a protein-coding region, the value of \( \arg(V) \) (the vector sum of the rotated spectra)
will be close to one of 3 possible values (0, -2\( \pi/3 \) or 2\( \pi/3 \), according to the reading frame), and random
in introns or between genes. Furthermore, the slope of the curve will be close to zero in sections
corresponding to protein coding regions, and will have a noisy unpredicted appearance elsewhere.
Therefore the plot of \( \arg(V) \) can be a tool for finding the reading frame. Moreover, as shown in section 6,
plotting the graph of the SR measure, along with \( \arg(V) \) can help to distinguish between one long exon
and multiple exons spaced by short introns. While the angle’s slope will tend to be close to zero in the
former case, it will have a noisy structure in the intron sections in the latter. This feature was also shown
to help in the exact demarcation of the exon-intron boundaries.
Last, a comment about the length of the analysis frame. A short analysis frame (less than 180 bp) may detect short exons and short introns, while frames of over 300 bp may miss. However, there is a tradeoff, since the use of shorter analysis frames causes more statistical fluctuations, resulting in more false negatives and false positives. Hence, it is important to have a measure that still performs reasonably with short frames.

In summary, we suggest that considering the arguments of the Fourier spectra at k=N/3 yields more information about a DNA sequence than the corresponding magnitudes alone. However, it should be noted that these two values (namely, the magnitude and the argument) are not independent. A large magnitude of a Fourier spectrum at k=N/3 is a result of a sharp position asymmetry in the corresponding base. If a sharp position asymmetry is characteristic of the coding regions of an organism, then the value of \( \text{arg}(U_b(N/3)) \) will be more stable, that is, its distribution over the genes of the organism will have low variance. However, as shown in this work, incorporating data about the distribution of the arguments of the Fourier spectra at k=N/3, along with their magnitudes, into a measure, yields a measure that is more sensitive to exon-intron transition than a measure that uses the magnitudes alone.

ACKNOWLEDGEMENTS

We thank Mr. Efim Yakir for preparing part of the figures and for technical support. We also thank Dr. Dorit Shweiki for reading the manuscript and for valuable discussions.

REFERENCES


WEBSITE REFERENCES

Figure legends

Figure 1: Computing $U_b(N/3)$ in the case $f_{\text{min}} = f(b,1)$

Figure 2: Argument distributions for all experimental genes in all chromosomes in *S. Cerevisiae*

Figure 3: Argument distribution for non-coding regions in all chromosomes in *S. Cerevisiae*

Figure 4a: Argument distribution for all experimental genes in all chromosomes of *S. Cerevisiae*.

Figure 4b: Argument distribution for all genes in chromosomes 2 and 3 of *S. Pombe*

Figure 4c: Argument distribution for all genes in chromosome 1 of *Guillardia theta*

Figure 5a: Rotation and alignment of the vectors $G(s)$ and $T(s)$, when $\arg(T(s)) = \mu_r$ and $\arg(G(s)) = \mu_c$

Figure 5b: Rotation and alignment of the vectors $G(s)$ and $T(s)$, when $\arg(T(s))$ and $\arg(G(s))$ are any random values

Figure 6: Probability density functions for *Spectral Rotation* (bold) and Spectral Content (fine) measures (solid lines represent exons and dashed lines represent non-coding regions)

Figure 7: Argument distribution of coding DNA strands of length 120 bp in *S. Cerevisiae*

Figure 8: Graphs of gene prediction applied on the gene “SPBC582.08” in chromosome II of *S. Pombe*, using a sliding window of 180 bp: (a) TG-Rotation measure; (b) Codon Usage measure. The horizontal segments represent the actual location of the three exons. To get the actual base location in the chromosome add 300,000 bp to the numbers on the horizontal axis.

Figure 9: Graphs of the SR- measure on the gene *SPBC1685.08* in chromosome II of *S. Pombe*, using a sliding window of 351 bp. (a) The measure; (b) $\arg(V)$. The horizontal segments represent the actual location of the exons. To get the actual base location in the chromosome, add 400,000 bp to the numbers on the horizontal axis.

Figure 10: Graphs of the SR measure on the gene “SPBC1709.08” in chromosome II of *S. Pombe*, using a sliding window of 351 bp. (a) The measure; (b) $\arg(V)$. The horizontal segment represents the actual location of the gene. To get the actual base location in the chromosome, add 1,000,000 bp to the numbers on the horizontal axis.
Figures:

Figure 1:
Figure 2:
Figure 3:
Figure 4:
Figure 5:
Figure 7:
Figure 8:
Figure 9:
Tables:

Table 1:

<table>
<thead>
<tr>
<th>Measure</th>
<th>90 bp</th>
<th>120 bp</th>
<th>180 bp</th>
<th>240 bp</th>
<th>300 bp</th>
<th>351 bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral Rotation</td>
<td>84.5</td>
<td>88.0</td>
<td>90.8</td>
<td>91.9</td>
<td>92.7</td>
<td>93.0</td>
</tr>
<tr>
<td>Optimized Spectral Content (Anastassiou)</td>
<td>82.1</td>
<td>86.0</td>
<td>89.3</td>
<td>91.0</td>
<td>91.7</td>
<td>92.2</td>
</tr>
<tr>
<td>Spectral Content (Tiwari et al.)</td>
<td>76.0</td>
<td>79.4</td>
<td>86.3</td>
<td>89.4</td>
<td>90.0</td>
<td>90.4</td>
</tr>
<tr>
<td>G Rotation</td>
<td>83.3</td>
<td>86.2</td>
<td>89.6</td>
<td>90.4</td>
<td>90.8</td>
<td>90.7</td>
</tr>
</tbody>
</table>

Table 1: Performance of Fourier spectrum measures on all experimental exons and all non-coding strands of length greater than 50 bp, in *S. Cerevisiae*, using different window sizes.

Table 2:

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start base</th>
<th>End base</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>352249</td>
<td>352711</td>
<td>463</td>
</tr>
<tr>
<td>2</td>
<td>352903</td>
<td>353702</td>
<td>800</td>
</tr>
<tr>
<td>3</td>
<td>353756</td>
<td>354010</td>
<td>255</td>
</tr>
</tbody>
</table>

Table 2: Gene SPBC582.08 in chromosome II of *S. Pombe*

Table 3:

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start base</th>
<th>End base</th>
<th>Length</th>
<th>Reading frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>433915</td>
<td>434252</td>
<td>338</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>434423</td>
<td>434626</td>
<td>204</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>434671</td>
<td>435403</td>
<td>733</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3: Gene SPBC1685.08 in chromosome II of *S. Pombe*