A FAST DFT BASED GENE PREDICTION ALGORITHM FOR IDENTIFICATION OF PROTEIN CODING REGIONS

Suprakash Datta and Amir Asif

Department of Computer Science and Engineering
York University, Toronto, ON, Canada M3J 1P3

ABSTRACT

The paper provides theoretical justification for the "3-periodicity property" observed in protein coding regions within genomic DNA sequences. We propose a new classification criteria improving upon traditional frequency based approaches for identification of coding regions. Experimental studies indicate superior performance compared with other algorithms that use the 3-periodicity property.

1. INTRODUCTION

Automated identification of protein coding regions in genomic DNA sequences is a fundamental step in the computational annotation of genes. Most gene prediction algorithms [1]-[4] exploit short range correlations in the nucleotide arrangement within coding regions. In particular, the discrete Fourier transform (DFT) of a protein coding region of length N exhibits a significant peak at frequency $(k = N/3)$. No such peak is observed in noncoding region. This characteristic, referred to as the 3-periodicity property [1, 2], has been used in [1] and [3] to design gene prediction algorithms. In this paper, we provide theoretical justification for the 3-periodicity property by defining a new parameter referred to as the position count function (PCF), which measures the number of times different nucleotides appear in the three phases within a DNA codon.

A second contribution of the paper is to improve the DFT based gene prediction algorithm [3].

- In the proposed algorithm, the DFT spectrum at $k = N/3$ is computed by parsing the DNA sequence in codons and counting the number of different nucleotides at each of three phases. The approach is of $O(N)$ and is computationally efficient than the fast Fourier transform (FFT).
- The proposed algorithm normalizes the DFT spectrum with average energy present in the DFT coefficients. The resulting parameter, referred to as the signal-to-noise ratio, provides an automated approach in predicting coding regions.
- Reference [3] applies rectangular windows to parse the DNA sequence, which causes abrupt truncations of the DNA sequence and results in extraneous peaks in the DFT spectrum. We show that the Bartlett window provides improved results and removes most of these extraneous peaks.

Section 2 defines the position count function and derives important properties for binary sequences. Section 3 provides theoretical justification for the 3-periodicity property, while section 4 introduces the proposed DFT based splicing algorithm. Section 5 presents our experimental results. Finally, in section 6, we conclude the paper.

This work was supported in part by the Natural Science and Engineering Research Council (NSERC), Canada under Grant No. 228415-03.

2. NUMERICAL REPRESENTATION

Viewed at the primary level, a DNA sequence $D[i]$ consists of four nucleotides $\{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\} \in \mathcal{A}$. The DNA sequence is mapped into binary signals $A[i]$, $T[i]$, $C[i]$ and $G[i]$, which indicate the presence or absence of these nucleotides at location i. For example, the binary signal $A[i]$, attributed to nucleotide A, takes a value of 1 at $i = i_o$ if $D[i_o] = \mathtt{A}$. Else, $A[i_o]$ is 0. For the DNA sequence

$$D[i] = [\mathtt{A\ T\ G\ A\ C\ T\ A\ A\ G\ A\ G\ A\ T\ C\ C\ G\ G}], \quad (1)$$

the numerical representation is given by

$$\begin{aligned} A[i] &= [1\ 0\ 0\ 1\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 0\ 0\ 0\ 0\ 0] \\ T[i] &= [0\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0] \\ C[i] &= [0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0] \\ \text{and } G[i] &= [0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 0\ 0\ 1\ 1]. \end{aligned}$$

We derive the properties for binary signal $A[i]$ in terms of a new parameter, position count function, which is defined next.

Position Count Function: A binary signal $A[i]$, $(0 \leq i < N)$ is parsed into nonoverlapping words of length w, $(3 \leq w < N)$. The position count function (PCF) for $A[i]$ is defined as

$$\mathcal{C}_w^A(s) = \sum_{i=0}^{\lfloor \frac{N-1}{w} \rfloor} A[wi + s] \text{ for } (0 \leq s < w), \quad (2)$$

and counts the number of 1's at phase s in the w-bit parsed words. For the DNA sequence given in (1), $\mathcal{C}_3^A(0) = 4$, $\mathcal{C}_3^A(1) = 1$, and $\mathcal{C}_3^A(2) = 1$. We now present important properties for binary sequences in terms of the PCF's.

Theorem 1: The magnitude of the DFT $\widetilde{A}[k]$ of the binary signal $A[i]$, at discrete frequency $k = N/3$, is given by

$$\begin{aligned} |\widetilde{A}[N/3]|^2 &= \frac{1}{2}\left[(\mathcal{C}_3^A(0) - \mathcal{C}_3^A(1))^2 + (\mathcal{C}_3^A(1) - \mathcal{C}_3^A(2))^2\right. \\ &+ \left.(\mathcal{C}_3^A(2) - \mathcal{C}_3^A(0))^2\right]. \quad (3) \end{aligned}$$

Proof: By definition $\widetilde{A}[N/3] = \sum_{n=0}^{N-1} A[i]e^{-\frac{j2\pi n}{3}}$, which is rearranged as

$$\begin{aligned} \widetilde{A}[N/3] &= \underbrace{\sum_{n=0,3,\ldots} A[i]e^{-\frac{j2\pi n}{3}}}_{\text{Term 1}} + \underbrace{\sum_{n=1,4,\ldots} A[i]e^{-\frac{j2\pi n}{3}}}_{\text{Term 2}} + \\ &\ldots + \underbrace{\sum_{n=2,5,\ldots} A[i]e^{-\frac{j2\pi n}{3}}}_{\text{Term 3}}. \quad (4) \end{aligned}$$

Substituting $n = mw$ in Term 1, $n = mw + 1$ in Term 2, and $n = mw + 2$ Term 3, leads to the following expression

$$\widetilde{A}[N/3] = \mathcal{C}_w^A(0) + e^{-\frac{j2\pi}{3}}\mathcal{C}_w^A(1) + e^{-\frac{j4\pi}{3}}\mathcal{C}_w^A(2),$$

which simplifies to Eq. (3). ∎

Theorem 1 computes the magnitude of the DFT coefficient $\widetilde{A}[N/3]$ directly from the PCF's without any complex algebra. It provides an efficient algorithm for computing $|\widetilde{A}[N/3]|$, which has a computational complexity of $O(N)$. We also observe that Eq. (3) is a *symmetric* function of the PCF's. The differences in the PCF's contribute to $|\widetilde{A}[N/3]|$ rather than the direct number of counts. The following two corollaries are derived directly from Theorem 1.

Corollary 1.1: If the PCF's for the 3-bit parsed words are equal, i.e., $\mathcal{C}_3^A(0) = \mathcal{C}_3^A(1) = \mathcal{C}_3^A(2)$, then DFT $\widetilde{A}[N/3]$ of sequence $A[i]$ is zero. If the PCF's are not equal, then $\widetilde{A}[N/3] \neq 0$. ∎

Corollary 1.2: Any permutation of the 3-bit parsed words in a binary sequence does not change the value of $|\widetilde{A}[N/3]|$. ∎

We now state Theorem 2, which expresses the average value of the DFT coefficients in terms of the PCF's. Theorem 2 is derived using the Parseval's property. To save on space, its proof is not included.

Theorem 2: The average value $|\widetilde{A}_{\text{av}}^{(1)}|^2$ of the squared magnitude, $|\widetilde{A}[k]|^2$, $(1 \leq k < N)$ of the DFT of a binary sequence (i.e., excluding the dc component $\widetilde{A}[0]$) is given by

$$|\widetilde{A}_{\text{av}}^{(1)}|^2 = \frac{1}{(N-1)}\left(N - \sum_{s=0}^{w-1} C_w^A(s)\right)\sum_{s=0}^{w-1} C_w^A(s). \quad (5)$$

Next, we generalize Theorem 1 to $3m$-bit parsed words. ∎

Theorem 3: The magnitude of the DFT $\widetilde{A}[k]$ of the binary signal $A[i]$, at discrete frequency $k = N/3m$, is given by

$$\widetilde{A}\left[\frac{N}{3m}\right] = \sum_{p=0}^{m-1} e^{-\frac{j2\pi p}{m}}\left[C_{3m}^A(3p) + e^{-\frac{j2\pi}{3m}}C_{3m}^A(3p+1) + e^{-\frac{j4\pi}{3m}}C_{3m}^A(3p+2)\right]. \quad (6)$$

The following corollary results directly from Theorem 5. ∎

Corollary 3.1: The magnitude of the DFT $\widetilde{A}[N/3m]$ in a $3m$-bit parsed binary sequence is zero if

$$C_{3m}^A(s) = C_{3m}^A(s+3) = \ldots = C_{3m}^A(s+3k) \quad (7)$$

for $(0 \leq s \leq 2)$ and $k = 0, 1, 2, \ldots$ such that $(s + 3k) < w$. ∎

The above results are also valid for the DFT's $\widetilde{T}[k]$, $\widetilde{C}[k]$, and $\widetilde{G}[k]$ of the binary signals $T[i]$, $C[i]$, and $G[i]$. In the next section, we use Theorems 1–3 to prove the "3-periodicity" property observed in the DNA sequences.

3. 3-PERIODICITY PROPERTY

The 3-periodicity property [2] states that the spectral energy

$$|\widetilde{S}[k]|^2 \triangleq |\widetilde{A}[k]|^2 + |\widetilde{T}[k]|^2 + |\widetilde{C}[k]|^2 + |\widetilde{G}[k]|^2, \quad (8)$$

derived from the DFT's of the four binary signals representing a DNA protein coding region of length N, exhibits a peak at discrete frequency $k = N/3$. No such peak is observed in the spectral energy of noncoding regions. To verify the 3-periodicity property, we report two observations made from protein coding and noncoding regions. Though these observations are verified for a number of eukaryotic organisms, we include results from chromosome III of C. elegans (Accession no. NC_003281) downloaded from the NCBI database [5]. Tables 1 and 2 are constructed from protein coding regions (cumulative length of about 4 million nucleotides), while Tables 3 and 4 are based on noncoding regions (cumulative length of about 9 million nucleotides).

Position	A	C	G	T
0	0.3189	0.3291	0.4545	0.2488
1	0.3642	0.3664	0.2382	0.3523
2	0.3168	0.3044	0.3073	0.3989

Table 1. Fraction of nucleotides at locations $(s = 0, 1, 2)$ in words of length (w = 3) for *coding regions* obtained from C. elegans.

Position	A	C	G	T
0	0.1603	0.1647	0.2271	0.1241
3	0.1587	0.1645	0.2274	0.1247
1	0.1825	0.1824	0.1191	0.1771
4	0.1818	0.1840	0.1192	0.1751
2	0.1583	0.1517	0.1551	0.1994
5	0.1585	0.1527	0.1522	0.1995

Table 2. Same as Table 1 but for locations $(s = 0, 1, \ldots, 5)$ in *coding* DNA words of length (w = 6).

Observation 1: When *coding regions* are parsed in words of length w that is a multiple of 3, $w = 3m$, we observe that:

Part I: The PCF's for nucleotide A

$$\mathcal{C}_w^A(s) \approx \mathcal{C}_w^A(s+3) \approx \ldots \approx \mathcal{C}_w^A(s+3k), \quad (9)$$

for $k = 0, 1, 2, \ldots$ such that $(s + 3k) < w$.

Part II: The PCF's for nucleotide A

$$\mathcal{C}_w^A(s) \neq \mathcal{C}_w^A(s+1) \neq \mathcal{C}_w^A(s+2), \quad (10)$$

for $(0 \leq s \leq w-3)$. ∎

Observation 1 is also valid for the PCF's $\mathcal{C}_w^T(s)$, $\mathcal{C}_w^G(s)$, and $\mathcal{C}_w^C(s)$. Table 1 records the PCF's for the four nucleotides within coding regions parsed in words of length $w = 3$. In each case, the PCF is expressed as a fraction of the total number of the nucleotide of the type being counted in the PCF. In Table 1, we observe that the entries in each column are significantly different from each other. This is in accordance with Observation 1, which states that the PCF's $\mathcal{C}_w^*(0)$, $\mathcal{C}_w^*(1)$, and $\mathcal{C}_w^*(2)$ are not equal for $w = 3$.

Table 2 repeats the earlier experiment performed in Table 1 for protein coding regions except that the DNA segment is parsed in words of length $w = 6$. To show the similarity between positions 0 and 3, 1 and 4, and 2 and 5, we have rearranged the order of the six rows. We observe that the PCF's $\mathcal{C}_w^*(0) = \mathcal{C}_w^*(3)$, $\mathcal{C}_w^*(1) = \mathcal{C}_w^*(4)$ and $\mathcal{C}_w^*(2) = \mathcal{C}_w^*(5)$. However, $\mathcal{C}_w^*(0) \neq \mathcal{C}_w^*(1) \neq \mathcal{C}_w^*(2)$ and $\mathcal{C}_w^*(3) \neq \mathcal{C}_w^*(4) \neq \mathcal{C}_w^*(5)$. This verifies Observation 1.

Observation 2: The PCF's obtained by parsing *noncoding regions* in words of length w are equal. Explicitly, for nucleotide A,

$$\mathcal{C}_w^A(0) \approx \mathcal{C}_w^A(1) \approx \ldots \approx \mathcal{C}_w^A(w-1), \quad (11)$$

for $(3 \leq w \leq N)$. Eq. (11) is also valid for the PCF's for nucleotides C, G, and T. ∎

Tables 3 and 4 record the normalized PCF's for nucleotides A, C, T, and G at locations 0 to $(w-1)$ for DNA words of lengths $w = 3$ and $w = 7$ parsed from noncoding regions obtained from the C. elegans dataset. The PCF's for each nucleotide are equal. Observation 2 is, therefore, confirmed.

Position	A	C	G	T
0	0.3334	0.3331	0.3336	0.3334
1	0.3333	0.3331	0.3339	0.3333
2	0.3333	0.3338	0.3325	0.3333

Table 3. Fraction of nucleotides at locations ($s = 0, 1, 2$) in words of length (w = 3) for *noncoding regions* obtained from C. elegans.

Position	A	C	G	T
0	0.1451	0.1430	0.1420	0.1420
1	0.1422	0.1420	0.1426	0.1453
2	0.1420	0.1429	0.1449	0.1424
3	0.1428	0.1425	0.1428	0.1432
4	0.1423	0.1441	0.1427	0.1422
5	0.1426	0.1426	0.1424	0.1429
6	0.1428	0.1429	0.1425	0.1419

Table 4. Same as Table 3 except for locations ($s = 0, 1, \ldots 7$) in *noncoding* DNA words of length (w = 7).

Explanation: Within a protein coding region, we show that:
(a) The value of the spectral energy $\widetilde{S}[k]$ is not zero at $k = N/3$.
(b) Elsewhere, except for k = 0, the spectral energy $\widetilde{S}[k]$ is zero.

To verify claim (a), we substitute $s = 0$ in Part II of Observation 1. The results show that the PCF's $\mathcal{C}_w^A(0)$, $\mathcal{C}_w^A(1)$, and $\mathcal{C}_w^A(2)$ are not equal. Under such conditions, Corollary 1.1 states that the DFT coefficients $\widetilde{A}[N/3] \neq 0$. By a similar analysis, we can show that the values of the DFT coefficients $\widetilde{T}[N/3]$, $\widetilde{C}[N/3]$, and $\widetilde{G}[N/3]$ are also nonzero. Therefore, the spectral energy $|\widetilde{S}[N/3]|$, defined in (9), has a nonzero value within *coding regions*.

We verify claim (b) next. Coupling Part I of Observation 1 with Corollary 3.1, shows that the DFT $\widetilde{A}[k]$ equals zero for all frequency components $k = N/3m$ subject to N being divisible by $3m$. In cases where N is not divisible by $3m$, we can always zero pad the sequence such that the new length $N_1 > N$ is divisible by $3m$. By approximating $\widetilde{A}[N/3m]$ with $\widetilde{A}[N_1/3m]$ in such cases, we conclude that $\widetilde{A}[k] \approx 0$ for all k except for $k = 0$ and $k = N/3$. The same reasoning is extended to nucleotides T, C, and G proving that $|\widetilde{S}[k]| \approx 0$ for all values of k except for $k = 0$ and $k = N/3$ within protein *coding regions*.

Finally, the value of $\widetilde{S}[N/3]$ within *noncoding regions*, is shown to be 0 by combining Observation 2 with Corollary 1.1. Collectively, the above explanation proves the 3-periodicity property.

4. SPLICING ALGORITHM

Our splicing algorithm exploits the 3-periodicity property. In describing the splicing algorithm, we use two binary signals, $R[i]$ and $W[i]$. Signal $R[i]$ is 1 if the nucleotide at location i in the DNA sequence is either G or C. Similarly, signal $W[i]$ is 1 if the nucleotide at location i is either A or T.
Initialization: Set $\ell = 0$ to indicate the number of window.
Step 1: Apply a rectangular window of length N_1 to select the first N_1 nucleotides of the DNA sequence. In our experiments, the length N_1 of the parsing window is set to 351.
Step 2: For the ℓ'th DNA subsequence obtained from step 1, compute the values of the two binary signals $R[i]$ and $W[i]$.
Step 3: By parsing $R[i]$ and $W[i]$, compute the PCF's $C_3^R(s)$ and $C_3^W(s)$, for $(0 \leq s \leq 2)$. Use Theorem 1 to evaluate the DFT's, $\widetilde{R}[N/3]$ and $\widetilde{W}[N/3]$ of binary signals $R[i]$ and $W[i]$.
It is straightforward to show that the parameter $|\widetilde{R}[N_1/3]|$ is a scaled version of the linear combination $|\beta_r \widetilde{R}[N_1/3]+\beta_w \widetilde{W}[N_1/3]|$, used in [3]. Using parameter $|\widetilde{R}[N_1/3]|$ eliminates the need to compute β_r and β_w.
Step 4: According to the 3-periodicity property, the spectral energy $|\widetilde{R}[N_1/3]|^2$ peaks within protein coding regions. However, the values of these peaks vary significantly across DNA specimens obtained from different organisms. We use

$$\text{SNR}[\ell_1] \triangleq \frac{|\widetilde{R}[N_1/3]|^2}{|\widetilde{R}_{\text{av}}^{(1)}|^2 + |\widetilde{W}_{\text{av}}^{(1)}|^2} = \frac{|\widetilde{R}[N_1/3]|^2}{2|\widetilde{R}_{\text{av}}^{(1)}|^2} \qquad (12)$$

as the classification criteria. Terms $|\widetilde{R}_{\text{av}}^{(1)}|^2$ and $|\widetilde{W}_{\text{av}}^{(1)}|^2$ are the average values of the squared magnitudes of the DFT's $\widetilde{R}[k]$ and $\widetilde{W}[k]$. These values are computed using Theorem 2.
Step 5: The rectangular window is moved forward by 3 nucleotides and the value of ℓ is incremented by 1. Starting from step 2, the procedure is repeated till the entire DNA sequence is scanned.
Step 6: Plot SNR$[\ell]$ as a function of ℓ. The peaks in the plot identify protein coding regions. To automate our classification, we use a threshold value η such that SNR$[\ell] \geq \eta$ corresponds to protein coding regions, while SNR$[\ell] < \eta$ corresponds to noncoding regions. The value of threshold η is computed in section 5.
Bartlett Window: The aforementioned algorithm uses a rectangular window to partition the DNA sequence into subsequences of length N_1. Rectangular windows introduce discontinuities by abruptly truncating the DNA sequences and tend to spread the spectrum of the original DNA sequence in the frequency domain. This causes the power of the DFT to *leak* over into adjacent frequencies. To minimize the leakage, we use the Bartlett window

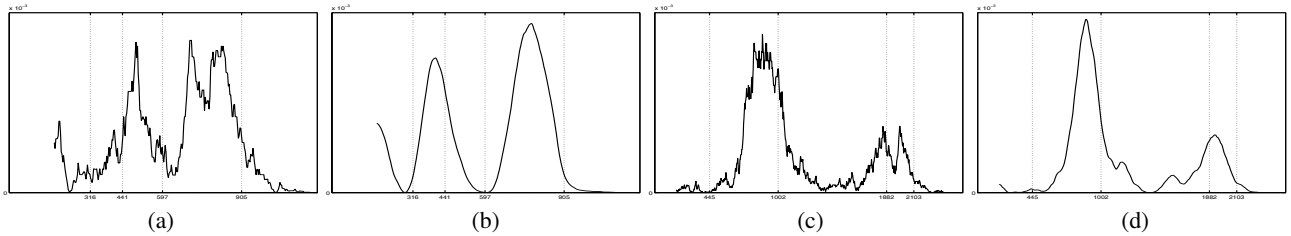
$$w[n] = \begin{cases} \frac{2n}{N_1-1} & 0 \leq n \leq \frac{1}{2}(N_1-1) \\ 2 - \frac{2n}{N_1-1} & \frac{1}{2}(N_1-1) \leq n \leq (N_1-1) \end{cases} \qquad (13)$$

to partition the DNA sequence. The proposed algorithm uses the Bartlett window and involves steps 1–6 except step 3 uses $R_b[n] = w[n]R[n]$ instead of $R[n]$.

5. EXPERIMENTS

The experiments are designed to make three major points. First, we show that the use of the Bartlett window improves the performance of the splicing algorithm by removing the extraneous peaks introduced by abrupt truncations of the rectangular window. Second, we determine the value of threshold η to classify the peaks in SNR, (13), as coding versus noncoding peaks. Third, we quantify the performance of the proposed splicing algorithm.

Fig. 1 illustrates the differences between the magnitudes of $|\widetilde{R}[N/3]|^2$ for the proposed splicing algorithm with rectangular and Bartlett windows. Results from two different DNA sequences, obtained from the C. elegans dataset, are included. Each DNA stretch has two coding regions at positions enclosed within the vertical dotted lines. The algorithm using rectangular window (Figs. 1(a) and 1(c)) is not accurate as it produces peaks at wrong locations. In Figs. 1(a) and 1(c), we also observe more than one peak within a single coding region. On the other hand, the algorithm using



(a) (b) (c) (d)

Fig. 1. Comparison of proposed splicing algorithm with rectangular window (plots (a) and (c)) and the Bartlett window (plots (b) and (d)).

the Bartlett window (Figs. 1(b) and 1(d)) produces a single peak within each coding region. Figs. 1(b) and 1(d) removes the extraneous peaks observed with the rectangular window.

To determine the value of threshold η that discriminates coding regions from noncoding regions, Fig. 2 plots the cumulative distribution of the SNR for both coding and noncoding regions for the three organisms: Chromosome III of C. elegans (Accession number NC_003281); Complete genome of E. coli (Accession number NC_002695); and Complete genome of Pirellula sp. (Accession number NC_005027). The solid curve in Fig. 2 shows the fraction of coding regions with SNR less than the abscissa, while the dotted curve shows the fraction of noncoding regions with SNR greater than the abscissa. We select a threshold value of 1.75 to distinguish between protein coding regions from noncoding regions. Note that by setting $\eta = 1.75$, Fig. 2 bounds the decoding capability of the splicing algorithm to about 85% in correctly identifying protein coding and noncoding regions.

To quantify the performance of the proposed algorithm, we process protein coding regions in chromosome III of C. elegans. The chromosome has a total of 13783268 nucleotides with 8172 coding regions. The minimum length of coding regions is 2 nucleotides, while the maximum length is 7204 nucleotides. Table 5 lists the number of successfully detected coding regions, arranged in order of the increasing length. We observe that the performance of the algorithm is better in detecting coding regions whose lengths are comparable to the window size ($N_1 = 351$). For example, coding regions with lengths equal to 250 nucleotides are correctly identified at a detection rate of about 80%. With larger coding regions, the detection rate improves even further. However, when the length of coding regions is smaller than 150 nucleotides, the DFT based splicing algorithm does not perform as well. In such cases, the data extracted by the window contains both coding and noncoding nucleotides. The "3-periodicity" condition is no longer valid and the proposed DFT based algorithm is relatively inaccurate. In terms of the detection of noncoding regions, the proposed algorithm correctly identifies 7053 or roughly 85% of noncoding regions.

6. SUMMARY

In this paper, we provide a theoretical justification for the 3-periodicity property, which results in a significant peak in the DFT spectrum within protein coding regions of genomic DNA sequences. A new classification criteria, which normalizes the DFT spectrum with average energy present in the DFT coefficients, results in an automated approach in predicting coding regions. In our experiments, the proposed approach provides considerable improvement over other DFT based algorithms that we tested.

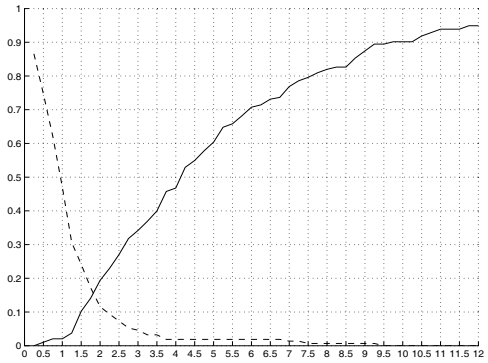


Fig. 2. Cumulative distribution of SNR for both coding and noncoding regions obtained from the C. elegans, E. coli, and Pirellula sp. datasets. The solid curve corresponds to coding regions while the dotted curve corresponds to noncoding regions.

Exons with length	Total Number	Total Detected
$L \geq 100$	7157	3004 (42%)
$L \geq 150$	4177	2513 (60%)
$L \geq 200$	2949	2080 (71%)
$L \geq 250$	2099	1648 (79%)
$L \geq 300$	1534	1270 (83%)
$L \geq 350$	1177	1010 (86%)
$L \geq 400$	919	826 (90%)

Table 4. Number of protein coding regions successfully detected.

7. REFERENCES

[1] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of Probable Genes by Fourier Analysis of Genomic Sequences," *Computer Applications in Biosciences*, vol. 113, 1997, pp. 263-70.

[2] V. R. Chechetkin and A. Y. Turygin, "Size-dependence of Three-periodicity and Long-range Correlations in DNA Sequences," *Physics Letters A*, vol. 199, 1995, pp. 75-80.

[3] D. Anastassiou, "Frequency Domain Analysis of Biomolecular Sequences," *Bioinformatics*, 2000, pp. 1073-81.

[4] D. Kotlar and Y. Lavner, "Gene Prediction by Spectral Rotation Measure: A New Method for Identifying Protein- Coding Regions," *Genome Research*, vol. 13(8), 2003, pp. 1930-37.

[5] National Centre for Biotechnology Information (NCBI). [Online]. Available: `http://www.ncbi.nlm.nih.gov/`.