Deng et al. BMC Genomics 2012, 13(Suppl 8):S19
http://www.biomedcentral.com/1471-2164/13/S8/S19

BMC Genomics

RESEARCH Open Access

Detecting the borders between coding and non-coding DNA regions in prokaryotes based on recursive segmentation and nucleotide doublets statistics

Suping Deng[1,2], Yixiang Shi[3], Liyun Yuan[3,4], Yixue Li[2,3], Guohui Ding[2*]

From The International Conference on Intelligent Biology and Medicine (ICIBM)
Nashville, TN, USA. 22-24 April 2012

Abstract

Background: Detecting the borders between coding and non-coding regions is an essential step in the genome annotation. And information entropy measures are useful for describing the signals in genome sequence. However, the accuracies of previous methods of finding borders based on entropy segmentation method still need to be improved.

Methods: In this study, we first applied a new recursive entropic segmentation method on DNA sequences to get preliminary significant cuts. A 22-symbol alphabet is used to capture the differential composition of nucleotide doublets and stop codon patterns along three phases in both DNA strands. This process requires no prior training datasets.

Results: Comparing with the previous segmentation methods, the experimental results on three bacteria genomes, *Rickettsia prowazekii, Borrelia burgdorferi* and *E.coli*, show that our approach improves the accuracy for finding the borders between coding and non-coding regions in DNA sequences.

Conclusions: This paper presents a new segmentation method in prokaryotes based on Jensen-Rényi divergence with a 22-symbol alphabet. For three bacteria genomes, comparing to A12_JR method, our method raised the accuracy of finding the borders between protein coding and non-coding regions in DNA sequences.

Background

The prediction of protein coding regions in DNA sequences is a major goal and a long-lasting topic in molecular biology, especially for the genome projects [1-6]. Lots of methods for finding probable borders are based on strong signals between the coding regions and the non-coding ones [7,8]. Staden [9] used the intersection method to detect the borders between coding and non-coding regions. The information entropy measures for signals are useful for identifying the homogeneous regions and evaluating the genomic complexity [10-12]. The entropy-based segmentation methods can be used to identify the borders between coding and non-coding regions [10,13,14]. The Jensen-Shannon divergence measure has provided an impelling tool in doing this [8,9,15]. Bernaola-Galvan et al. presented an entropic segmentation method to search the borders [12]. The accuracy of their results was higher than those obtained with the intersection method [9,12]. The segmentation method presented by Nicorici et al. [8] was based on the Jensen-Rényi divergence measure in both DNA strands. In 2007, Zhang et al. [16] introduced a segmentation method based on a R14 alphabet and the β-*KL* divergence. However, its accuracy is not higher than Nicorici's method [8].

* Correspondence: gwding@sibs.ac.cn
[2]Key lab of systems biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai, 200031, P.R. China
Full list of author information is available at the end of the article

In this study, we constructed a 22-symbol alphabet to represent DNA sequences. Based on the entropy theory, we used recursive segmentation to detect the borders between coding and non-coding DNA regions. Comparing to previous methods, it is shown that our accuracy was well improved.

Materials and methods

The data set

Three tested genomes were downloaded from the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/): *Rickettsia prowazekii* (GenBank: AJ235269), *Borrelia burgdorferi* (GenBank: NC_000948 and AE000783) and *E.coli* (GenBank: NC_009837, NC_008563 and NC_010468).

A22 alphabet

The statistical properties of DNA sequences were commonly used to recognize protein coding regions [11,12,14,17-20]. The statistical properties of doublets of nucleotides (called di-nucleotide for short) in coding regions are also different from those in non-coding regions. This may be used to predict the coding DNA regions. Each di-nucleotide of the DNA sequence is substituted by the symbols from A16 = {A_A, A_C, A_G, A_T, C_A, C_C, C_G, C_T, G_A, G_C, G_G, G_T, T_A, T_C, T_G, T_T} (Table 1).

The distribution of stop codon patterns (SCPs for short) in DNA coding regions differs from that in the non-coding regions [3,21]. It is well known that the SCPs are strong signals in DNA sequences, so that we can effectively use these signals to detect borders between coding and non-coding DNA regions [22]. The SCPs, TGA, TAG and TAA appear in one given DNA strand, and the three SCPs corresponding to TCA, CTA and TTA appear on the reverse strand. In this way, the SCPs statistics on both DNA strands is the same as the statistics of the six codons TAA, TAG, TGA, TCA, CTA, and TTA on a single DNA strand.

In our study, we introduced a 22-symbol alphabet (called A22 for short) that took into account the non-uniform distribution of di-nucleotides and SCPs in both DNA strands (Table 1 and Table 2). Thus the di-nucleotides and the SCPs are substituted by the symbols from A22 = {A_A, A_C, A_G, A_T, C_A, C_C, C_G, C_T, G_A, G_C, G_G, G_T, T_A, T_C, T_G, T_T, S_1, S_2, S_3, S'_1, S'_2, S'_3}(Table 1 and Table 2). The phase of the nucleotide is defined as m = ((n-1)mod 3)+1, where m ∈ {1,2,3}, and n is the position of the nucleotide in the DNA sequence. The phase of a SCP is defined in the same way with the exception that n represents the position of the first nucleotide of the given codon. For example, the DNA sequence ACGTAATC is converted using the A22 alphabet as A_C, C_G, G_T, T_A, S_1, A_A, S'_2, A_T, T_C.

Table 1 Di-nucleotides mapping in 22-symbol alphabet.

Di-nucleotide	Symbol	Di-nucleotide	Symbol
AA	A_A	GA	G_A
AC	A_C	GC	G_C
AG	A_G	GG	G_G
AT	A_T	GT	G_T
CA	C_A	TA	T_A
CC	C_C	TC	T_C
CG	C_G	TG	T_G
CT	C_T	TT	T_T

Table 2 SCPs mapping in 22-symbol alphabet.

Codons	Phase	Symbol
TGA, TAG, or TAA	1	S_1
	2	S_2
	3	S_3
TCA, CTA, or TTA	1	S'_1
	2	S'_2
	3	S'_3

Detecting borders between coding and non-coding DNA regions

In order to partition a DNA sequence, we used the approach proposed by Nicorici et al [8]. and Li [10,13]. A sliding pointer is moving along the sequence. At each position, the pointer divided the sequence into two subsequences and we computed the Jensen-Rényi divergence D_{JR_a}. Then, we found the maximum D_{JR_a} and computed its segmentation strength s (see below). If this segmentation strength s exceeded a given threshold s_0, the position was identified as a significant cut (or a probable border) between coding and non-coding DNA regions. The procedure continued recursively for each of the two resulting subsequences created by each cut until none of the cuts had a segmentation strength level exceeding the s_0. Then such a sequence was segmented at the segmentation strength level s_0.

In this study, the Jensen-Rényi divergence [23,24] is defined as follows:

$$D_{JR_a} = \max_i D_{JR_a}(i) = \left[R_a - \frac{i}{N} R_{a,l} - \frac{N-i}{N} R_{a,r} \right] \quad (1)$$

Where R_a, $R_{a,\,l}$ and $R_{a,\,r}$ are the Jensen-Rényi entropies of the whole, left, and right subsequences, respectively.

Stopping criterion

To decide when the segmentation process has to be stopped, we adopted the method proposed in references [8,25,26] and introduced a segmentation strength, derived empirically, as

$$s = \frac{2 \cdot N \cdot D_{JR_a} - K \cdot log_2 N}{K \cdot log_2 N} \quad (2)$$

The recursive segmentation continues as long as $s \geq s_0$ and the segmented sequence has SCPs in all three phases, where s_0 can be set by the user. K is a constant, which was set as 16 [8].

Actually, the probable borders (the significant cuts) predicted by the recursive segmentation method is generally not the actual borders but are close to them. Since it is well known that the codons at the real borders between coding and non-coding DNA regions must be one kind of start or stop codons, we could use start or stop codons like nuclear acid pattern around the border as border cut. Then, we filter the segment region less than 20 bp. The procedure for finding borders between coding and non-coding DNA regions can be described by the flow chart (Figure 1).

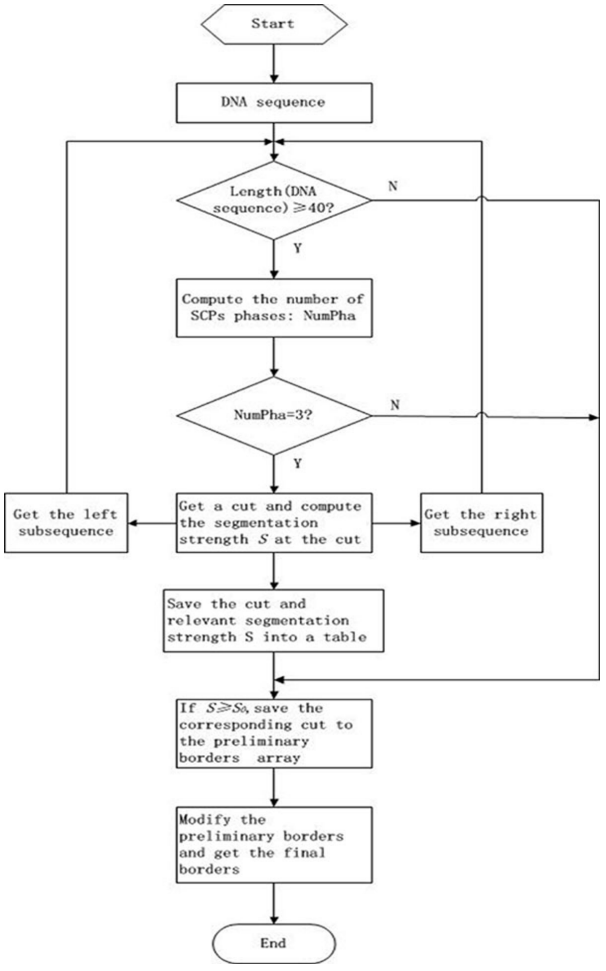

Figure 1 The flow chart of the procedure for finding borders between coding and non-coding DNA regions.

Evaluation

In order to evaluate how well the predicted borders matching the actual borders between coding and non-coding regions, we use the following measure introduced by Bernaola-Galvan et al. [12].

$$CBC = \frac{1}{2}\left[\sum_i \frac{min_j |b_i - c_j|}{N_T} + \sum_j \frac{min_i |b_i - c_j|}{N_T}\right] \quad (3)$$

Where $\{b_i\}$ is the set of all known borders (called KBs for short) between coding and non-coding regions, and $\{c_j\}$ is the set of all predicted borders (called PBs for short), and N_T is the total length of the DNA sequence. The first summation measures the discrepancy between PBs and KBs by adding the distance from each KB to the closest PB, and the second summation performs the same operation, but includes the distance from each PB to the closest KB. Both are required to take into account not only the correctness for the cutting position (CBC would be zero only when the PBs overlap the KBs), but also the difference between the number of PBs and KBs. CBC can be viewed as an average of the error in determining the correct boundaries between coding and non-coding regions, so (1-CBC) is a reasonable measure of the accuracy of the method.

Results and discussion

In Figure 2, we plotted the Jensen-Rényi divergence (α = 0.5 and used in the following experiments as the prediction results have no change when α is adjusted from 0 to 1) with A12 [12] and A22 alphabets along a DNA segment. The DNA segment was randomly chosen from the bacterium genome *Borrelia burgdorferi* and *Rickettsia prowazekii*. In Figure 2(a), the analyzed DNA segment was chosen from bacterium *Rickettsia prowazekii* (AJ235269, 3757-6226 bp). The left part (length 2121 bp) belongs to a coding region and the right part (length 350 bp) belongs to a non-coding region. In Figure 2(b), the analyzed DNA segment was chosen from bacterium *Rickettsia prowazekii* (AJ235269, 10683-11820 bp). The left part (length 1074 bp) belongs to a coding region and the right one belongs to a non-coding region. From Figure 2, the cuts predicted by A22-JR (the method with A22 alphabet, Jensen-Rényi divergence) are closer to the real borders than those by A12-JR.

We also applied the two methods to whole genome respectively. There are multiple coding and non-coding regions in those sequences. The results are summarized in Table 3. The accuracy of A22_JR is better than that of A12_JR for each DNA sequence (p = 0.0015, Table 3).

For visualizing the borders predicted by our proposed method, we plotted the known coding regions in the first 22000 bp of the bacterium genome *Borrelia*

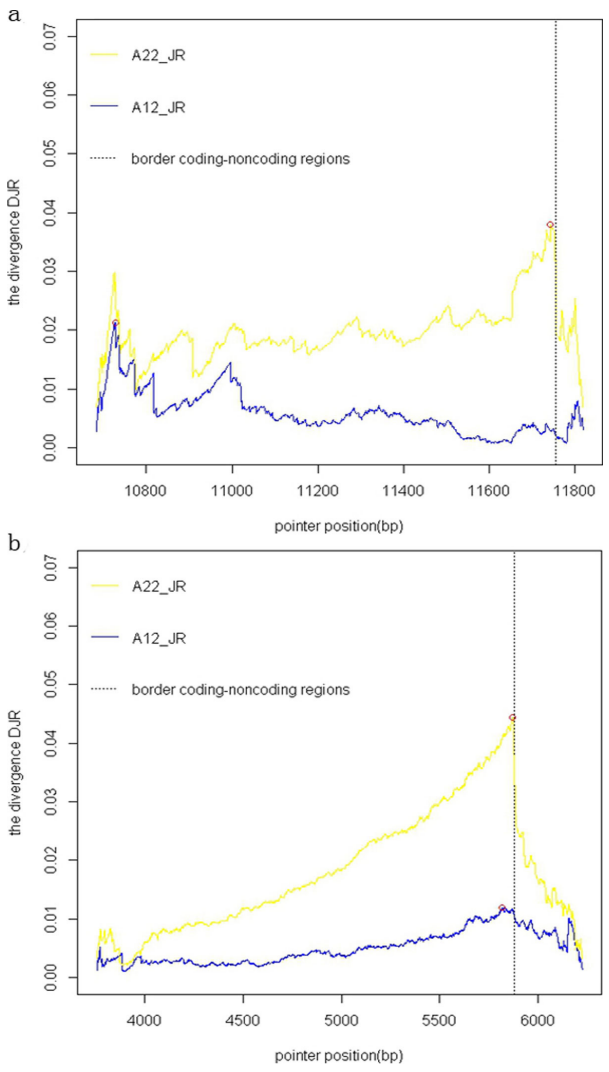

Figure 2 Jensen-Rényi divergence versus cutting position for a DNA sequence. The DNA sequence contains a coding region followed by a non-coding region. The maximum values for the divergences are circled on the graph. (a) The analyzed DNA segment was chosen from bacterium *Rickettsia prowazekii* (AJ235269, 3757-6226 bp). (b) The analyzed DNA segment was chosen from bacterium *Rickettsia prowazekii* (AJ235269, 10683-11820 bp).

Table 3 The maximum accuracy of different methods applied to different data sets.

Organism	GenBank ID	1-CBC(×100%)	
		A12-JR	A22-JR
Rickettsia prowazekii	AJ235269	62.50	63.85
Borrelia burgdorfer	NC_000948	69.18	70.57
	AE000783	70.48	73.18
Escherichia coli	NC_010468	72.26	75.44
	NC_008563	73.39	77.70
	NC_009837	71.45	75.68

burgdorferi (AE000783) and the unmodified predicted borders from our results (Figure 3).

Finally, we described how to choose an appropriate threshold s_0 of segmentation strength. After having gotten the cuts and their corresponding segmentation strength, s_0 ranged from 0.30 to 1.00 stepping by 0.01. For each s_0, the accuracy was computed. From Figure 4, we can find that the accuracy is much higher when s_0 is about -0.50. Thus the appropriate threshold s_0 of segmentation strength can be set as -0.50.

In this study, we introduced a new segmentation method for finding the borders between coding and non-coding regions. It is based on the Jensen-Rényi divergence, a 22-symbol alphabet, and a new stopping criterion. Tested on three bacteria genomes, our method improved the accuracies of the borders detection over the previously reported A12-JR segmentation approach. Most of the existing segmentation algorithms [10,12,13] rely heavily on statistical properties of the coding, non-coding or other interested regions in DNA sequences. Moreover, since the gene-finding systems [24,26,27] use biological knowledge regarding functional sites, together with statistics for finding genes, they require extensive training on known datasets. The recursive segmentation needs no prior training. It should be noted that the value of the segmentation strength threshold s_0 is generally set as -0.50 for bacterium and may be adjusted accordingly in different species. For a new unknown genomic sequence, the optimal threshold s_0 of segmentation strength or significance level can be computed using the genomic sequence of the same or the closest organism.

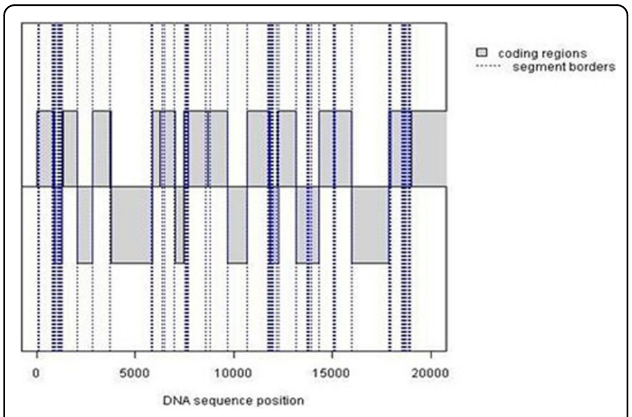

Figure 3 Comparison between the known coding regions and the predicted borders of a DNA sequence. The known coding regions are gray regions with solid lines as borders. The predicted borders (vertical dotted lines) is obtained through recursive segmentation using A22_JR (a = 0.5). The DNA sequence is from bacteria Rickettsia prowazekii and the borders. The coding regions shown downwards are on the opposite DNA strand.

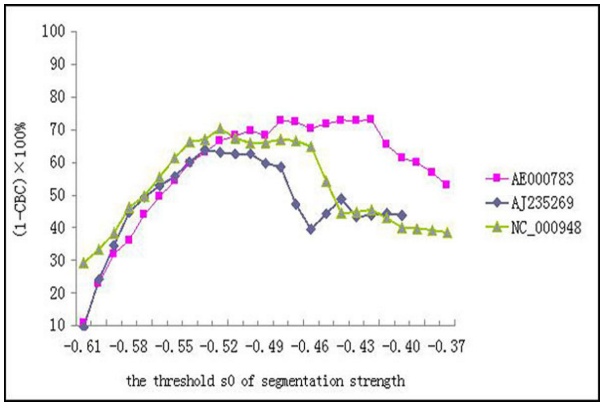

Figure 4 Accuracies of recursive segmentation for different thresholds of segmentation strength. The DNA sequences are from the genomes of *Rickettsia prowazekiiand* and *Borrelia burgdorferi* (the first 30000 bp).

Conclusions

The borders between coding and non-coding regions are found more efficient and accurate will raise the vital effect for DNA sequences annotation. This paper presents a new segmentation method based on Jensen-Rényi divergence with a 22-symbol alphabet, new stopping criterion for finding the borders between coding and non-coding DNA regions in prokaryotes. For three bacteria genomes, comparing to A12_JR method, our method raised the accuracy of finding the borders between coding and non-coding regions in DNA sequences. The success comes from the utilization of the di-nucleotides and SCPs statistics in all three phases along the DNA sequence, and use of Jensen-Rényi divergence.

Acknowledgements

This research was supported by grants from from National High-Tech R&D Program (863) (2009AA02Z304, 2012AA020404), State key basic research program (973) (2006CB910705, 2010CB529206, 2011CBA00801), Research Program of CAS (KSCX2-YW-R-112, KSCX2-YW-R-190, 2011KIP204), National Natural Science Foundation of China (30900272) and SA-SIBS Scholarship Program.

This article has been published as part of *BMC Genomics* Volume 13 Supplement 8, 2012: Proceedings of The International Conference on Intelligent Biology and Medicine (ICIBM): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S8.

Author details

[1]School of Life Sciences and Technology, Tongji University, 1239 Siping Road, Shanghai, 200092, P.R. China. [2]Key lab of systems biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai, 200031, P.R.China. [3]Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai, 200235, P. R. China. [4]Shanghai Jiaotong University, Shanghai, 200240, P. R. China.

Competing interests

The authors declare that they have no competing interests.

Published: 17 December 2012

References

1. Li W: **The complexity of DNA.** *Complexity* 1997, **3**:33-37.
2. Zhang CT, W J: **Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on Z curve.** *Nucleic Acids Res* 2000, **28**:2804-2814.
3. Stanke M, W S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics* 2003, **19**:ii215-ii225.
4. Haimovich AD, B B, Ramaswamy R, Welsh WJ: **Wavelet analysis of DNA walks.** *J Comput Biol* 2006, **13**:1289-1298.
5. Orlov YL, T R, Abnizova I: **Statistical measures of the structure of genomic sequences: entropy, complexity and position information.** *JBioinform Comput Biol* 2006, **4**:523-526.
6. TeBoekhorst R, A I, Nehaniv C: **Discriminating coding, non-coding and regulatory regions using rescaled range and detrended fluctuation analysis.** *BioSystems* 2008, **91**:183-194.
7. Bennetzen JL, H BD: **Codon selection in yeast.** *J Biol Chem* 1982, **257**:3026-3031.
8. Nicorici Daniel, A J: **Segmentation of DNA into Coding and Noncoding Regions Based on Recursive Entropic Segmentation and Stop-Codon Statistics.** *EURASIP Journal on Applied Signal Processing* 2004, **1**:81-91.
9. Staden R: **Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes.** *Nucleic Acid Res* 1984, **12**:551-567.
10. Li Wentian , P B-G, Haghighi F, Grosse I: **Applications of recursive segmentation to the analysis of DNA sequences.** *Computers and Chemistry* 2002, **26(5)**:491-510.
11. Nicorici D, B JA, Astola J, Mitra SK: **Finding borders between coding and noncoding DNA regions using recursive segmentation and statistics of stop codons.** *Proceedings of the 2003 Finnish Signal Processing Symposium: May 2003 2003; Tampere, Finland* 231-235.
12. Bernaola-Galvan P, G I, Carpena P, Oliver JL, Roman-Roldan R, Stanley HE: **Finding borders between coding and noncoding DNA regions by an entropic segmentation method.** *Phys Rev Lett* 2000, **85(6)**:1342-1345.
13. Li W: **New stopping criteria for segmenting DNA sequences.** *PhysRevLett* 2001, **86(25)**:5815-5818.
14. Bernaola-Galvan P, R-R R, Oliver JL: **Compositional segmentation and long-range fractal correlations in DNA sequences.** *PhysRevE* 1996, **53(5)**:5181-5189.
15. Ramaswamy R: **Prediction of probable genes by Fourier analysis of genomic sequences.** *CABIOS* 1997, **13(3)**:263-270.
16. Zhang Jingxiang , X Z: **Finding Borders Between Coding and Noncoding DNA Regions By β-KL Divergence.** *ICBBE 2007* 2007, **77**:286-289.
17. Fickett JW: **Recognition of protein coding regions in DNA sequences.** *Nucleic Acids Research* 1982, **10(17)**:5303-5318.
18. Staden R, M AD: **Codon preference and its use in identifying protein coding regions in long DNA sequences.** *Nucleic Acids Research* 1982, **10**:141-156.
19. Shepherd JCW: **Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification.** *Proc Natl Acad Sci* 1981, **78**:1596-1600.
20. Herzel H, G I: **Measuring correlations in symbolic sequences.** *Physica A* 1995, **216**:518-542.
21. Grantham R, G C, Gouy M, Jacobzone M, Mercier R: **Codon catalog usage is a genome strategy modulated for gene expressivity.** *Nucleic Acids Res* 1981, **9(1)**:R43-74.
22. Grosse I, H H, Buldyrev SV, Stanley HE: **Species independence of mutual information in coding and noncoding DNA.** *Phys Rev E* 2000, **61(5)**:5624-5629.
23. Voss RF: **Evolution of long-range fractal correlations and 1/f noise in DNA base sequences.** *Phys Rev Lett* 1992, **68(1)**:3805-3808.
24. Nicorici D, A J, Tabus I: **Computational identification of exons in DNA with a hidden Markov model.** *Work shop on Genomic Signal Processing and Statistics* Raleigh, NC, USA; 2002.
25. He Y, H AB, Krim H: **A generalized divergence measure for robust image registration.** *IEEE Trans Signal Process* 2003, **51(5)**:1211-1220.
26. Henderson J, S S, Fasman KH: **Finding genes in DNA with a hidden Markov model.** *Journal of Computational Biology* 1997, **4(2)**:127-141.
27. Salzberg S, D A, Fasman K, Henderson J: **A decision tree system for finding genes in DNA.** *Journal of Computational Biology* 1998, **5(4)**:667-680.

doi:10.1186/1471-2164-13-S8-S19
Cite this article as: Deng *et al.*: Detecting the borders between coding and non-coding DNA regions in prokaryotes based on recursive segmentation and nucleotide doublets statistics. *BMC Genomics* 2012 **13** (Suppl 8):S19.