Segmentation of DNA into coding and noncoding regions based on inter-STOP symbols distances

Carlos A. C. Bastos, Vera Afreixo, Sara P. Garcia, Armando J. Pinho

Abstract In this study we set to explore the potentialities of the inter-genomic symbols distance for finding the coding regions in DNA sequences. We use the distance between STOP symbols in the DNA sequence and a chi-square statistic to evaluate the nonhomogeneity of the three possible reading frames. The results of this exploratory study suggest that inter-STOP symbols distance has strong ability to discriminate coding regions.

Key words: inter-STOP symbols distance, DNA, coding regions, chi-square.

1 Introduction

It is well known that DNA sequences present a nonhomogenous distribution along the sequence (e.g. coding regions have a tendency to reveal three-base periodicity [1, 2, 3]). There are many published algorithms for coding regions location (e.g. [3, 4, 5, 6, 7]). However, their accuracy needs improvement [5, 6] and there is room for improvement.

In previous work, we explored the inter-nucleotide and inter-dinucleotide distance, i.e., the distance to the first occurrence of the same nucleotide (dinucleotide), to perform a comparative analysis between species [8, 9]. In this work, we extend the concept and explore the inter-STOP symbols distance over different reading frames in the DNA sequence.

Carlos A. C. Bastos, Sara P. Garcia, Armando J. Pinho
Signal Processing Lab, IEETA and Department of Electronics Telecommunications and Informatics, University of Aveiro, 3810-193 Aveiro, Portugal, e-mail: cbastos@ua.pt, spgarcia@ua.pt, ap@ua.pt

Vera Afreixo
Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal, e-mail: vera@ua.pt

It is well known that the distributions of STOP symbols in coding regions and noncoding regions are different. In the correct reading frame of coding regions the STOP symbols occur only at the end [5]. Motivated by the expectation that the distance between STOP symbols has higher values in the correct reading frame than in the other reading frames, we study, in this work, the potentiality of inter-STOP symbols distance distribution for DNA segmentation.

2 Methods

2.1 Inter-STOP symbols distance sequence

The inter-STOP symbols distance sequence is a special case of the inter-trinucleotide distance sequence. It is possible to generate three trinucleotide sequences, one for each reading frame, from a single genomic sequence.

As an illustrative example consider a genomic sequence starting by

AAACAAACTGACACAAAACACTAATAGTTTAAAATAATAATGA ....

Then, the three trinucleotide reading frames (R_1, R_2 and R_3) produce the following trinucleotide sequences,

R_1: $\underbrace{\mathit{AAA}}\underbrace{\mathit{CAA}}\underbrace{\mathit{ACT}}\underbrace{\mathit{GAC}}\underbrace{\mathit{ACA}}\underbrace{\mathit{AAA}}\underbrace{\mathit{CAC}}\underbrace{\mathbf{TAA}}\underbrace{\mathbf{TAG}}\underbrace{\mathit{TTT}}\underbrace{\mathit{AAA}}\underbrace{\mathit{ATA}}\underbrace{\mathit{ATA}}\underbrace{\mathit{ATG}}\mathit{A}\cdots$

R_2: $\mathit{A}\underbrace{\mathit{AAC}}\underbrace{\mathit{AAA}}\underbrace{\mathit{CTG}}\underbrace{\mathit{ACA}}\underbrace{\mathit{CAA}}\underbrace{\mathit{AAC}}\underbrace{\mathit{ACT}}\underbrace{\mathit{AAT}}\underbrace{\mathit{AGT}}\underbrace{\mathit{TTA}}\underbrace{\mathit{AAA}}\underbrace{\mathbf{TAA}}\underbrace{\mathbf{TAA}}\underbrace{\mathbf{TGA}}\cdots$

R_3: $\mathit{AA}\underbrace{\mathit{ACA}}\underbrace{\mathit{AAC}}\underbrace{\mathbf{TGA}}\underbrace{\mathit{CAC}}\underbrace{\mathit{AAA}}\underbrace{\mathit{ACA}}\underbrace{\mathit{CTA}}\underbrace{\mathit{ATA}}\underbrace{\mathit{GTT}}\underbrace{\mathbf{TAA}}\underbrace{\mathit{AAT}}\underbrace{\mathit{AAT}}\underbrace{\mathit{AAT}}\mathit{GA}\cdots$

The distance sequence for each trinucleotide is a vector containing the distances between consecutive occurrences of that trinucleotide. In this work we are interested in the inter-STOP symbols distance, i.e. the distance between consecutive stop symbols: TAA, TAG, TGA. Any of these three symbols signals the end of genes.

As an example, and using the previous nucleotide sequence, we present the beginning of inter-STOP distance sequences for each of the three reading frames:

$$\begin{aligned} d_{R_1}^{STOP} &= (1,\cdots) \\ d_{R_2}^{STOP} &= (1,1,\cdots) \\ d_{R_3}^{STOP} &= (7,\cdots) \end{aligned}$$

2.2 Chi-square statistic

We use the chi-square statistic to measure the lack of homogeneity of the inter-STOP distance distribution between the three possible reading frames. In order to compute the chi-square statistic along the trinucleotide sequences we use a sliding window of fixed length (w) in each frame, and the distances within each window are classified into 2 categories: short distance and long distance. The value used to separate the short and long distances was called cut-off (note: the long distances include the distance corresponding to the cut-off value). We also include an extra category with the number of non stop symbols within the window.

For each DNA sequence we construct contingency tables at each trinucleotide with a window of w trinucleotides. Table 1 shows the structure of the contingency tables.

	Frame 1	Frame 2	Frame 3	Total
non STOP	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
short distance	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
long distance	n_{31}	n_{32}	n_{33}	$n_{1\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	N

Table 1 Contingency table for each window. Note: $n_{\cdot 1} = n_{\cdot 2} = n_{\cdot 3} = w$ and $n_{1j} = w - n_{2j} - n_{3j}$.

The chi-square statistic is given by

$$X^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{\cdot j} n_{i\cdot}}{N}\right)^2}{\frac{n_{\cdot j} n_{i\cdot}}{N}}.$$

When $\frac{n_{\cdot j} n_{i\cdot}}{N} = 0$, we consider $X^2 = 0$, with $i, j \in \{1, 2, 3\}$. This value means that the inter-STOP symbols distributions in the three reading frames are homogenous.

2.3 DNA data

We used genomic data files obtained from the European Bioinformatics Institute site (http://www.ebi.ac.uk/genomes/) for 5 bacteria and the 16 chromosomes of *Saccharomyces cerevisiae* S288c. The bacteria were: Aster yellows witches-broom phytoplasma AYWB; *Borrelia burgdorferi* B31; *Buchnera aphidicola* (Cinara tujafilina); *Candidatus Carsonella* ruddii CE isolate Thao2000; and *Mycoplasma gallisepticum* CA06_2006.052-5-2P.

We extract the genomic sequences and the information of the position of the coding regions from the data files. This information is used to compare with the results of the chi square statistic and to evaluate its discrimination capacity.

We only considered the 5’ to 3’ strand and consequently we did not use the information for the genes on the complement strand.

2.4 *Procedure*

We obtain the chi-square statistic for each symbol of the three reading frames for a sliding window with fixed length (1000 symbols) and we vary the cut-off distance from 50 to 350 symbols. We use the ROC (receiver operating characteristic) curve and compute the area under the ROC curve (AUC) to evaluate the discrimination accuracy of the chi-square statistic. The procedures with higher AUC have better performance. Note that if the AUC is 1 the discrimination is perfect, and if the AUC is 0.5 the discrimination is worthless.

3 Results

Figure 1 shows the position of the coding regions in each of the trinucleotide reading frames and the inter-STOP symbols distances at the positions where the STOP symbols occur. As can be seen from the figure, there is a long inter-STOP distance close to the beginning of most of the contiguous coding regions in one (and only one) of the reading frames.

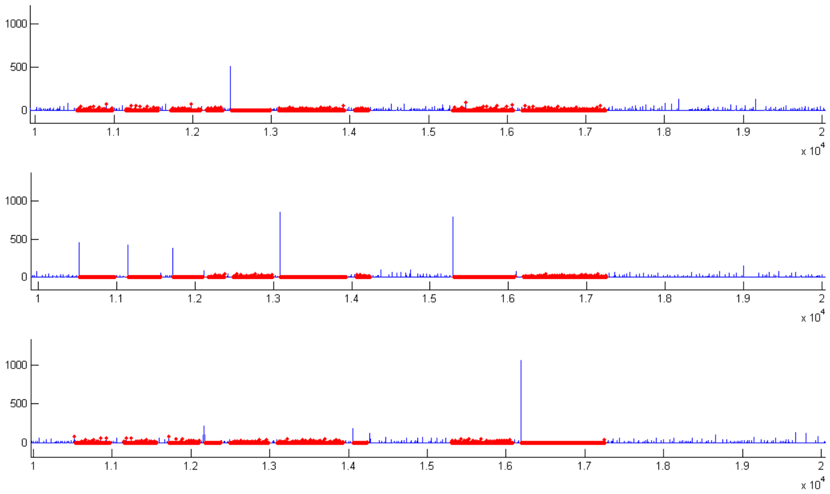

Fig. 1 Plot of the inter-STOP symbols distances for 10000 trinucleotides of the *Saccharomyces cerevisiae* chromosome I in the three frames. The coding regions are marked with thick lines.

We used a sliding window of length 1000 (corresponding to 1000 trinucleotides) which is a reasonable compromise for the genomic sequences considered in this work. In all the genomic sequences used in this study, the percentage of contiguous coding regions whose length ≤ 1000 (trinucleotides) is at least 90%.

As mentioned previously, we varied the cut-off distance that separates the short and long distances. Table 2 shows the AUC for the various cut-off distances studied. In order to reduce the size of the table, we show only the mean and standard deviation (sd) of the AUC for the *Saccharomyces cerevisiae* chromosomes and the bacteria under study.

	cut-off													
Species	50		100		150		200		250		300		350	
	mean	(sd)	mean	(sd)	mean	(sd)	mean	(sd)	mean	(sd)	mean	(sd)	mean	(sd)
Saccharomyces cerevisiae (16 Chr)	62%	3%	67%	2%	76%	2%	79%	1%	80%	1%	80%	1%	79%	1%
Bacteria (5)	67%	6%	79%	5%	82%	4%	82%	2%	80%	2%	78%	5%	75%	8%

Table 2 Area Under the ROC Curve (AUC) to discriminate coding regions using the inter-STOP distance and the chi-square statistic.

The discrimination capacity of the chi-square procedure varies with the cut-off distance; it has a maximum around 250 for the *Saccharomyces cerevisiae* and around 200 for the bacteria.

Figure 2 shows, as an example, the behavior of the chi-square statistic in a section of the *Saccharomyces cerevisiae* chromosome I. The coding regions are highlighted with a thick line. The method seems to have some difficulty in separating coding regions that are very close together. However, the chi-square statistic has non zero values in most of the coding regions showing heterogeneous inter-STOP distance distributions for the three reading frames .

4 Conclusion and future work

In this work we evaluated the possibility of the inter-STOP symbols distance for discriminating coding and noncoding regions in DNA sequences.

We conclude that the inter-STOP symbols distance combined with the chi-square statistic has potential for discriminating coding regions. We believe that this exploratory study may be extended to improve the performance of the procedure presented. In the future, we intend to study the effect of various parameters (e.g., window length, cut-off distance, number of categories in the chi-square statistic) on the discrimination accuracy of the procedure. We also intend to study the association between the coding regions lengths and the parameters of the procedure. The difficulty in separating very close coding regions may be limiting the overall quality of the results. Consequently, we intend to implement a multi-scale procedure based on chi square statistics and the variation of the window length and the cut-off distance. For eukaryotes with genes whose nucleotide number is not a multiple of 3, the algo-

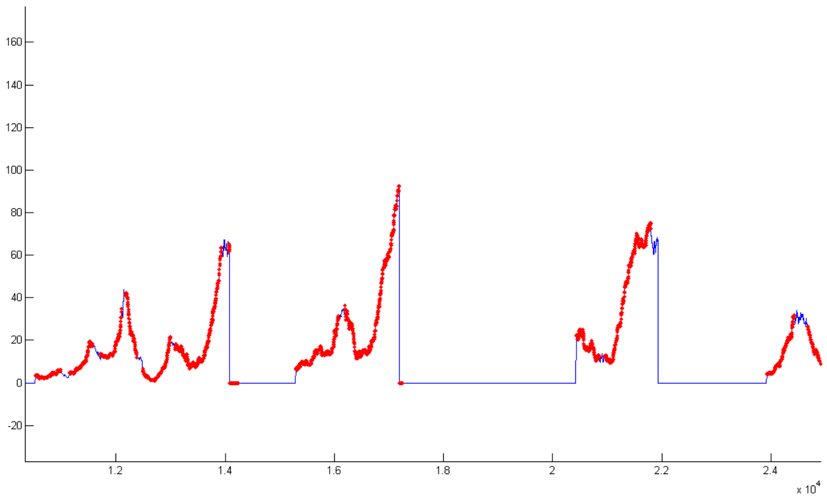

Fig. 2 Plot of chi-square values at each trinucleotide position for part of the *Saccharomyces cerevisiae* chromosome I DNA sequence. The thick lines highlight the positions corresponding to coding regions.

rithm will have to be improved to account for the change of phase in the 3 reading frames.

We expect that the inter-STOP symbols distance will be able to complement existing methods to increase the overall performance of gene finding algorithms.

5 Acknowledgments

This work was supported in part by *FEDER* funds through *COMPETE*– Operational Programme Factors of Competitiveness ("Programa Operacional Factores de Competitividade") and by Portuguese funds through the *Center for Research and Development in Mathematics and Applications* and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project PEst-C/MAT/UI4106/2011 with COMPETE number FCOMP-01-0124-FEDER- 022690.

References

1. Vera M. A. Afreixo, Paulo J. S. G. Ferreira, and Dorabella M. S. Santos. Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*, 14(6):523–530, November 2004.

2. F.E. Frenkel and E.V. Korotkov. Using triplet periodicity of nucleotide sequences for finding potential reading frame shifts in genes. *DNA Research*, 16(2).
3. Omid Abbasi, Ali Rostami, and Ghader Karimian. Identification of exonic regions in dna sequences using cross-correlation and noise suppression by discrete wavelet transform. *BMC Bioinformatics*, 12:430, 2011.
4. Wei Wang and Don H. Johnson. Computing linear transforms of symbolic signals. *IEEE Trans. Signal Processing*, 50(3):628–634, March 2002.
5. Daniel Nicorici and Jaakko Astola. Segmentation of DNA into coding and noncoding regions based on recursive entropic segmentation and stop-codon statistics. *EURASIP Journal on Applied Signal Processing*, 1:81–91, 2004.
6. Suping Deng, Yixiang Shi, Liyun Yuan, Yixue Li, and Guohui Ding. Detecting the borders between coding and non-coding dna regions in prokaryotes based on recursive segmentation and nucleotide doublets statistics. *BMC Genomics*, 13(Suppl 8):S19, 2011.
7. A.A. Tsonis, P. Kumar, J.B. Elsner, and P.A. Tsonis. Wavelet analysis of DNA sequences. *Phys. Rev. E*, 53(2):1828–1834, February 1996.
8. Vera Afreixo, Carlos A. C. Bastos, Armando J. Pinho, Sara P. Garcia, and Paulo J. S. G. Ferreira. Genome analysis with inter-nucleotide distances. *Bioinformatics*, 25(23):3064–3070, December 2009.
9. Carlos A. C. Bastos, Vera Afreixo, Armando J. Pinho, Sara P. Garcia, Joo M. O. S. Rodrigues, and Paulo J. S. G. Ferreira. Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions. *Journal of Integrative Bioinformatics*, 8(3):172, 2011.