International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:1, No:5, 2007

A New Predictor of Coding Regions in Genomic Sequences using a Combination of Different Approaches

Aníbal Rodríguez Fuentes, Juan V. Lorenzo Ginori, and Ricardo Grau Ábalo

***Abstract*—**Identifying protein coding regions in DNA sequences is a basic step in the location of genes. Several approaches based on signal processing tools have been applied to solve this problem, trying to achieve more accurate predictions. This paper presents a new predictor that improves the efficacy of three techniques that use the Fourier Transform to predict coding regions, and that could be computed using an algorithm that reduces the computation load. Some ideas about the combination of the predictor with other methods are discussed. ROC curves are used to demonstrate the efficacy of the proposed predictor, based on the computation of 25 DNA sequences from three different organisms.

***Keywords*—**Bioinformatics, Coding region prediction, Computational load reduction, Digital Signal Processing, Fourier Transform,

I. Introduction

THE genomic information is usually represented by sequences of nucleotide symbols in the strands of DNA molecules, by symbolic codons (triplets of nucleotides), or by symbolic sequences of amino acids in the corresponding polypeptide chains. When representing by sequences of nucleotide symbols, the alphabet consists of the letters A, T, C and G; represent adenine, thymine, cytosine, and guanine respectively. The segments of the DNA molecule responsible for protein synthesis are the genes. The regions containing useful information from genes are called exons; in eukaryotes these regions are separated by introns, whereas in prokaryotes they are continuous.

The computational recognition of genes is one of the challenges in the analysis of newly sequenced genomes, and it is a basic step to an understanding of the genome. It is needed to find accurate and fast tools to analyze genomic sequences and annotate genes. A number of methods have been proposed for gene and exon detection, based on distinctive features of protein-coding sequences, and among them many techniques using digital signal processing [1]-[8] and entropic segmentation methods [9], [10] have been used and shown to be successful. All these techniques are mainly based on the period-3 periodicity present in most of genome exons due to the non-uniform distribution of codons in coding regions. The first methods try to detect where coding regions are located inside a large DNA strand; while the second ones try to find the borders between coding and noncoding regions.

In this paper a new coding region predictor based on a combination of other approaches that use the Short Time Fourier Transform (STFT) [1]-[3] is proposed, and that could be computed using an algorithm [11] designed by the authors to improve the computational load. Some ideas about how to combine entropy based methods with the proposed predictor to increase its efficacy are discussed. In order to validate the results of the proposed predictor, ROC curves are used, which show a slight increase of the efficacy of the predictor when compared with the others that use the STFT.

Aníbal Rodríguez Fuentes (e-mail: anibalr@uclv.edu.cu) and Juan V. Lorenzo Ginori (e-mail: juanl@uclv.edu.cu) are with the Center for Studies on Electronics and Information Technologies. http://www.fie.uclv.edu.cu.

Ricardo Grau Ábalo is with the Center for Studies on Informatics (e-mail: rgrau@uclv.edu.cu).

Universidad Central "Marta Abreu" de Las Villas,
Carretera a Camajuaní, km 5 ½, Santa Clara, VC, CP 54830, Cuba.

II. Materials and Methods

In the following paragraphs the new proposed predictor is presented, introducing firstly the technique that was used to convert the genomic information to a numerical sequence. In this work it has been extensively used the fast algorithm previously developed by the authors [11], in order to reduce the computational load associated to the use of the predictor. Later, some ideas about the use of entropy based methods and the possibility of combining them with the proposed predictor are discussed. At the end a brief presentation of ROC curves is made as a validation approach.

A. Obtaining numerical sequences from genomic information

There are several approaches [2], [6]-[8], [12], [13] to convert genomic information in numeric sequences using different representations. The most relevant for the application of signal processing tools is the assignation of complex numbers to each base of the nucleotide sequence, and the indicator sequences. The complex numbers to be assigned are selected according to their mathematical properties, their relation with the bases and the properties of the resulting numeric sequence. Indicator sequences are defined as binary sequences for each base, where 1 at position k indicates the presence of the base at that position, and 0 its absence. For example, for the DNA sequence $x[k] = TACAGAACTTAGC...$ the binary indicator sequences for each base are:

$$\begin{aligned} x_A[k] &= 0101011000\mathbf{1}00\ldots \\ x_T[k] &= 1000000010\mathbf{0}00\ldots \\ x_C[k] &= 0010000100\mathbf{0}01\ldots \\ x_G[k] &= 0000100000\mathbf{1}10\ldots \end{aligned} \quad (1)$$

One of the advantages of using indicator sequences lies in their simplicity, and in the possibility of providing a four-dimensional representation of the frequency spectrum of the character string, when computing the Discrete Fourier Transform of each indicator sequence.

B. Combining approaches based on Discrete Fourier Transform

The Discrete Fourier Transform (DFT) has been used by several authors to predict coding regions in large DNA sequences. As a consequence of the non-uniform distribution of codons in coding regions, a three-periodicity is present in most of genome coding regions, which show a notable peak at the frequency component *N/3* when calculating their DFT [14], [15]. Taking into account the validity of this result the Short Time Fourier Transform has been applied to large DNA sequences to predict coding regions, using a sliding window along the sequence, calculating the Fourier Transform of each subsequence, and taking only the *N/3* frequency component. In [1] Tiwari introduces the Spectral Content Measure (SCM), defined as:

$$S[k] = |X_A[k]|^2 + |X_T[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2 . \quad (2)$$

Here $X_A[k]$, $X_T[k]$, $X_C[k]$ and $X_G[k]$ are the frequency components at k of the Fourier Transform for the indicator sequences. In [2] Anastassiou introduces a new predictor, which he called "Optimized Spectral Content Measure (OSCM)," and that was defined as:

$$W = |aA(s) + tT(s) + cC(s) + gG(s)|^2 , \quad (3)$$

where *A(s)*, *T(s)*, *C(s)* and *G(s)* are the frequency components at *N/3* of the Fourier Transform for the sequence *s*. The values *a*, *t*, *c* and *g* are numerical complex constants obtained as a solution of an optimization problem proposed by the author to maximize the discriminatory capacity between coding and non-coding regions. This predictor demonstrated to be significantly better than the Spectral Content Measure for the sequences analyzed by Anasstasiou.

Using an expression similar to (3), Kotlar proposes in [3] the Spectral Rotation Measure (SRM) (4), where μ_A, μ_T, μ_C and μ_G are the approximated average values, in coding regions, of *arg(A(s))*, *arg(T(s))*, *arg(C(s))*, and *arg(G(s))* respectively; and σ_A, σ_T, σ_C and σ_G are the angular deviation corresponding to *A(s)*, *T(s)*, *C(s)*, and *G(s)*. The magnitude shown in equation (4), proposed by Kotlar, achieves an increase in the magnitude on coding regions when computing this measure.

$$|V|^2 = \left| \frac{e^{-i\mu_A}}{\sigma_A} A(s) + \frac{e^{-i\mu_T}}{\sigma_T} T(s) + \frac{e^{-i\mu_C}}{\sigma_C} C(s) + \frac{e^{-i\mu_G}}{\sigma_G} G(s) \right|^2 . \quad (4)$$

In his paper Kotlar demonstrates on all experimental exons, and for all non-coding strands of length greater that 50 bp from the first 15 Chromosomes of *S. cerevisiae*, that this predictor is more efficiently that the other two exposed methods. The measures were calculated using chromosome 16 of *S. cerevisiae*.

In order to show how these three predictors work, the algorithm which described how they are constructed and used will be presented, supposing that there is a DNA sequence *x[k]* of length *N*. The first step is to define the value *n<N* equal to the length of the sliding window used to move along the sequence *x[k]*. The use of the STFT involves completing the sequence with *2n/3* ceros at the beginning and with *n/3* ceros at the end of *x[k]*. The procedure is to compute the four indicator sequences and the frequency component at *n/3* (*A(s)*, *T(s)*, *C(s)* and *G(s)*) of their Discrete Fourier Transform for each subsequence. Then *S* is calculated as the sum of the square of the modules of the four frequency components at *n/3*, while *W* and *V* previously multiply the frequency components at *n/3* by four fixed complex values and then square the module of the result.

The approach described assures to associate a value to each position of the original sequence *x[k]* by each predictor, and decide if the base at position *k* belongs to a coding region according to the measure of those values, which are obtained when computing the predictors for subsequence *k*.

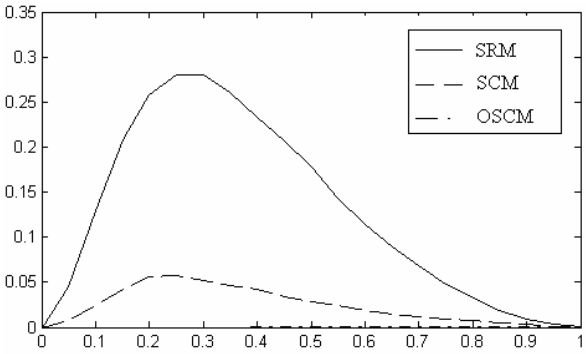

Fig. 1 Distribution of the true positive fraction detected only by each predictor for all possible decision thresholds

Based on these three predictors a new predictor was designed as a linear combination of them. In order to determine the linear combination coefficients, it was first considered the True Positive Fraction (The fraction of bases in the sequence that are predicted as coding regions, when they are truly inside a coding region) detected only by each predictor using a sample composed by 25 DNA sequences from different sizes and belonging to three different organisms. In Fig. 1 it is shown the distribution of these

fractions associated to each predictor for all possible decision thresholds using a sliding window of length 351, which is a typical value according to [2] when computing the DFT.

The distribution of the true positive fraction detected only by the Optimized Spectral Content Measure appeared as hardly noticeable. This result clearly demonstrates that this predictor must be eliminated from the linear combination. The performance of the two remaining predictors when using the ROC analysis are very similar (Fig. 2), and after analyzing the similarity of the mean squared error of each predictor, the following equation was obtained:

$$P[k] = mR[k] + nS[k], \tag{5}$$

where *R[k]* and *S[k]* are the Spectral Rotation Measure and the Spectral Content Measure respectively, and *m* and *n* are the inverse of the maximum value reached by the corresponding predictor when computing the sequence. The objective of the previous multiplication of the measures by the values *m* and *n* is to normalize these measures before adding them. Once *P[k]* is obtained, it can be normalized using the maximum norm, but it could be unnecessary because in practice some of the ways to calculate the threshold are based on the ratio of the average to the maximum of *P[k]*.

C. Reducing the computational load

The use of two predictors and a linear combination of these predictors increases the computational load of the approach. Here was used the algorithm proposed by the authors in a previous paper [11] to calculate the DFT for sliding windows, which reduced at great extent the computational load associated to this task. The general computation load is also reduced, considering that the *N/3* frequency component coefficient of the Discrete Fourier Transform for each window is computed once per indicator sequence.

D. Entropy-based methods

The entropy-based segmentation methods are computational methods used to identify the homogeneous regions based on entropy measures. They are important for DNA-sequence analysis when identifying the borders between coding and noncoding regions. In [9] Bernaola et al. use a 12-symbol alphabet and Jensen-Shannon divergence for finding the borders between coding and noncoding regions in DNA. The 12-symbol alphabet is based on nucleotide statistics inside codons, and in [10] Nicorici and Astola proposed the use of the Jensen-Rényi divergence measure based on the ideas of Bernaola to make the segmentation. Other ideas based on entropic methods were introduced in [16], where the authors propose a new representation of DNA sequences which is able to characterize certain random aspects of a DNA sequences using entropy. One of the advantages of these methods is that they do not need prior training on known databases to process the DNA sequence.

For a DNA sequence a frequency vector can be defined as $F = \{f_{l,j}\}$ where $l \in \{A,T,C,G\}$ and $j \in \{0,1,2\}$; $f_{l,j}$ is defined as the relative number of nucleotides of type l with phase *j*. Based on the definition of the frequency vector, given two sequences of lengths n_1 and n_2 with frequency vectors F_1 and F_2 the Jensen-Shannon divergence is defined as:

$$C(F_1, F_2) = 2\ln 2[NH(F) - n_1 H(F_1) - n_2 H(F_2)], \tag{6}$$

where $N = n_1+n_2$, $F = (n_1/N)F_1 + (n_2/N)F_2$ is the frequency vector of the entire sequence obtained concatenating both sequences, and *H(F)* is the Shannon entropy [17], defined as:

$$H(F) = -\sum_{l,j} f_{l,j} \log_2 f_{l,j} \ . \tag{7}$$

In case of using the Jensen-Rényi divergence, the Rényi entropy [18] is defined as:

$$H_\alpha(F) = \frac{1}{1-\alpha} \log_2 \sum_{l,j} (f_{l,j})^\alpha \ , \tag{8}$$

where $\alpha>0$ and $\alpha\neq1$, and can be considered as a generalization of Shannon entropy, which is a particular case of Rényi Entropy for $\alpha=1$.

In [9], [10] the authors propose a recursive algorithm to make the segmentation. Based on the basic ideas of the algorithm used to compute previous predictors presented, it is proposed the use of some entropic method to increase the efficacy of the combination, using a sliding rectangular window as in the STFT calculation. In this case, the divergence between both halves of each windowed subsequence when moving along the whole sequence was evaluated. To reduce the computation load involved in the calculation of frequency vectors associated to both halves of each subsequence it can be used a mathematical simplification consisting in the calculation of frequency vectors for the first subsequence and then calculate the other ones based on the prior frequency vectors as it is shown below.

Let ${f_{l,j}}^p$ the frequency matrix obtained from one of the two halves of a subsequence at step p, then ${f_{l,j}}^{p+1}$ can be obtained subtracting 1 from element ${f_{B0}}^p$ where index B is the first base of the subsequence at step p, adding 1 to ${f_{B0}}^p$ where index B is the last base of subsequence at step *p+1*, and finally it is only necessary to make the permutations $C_0 \leftrightarrow C_1$ and $C_1 \leftrightarrow C_2$.

Once the whole sequence has been computed, the peaks must correspond to the borders between coding and noncoding regions, and this result can be used to increase the efficacy of the proposed predictor.

Another idea related to entropy based methods is the combination of the representation proposed in [16] and an entropic measure which reaches its highest values in the coding regions. This idea could be integrated directly into the proposed predictor, proposing a linear combination which

involves a predictor based on an entropic measure.

E. Evaluation method: ROC Curves

In order to measure and compare the efficacy of the proposed predictor with the other three approaches, it is proposed the use of the receiver operating characteristic (ROC) curves [19], [20], which provide a global representation of the prediction accuracy.

When a dichotomic test is evaluated (results can be only interpreted as positives or negatives), the sensitivity is defined as the probability that an individual be correctly classified when its real status is the one defined as positive, regarding the condition studied by the test. This is also known as True Positive Fraction (TPF). The specificity is the probability of an individual to be correctly classified when its real status is the one defined as negative. It is the result to subtract the False Positive Fraction (FPF) from 1.

In Table 1 it is shown the statistical procedure to obtain the sensitivity and the specificity, considering the problem of coding region prediction in DNA sequences.

TABLE I
STATISTICAL PROCEDURE TO OBTAIN THE SENSIBILITY AND THE SPECIFICITY IN CODING REGION PREDICTION

	Coding region	**Non-coding region**
Positive Prediction	True Positive (TP)	False Positive (FP)
Negative Prediction	False Negative (FN)	True Negative (TN)

Sensitivity (Ss) = TP/(TP+FN) = TPF
Specificity (Sp) = TN/(TN+FP) = TNF = 1 – FPF

Basically the ROC curve plots for every possible decision threshold, which ranges from zero to the maximum value reached by the predictor, the pair (1-Sp, Ss) when computing the whole sequence and the results are compared with the real values. The closer the ROC curve is to a diagonal, the less useful is the predictor in order to discriminate coding and non-coding region of a DNA sequence. The more the curve moves to the upper left corner on the graph, the better the predictor.

III. RESULTS

For the validation of the experiments all the techniques were applied to 25 genomic sequences with different features and sizes, belonging to three organisms: *S. cerevisiae*, *S. pombe* and *C. elegans*. These sequences can be retrieved directly from the Genbank database, maintained by National Center for Biotechnology Information (NCBI) [21].

Fig. 2 shows the ROC curves associated with each predictor when computing the 25 selected DNA sequences, using a sliding window of size 351. It can be noticed that the graph corresponding to the proposed predictor (solid line) is more effective than the three other approaches. The approximate values measured by the area under the curve for each predictor are: Proposed Predictor: 0.7767, Spectral Content Measure: 0.7352, Optimized Spectral Content Measure: 0.7319 and Spectral Rotation Measure: 0.7351; demonstrating that the Proposed Predictor outperforms the efficacy of the Spectral Content Measure in 6.12%. Using sliding window of lengths 180, 480 and 702, similar results were obtained.

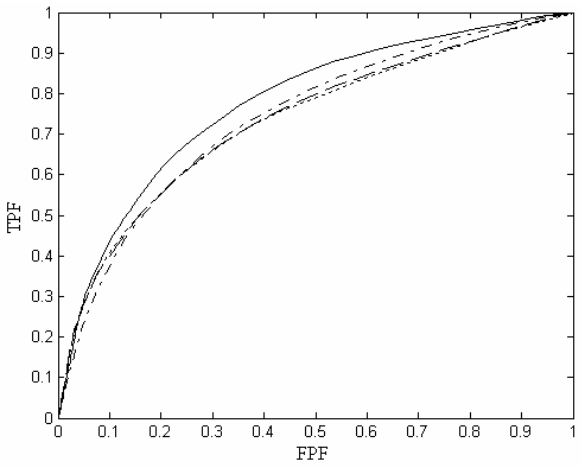

Fig. 2 ROC curves associated to each predictor. Proposed Predictor (Solid line), Spectral Content Measure (dash dot), Optimized Spectral Content Measure (dotted), Spectral Rotation Measure (dashed line)

Fig. 3 shows the graph obtained at using the proposed predictor to a DNA sequence composed by 16680 bp inside chromosome X of *S. cerevisiae*. Real coding regions are represented using the rectangles.

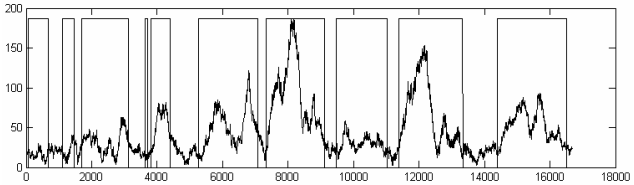

Fig. 3 Application of the proposed approach to a sequence contains 16680 bp inside chromosome X of S. cerevisiae. Rectangles indicate real coding regions.

In order to evaluate the computation time of the proposed predictor when using the fast algorithm described in [11], Table 2 shows the average execution time, in seconds, when computing the Spectral Rotation Measure using the direct (traditional) method and the proposed predictor using the fast algorithm for two sequences with different lengths and using two different window lengths, under the same conditions.

TABLE II
COMPUTATION TIME COMPARISON, IN SECONDS, BETWEEN THE SRM USING THE DIRECT METHOD AND THE PROPOSED PREDITOR USING THE FAST ALGORITHM FOR DIFFERENT DNA STRINGS

DNA stretch length	**8000 bp**		**16680 bp**	
Window Length	**351**	**702**	**351**	**702**
Spectral Rotation Measure using direct method	1.4210	2.1700	2.9030	4.6050
Proposed predictor using the fast algorithm	0.0470	0.0620	0.0930	0.1090

The percentage of time used by the proposed predictor using the fast algorithm is about 3% of the time employed to compute the Spectral Rotation Measure using the direct method.

IV. Conclusions

The prediction of coding regions in large DNA sequences is a basic problem to annotate genes. Digital Signal Processing techniques have been used successfully to solve this problem; however the current tools are still unable to predict all the coding regions present in a DNA sequence.

In this work, a new predictor is proposed based on the linear combination of two other methods that showed good efficacy individually and also on a fast algorithm previously developed by the authors to reduce the computational load. The efficacy of the proposed predictor was evaluated by means of ROC curves, which showed a better performance in coding regions detection when compared to the previous methods. A computation time comparison between the Spectral Rotation Measure using the direct method and the proposed predictor using the fast algorithm demonstrated that even when combining two predictors the computational load does not increase significantly.

References

[1] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," CABIOS, vol. 113, pp. 263-270, 1997.

[2] D. Anastassiou, "Genomic signal processing," IEEE Signal Processing Magazine, vol. 18, pp. 8-20, 2001.

[3] D. Kotlar and Y. Lavner, "Gene Prediction by Spectral Rotation Measure: A New Method for Identifying Protein-Coding Regions," Genome Research, vol. 13, pp. 1930-1937, 2003.

[4] P. P. Vaidyanathan and B.-J. Yoon, "Gene and exon prediction using allpass-based filters," ONR, 2002.

[5] M. Akhtar, E. Ambikairajah, and J. Epps, "Detection of Period-3 Behavior in Genomic Sequences Using Singular Value Decomposition," IEEE-International Conference on Emerging Technologies, pp. 13-17, 2005.

[6] J. A. Berger, S. K. Mitra, and J. Astola, "Power spectrum analysis for DNA sequences," Proceedings of the International Symposium on Signal Processing and its Applications (ISSPA 2003), Paris, France, pp. 29-32, 2003.

[7] G. Dodin, P. vanderghenynst, P. Levoir, C. Cordier, and L. Marcourt, "Fourier and Wavelet Transform Analysis, a Tool for Visualizing Regular Patterns in DNA Sequences," J. Theor. Biol, vol. 206, pp. 323-326, 2000.

[8] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri, "New approaches to genome sequence analysis based on digital signal processing," University of California, 2002.

[9] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán, and H. E. Stanley, "Finding Borders between Coding and Noncoding DNA Regions by an Entropic Segmentation Method," PHYSICAL REVIEW LETTERS, vol. 85, pp. 1342-1345, 2000.

[10] D. Nicorici and J. Astola, "Segmentation of DNA into Coding and Noncoding Regions Based on Recursive Entropic Segmentation and Stop-Codon Statistics," EURASIP Journal on Applied Signal Processing, pp. 81-91, 2004.

[11] A. R. Fuentes, J. V. L. Ginori, and R. G. Ábalo, "Detection of Coding Regions in Large DNA Sequences Using the Short Time Fourier Transform with Reduced Computational Load," In: Martínez-Trinidad, J.F., Carrasco Ochoa, J.A., Kittler, J. (eds.) CIARP 2006. LNCS, vol. 4225, pp. 902-909, 2006.

[12] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals," J. Cell. Mol. Med., vol. 6, pp. 279-303, 2002.

[13] S.-C. Su, C. H. Yeh, and C. J. Kuo, "Structural Analysis of Genomic Sequences with Matched Filtering," IEEE Signal Proccessing Magazine, vol. 3, pp. 2893-2896, 2003.

[14] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, "Periodicity in DNA coding sequences: Implications in gene evolution," J. Theor. Biol., vol. 151, pp. 323-331, 1991.

[15] V. R. Chechetkin and A. Y. Turygin, "Size-dependence of three-periodicity and long-range correlations in DNA sequences," Phys. Lett. A, vol. 199, pp. 75-80, 1995.

[16] J. Gao, Y. Cao, Y. Qi, and J. Hu, "Building Innovative Representations of DNA Sequences to Facilitate Gene Finding," IEEE INTELLIGENT SYSTEMS, pp. 34-39, 2005.

[17] C. E. Shannon, "A Mathematical Theory of Communication," The Bell System Technical Journal, vol. 27, pp. 379–423, 623–656, 1948.

[18] A. Rényi, "On measures of information and entropy," Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability, pp. 547-561, 1960.

[19] J. A. Swets and R. M. Pickett, "Evaluation of diagnostic systems: methods from signal detection theory.," Nueva York: Academic Press, 1982.

[20] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.," Clin Chem, vol. 39, pp. 561-577, 1993.

[21] "GenBank database," NCBI.