Inter-STOP symbol distances for the identification of coding regions

Carlos A. C. Bastos[1,*], Vera Afreixo[2], Sara P. Garcia[1] and Armando J. Pinho[1]

[1]Signal Processing Lab, IEETA and Department of Electronics, Telecommunications and Informatics, University of Aveiro, 3810-193 Aveiro, Portugal.

[2]CIDMA – Center for Research and Development in Mathematics and Applications, IEETA and Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal.

Summary

In this study we explore the potential of inter-STOP symbol distances for finding 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 and the occurrence of one long distance in one of the frames. The results of this exploratory study suggest that inter-STOP symbol distances have strong ability to discriminate coding regions in prokaryotes and simple eukaryotes.

1 Introduction

It is well known that DNA sequences present a nonhomogenous distribution of nucleotides along the sequence, an example being coding regions, which have a tendency to reveal a three-base periodicity [1, 2, 3]. Though algorithms for the discovery of coding regions have been published throughout the years (e.g. [3, 4, 5, 6, 7]), there is still room for improvement, particularly with respect to accuracy [5, 6].

In previous work, we explored inter-nucleotide and -dinucleotide distances. They are defined as the distance to the first occurrence of the same nucleotide or dinucleotide, respectively, and were used to perform a comparative analysis between species [8, 9]. Here, we extend the concept and explore the inter-STOP symbols distance over different reading frames in a DNA sequence. We use the term STOP symbols to designate stop codons irrespective of their function in effectively terminating transcription.

It is well known that the distributions of STOP symbols in coding regions and noncoding regions are different. In the correct reading frame the STOP symbols occur only at the end of coding regions [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 and that there are no small inter-STOP distances within the correct reading frame of a gene, we study, in this work, the potentiality of the inter-STOP symbols distance distribution for DNA segmentation.

*To whom correspondence should be addressed. Email: cbastos@ua.pt

2 Methods

2.1 Sequence of inter-STOP symbol distances

The sequence of inter-STOP symbol distances is a special case of the inter-trinucleotide distance sequence. In a previous work [9] we described and used the inter-dinucleotide distance. The extension to the inter-trinucleotide case is straightforward and starts by obtaining three trinucleotide sequences, one for each reading frame, from a single genomic sequence. Each reading frame starts in a different nucleotide.

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

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

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

The distance sequence for each trinucleotide is a vector containing the distances between consecutive occurrences of that trinucleotide in the primary structure of DNA sequences. In this work we are interested in the inter-STOP symbol distances, i.e. the distance between consecutive stop symbols: TAA, TAG, or TGA. Any of these three symbols signals the end of genes. Note that the distance used here is in terms of trinucleotides and not a physical distance.

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:

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

The inter-STOP distance distribution of a sequence of random and independently placed nucleotides is given by

$$f^{STOP}(k) = P(d^{STOP} = k) = p^{STOP}(1 - p^{STOP})^{k-1}, \qquad k = 1, 2, ...,$$

where $p^{STOP} = p^{TAA} + p^{TAG} + p^{TGA}$. If the four nucleotides have the same occurrence probability we obtain $p^{STOP} = 3/64$ and the expected distance value is $64/3 \approx 21$.

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

	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}	$n_{3\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	N

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 [10]. In order to compute the chi-square statistic along the trinucleotide sequences we use a sliding window of fixed length (w) in each frame. 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 is 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.

The chi-square statistic is given by

$$X^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{\cdot j} n_{i\cdot}}{N}\right)^2}{\frac{n_{\cdot j} n_{i\cdot}}{N}} = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i\cdot}}{3}\right)^2}{\frac{n_{i\cdot}}{3}}.$$

When $n_{i\cdot} = 0$, we consider $X^2 = 0$, with $i \in \{1, 2, 3\}$. This value of the statistic means, in this work, that the inter-STOP symbol distributions in the three reading frames are homogenous.

Note that it is impossible for the nucleotides of one STOP codon to be counted simultaneously as STOP symbols in two different reading frames. So, in each window, the total number of occurrences of STOP symbols is given by $n_{2\cdot} + n_{3\cdot}$.

2.3 DNA data

We used genomic data files obtained from the European Bioinformatics Institute site (`http://www.ebi.ac.uk/genomes/`) for 10 bacteria, 3 archaea 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; *Mycoplasma gallisepticum*, CA06_2006.052-5-2P; *Bacillus subtilis* subsp subtilis str 168; *Chlamydia trachomatis* D UW 3 CX; *Escherichia coli* str K 12 substr MG1655; *Mycoplasma genitalium* G37; *Streptococcus mutans* UA159. The archea were: *Acidianus hospitalis* W1; *Ferroglobus placidus* DSM 10642; *Aeropyrum pernix* K1.

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 the cut-off distance is varied from 100 to 400 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 and to establish the cut-point for prediction purposes. Higher AUC values mean better discrimination performance. Note that if the AUC is 1 the discrimination is perfect and if the AUC is 0.5 the discrimination is worthless. The point of the ROC curve closest to (0,1) will be the "optimal point" in terms of sensitivity, specificity and global accuracy of the prediction.

In order to improve the segmentation of coding regions of DNA, we introduce a new method that better sets the beginning and end of the coding regions. The method relies on the occurrence of two conditions: 1) the existence of a long distance (greater than a certain threshold value), 2) the value of the chi-square statistic being above a certain reference value. If a certain DNA position verifies the two previous conditions then we expect that there will be a start codon near the STOP codon in that reading frame. Then, starting at the STOP codon, we search the next ATG codon (the most frequent initiation codon) and consider it as the beginning of the coding region. As a consequence, the symbols between the STOP codon and the beginning of the coding region are marked as non coding symbols and the symbols between the beginning of the coding region and the next STOP codon in the same reading frame are marked as coding symbols. We denote this improvement by "adjusted method".

3 Results

Figure 1 shows the position of the coding regions in each of the trinucleotide reading frames and the inter-STOP symbol 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. We also observe a similar behavior of the inter-STOP distance distribution in the three reading frames of non coding regions, i.e., the three frames present a similar inter-STOP distance distribution. We also observe that in coding regions two of the reading frames present a similar behavior as those in non coding regions, but there is one reading frame in which there are no inter-STOP distances within the limits of the coding region.

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%. The use of a longer window may penalize the identification of close coding regions and of short coding regions.

As mentioned previously, we varied the cut-off distance that separates the short and long distances. Figure 2 shows the AUC for the various cut-off distances studied. In order to sintetize

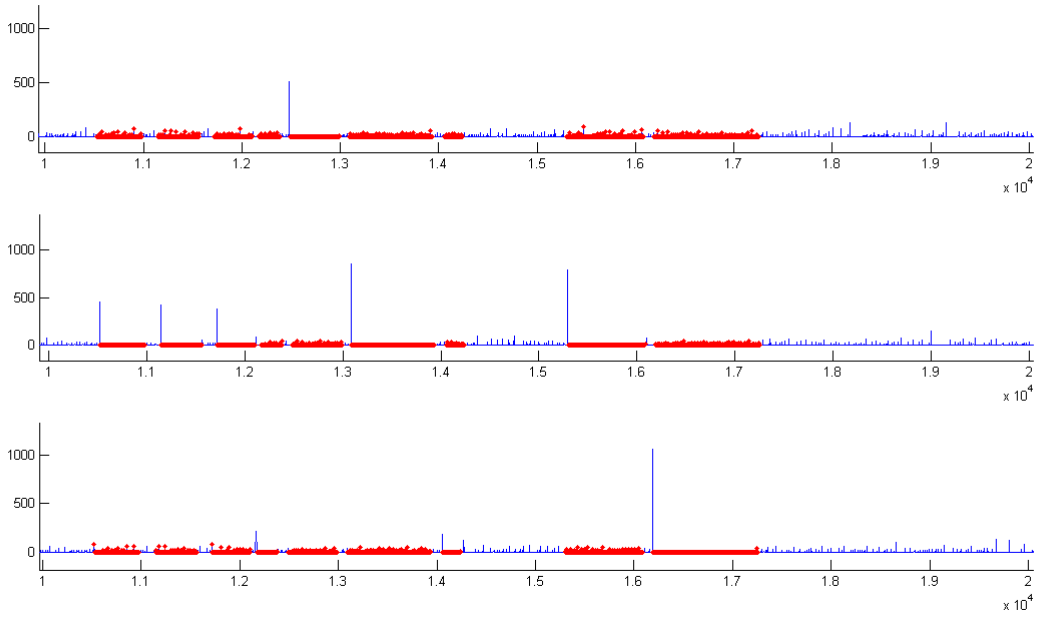

Figure 1: Plot of the inter-STOP symbol distances for 10000 trinucleotides of the *Saccharomyces cerevisiae* chromosome I in the three frames. The coding regions are marked with thick red lines.

Table 2: Description summary of cut-off distance corresponding to best AUC for each group of files.

	# files	mean	sd	min	max
all	29	252	51	150	330
S. cerevisiae	16	271	37	200	320
Bacteria	10	225	65	150	330
Archaea	3	250	30	220	280

the information, we show only the mean (line) and standard deviation (sd, shadow) of the AUC for the groups of DNA files: *Saccharomyces cerevisiae* chromosomes, the bacteria and the archaea under study. The worst AUC results were obtained for the archaea and those for the bacteria and the *Saccharomyces cerevisiae* were somehow similar.

The discrimination capacity of the chi-square procedure varies with the cut-off distance. Table 2 shows a description summary of cut-off distances corresponding to the best AUC for the different groups of data. The global mean cut-off distance corresponding to the best AUC is 252, and the group maxima are: around 271 for the *Saccharomyces cerevisiae*, around 225 for the bacteria and around 250 for the archea. Table 2 presents also the standard deviation and the minimum and maximum cut-off distances for each data group.

Figure 3 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.

We studied the association between the gene length and the best AUC and the association

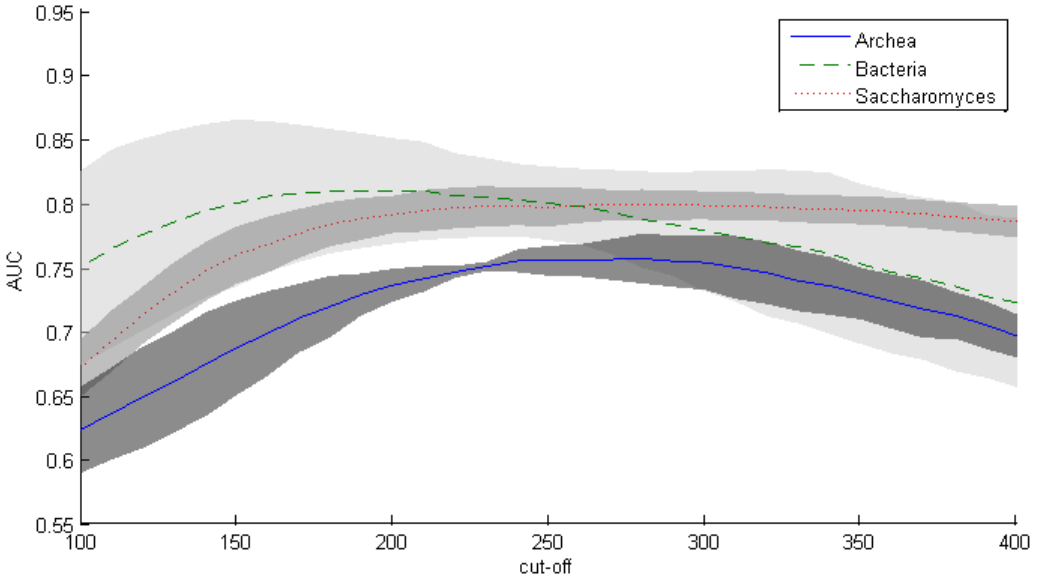

Figure 2: Plot of Area Under the ROC Curve (AUC) to discriminate coding regions using the inter-STOP distance and the chi-square statistic.

Table 3: Sensibility, specificity and accuracy mean values for four cut-point values over all used files, not using the adjusted method.

cut-point	sensitivity	specificity	accuracy
1	87%	72%	78%
5	77%	75 %	76%
10	64%	79%	73%
15	53%	83%	71%

between the cut-off distance and the best AUC. For each of the files used in this work, we obtained almost no association between the median of the gene length and the best AUC. The Pearson's correlation coefficient between gene length and AUC is -0.11. The cut-off distance and the best AUC show also weak negative association. Using all files under study we obtained a correlation coefficient of -0.59. Since there is only weak association we suggest to use a cut-off distance value around the mean cut-off distance (252) for all files (see Table 2).

One method commonly used for establishing the "optimal" cut-point for purposes of prediction is the point on the ROC curve closest to (0,1). Table 3 presents the sensibility, specificity and accuracy values for four of the best cut-point values: 1, 5, 10 and 15. The results obtained for the first 2 cut-points are good but there is room for improvement. Consequently, we introduced the adjusted method that improved the accuracy of the coding region segmentation. Table 4 presents the results obtained with the adjusted method which are better, namely in terms of sensibility and discrimination accuracy than those shown in Table 3. The adjusted method was applied considering a reference value of 252 for separating short and long distances, the chi squared statistic cut-points used were 1, 5, 10 and 15.

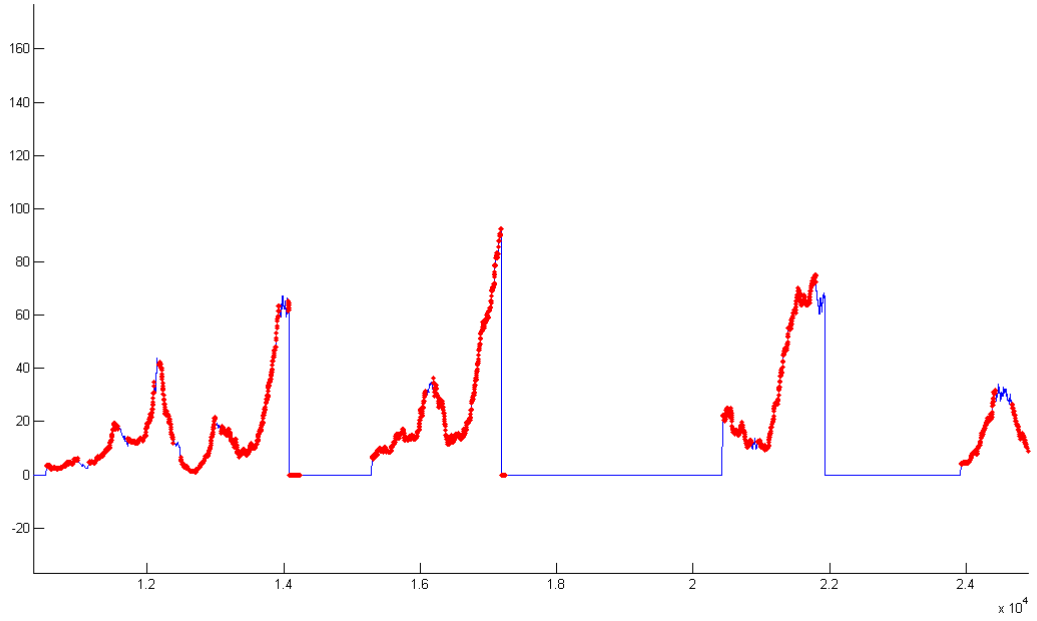

Figure 3: Plot of chi-square values at each trinucleotide position for part of the *Saccharomyces cerevisiae* chromosome I DNA sequence. The thick red lines highlight the positions corresponding to coding regions.

Table 4: Sensibility, specificity and accuracy mean values for four cut-point values over all used files, using the adjusted method.

cut-point	sensitivity	specificity	accuracy
1	91%	74%	81%
5	85%	76 %	80%
10	73%	80%	77%
15	62%	83%	75%

4 Conclusion

In this work, we evaluated the possibility of using the inter-STOP symbol distances for discriminating coding and non coding regions in DNA sequences.

We used the inter-STOP distance and a chi squared statistic to study the influence of various parameters on the AUC. The association between the cut-off, the coding region lengths and the best AUC were found to be weak. This weak association suggested the possibility of using a fixed cut-off distance and the introduction of the adjusted method. The overall algorithm provided good sensibility and accuracy regarding coding region discrimination.

We conclude that the inter-STOP symbol distances combined with the chi-square statistic and the adjusted method has the potential for contributing to the discrimination of coding regions within DNA.

We expect that the inter-STOP symbol distances will be able to complement existing methods to increase the overall performance of gene finding algorithms. We intend to develop methods for

finding introns that will allow the method described in this work to be applied to the genomes of more complex eucaryote species.

Acknowledgements

This work was supported 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 (University of Aveiro) and the Portuguese Foundation for Science and Technology (Fundação para a Ciência e a Tecnologia) within project PEst-C/MAT/UI4106/2011 with COMPETE number FCOMP-01-0124-FEDER-022690 and through the Institute of Electronics and Telematics Engineering of Aveiro within project FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011) and Incentivo/EEI/UI0127/2013.

References

[1] V. M. A. Afreixo, P. J. S. G. Ferreira and D. M. S. Santos. Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*, 14(6):523–530, 2004.

[2] F. Frenkel and E. Korotkov. Using triplet periodicity of nucleotide sequences for finding potential reading frame shifts in genes. *DNA Research*, 16(2):105–114, 2009.

[3] O. Abbasi, A. Rostami and G. Karimian. Identification of exonic regions in DNA sequences using cross-correlation and noise suppression by discrete wavelet transform. *BMC Bioinformatics*, 12(1):430, 2011.

[4] W. Wang and D. H. Johnson. Computing linear transforms of symbolic signals. *IEEE Trans. Signal Processing*, 50(3):628–634, 2002.

[5] 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*, 2004:81–91, 2004.

[6] S. Deng, Y. Shi, L. Yuan, Y. Li and G. 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. Tsonis, P. Kumar, J. Elsner and P. Tsonis. Wavelet analysis of DNA sequences. *Phys. Rev. E*, 53(2):1828–1834, 1996.

[8] V. Afreixo, C. A. C. Bastos, A. J. Pinho, S. P. Garcia and P. J. S. G. Ferreira. Genome analysis with inter-nucleotide distances. *Bioinformatics*, 25(23):3064–3070, 2009.

[9] C. A. C. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. M. O. S. Rodrigues and P. 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.

[10] R. R. Sokal and F. J. Rohlf. *Biometry*. W. H. Freeman and Company, 1994.