RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS
LABORATORY OF THEORETICAL GENETICS

PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON BIOINFORMATICS OF GENOME REGULATION AND STRUCTURE

Volume 2

BGRS' 2004
Novosibirsk, Russia
July 25 - 30, 2004

IC&G, Novosibirsk, 2004

SEVEN CLUSTERS AND UNSUPERVISED GENE PREDICTION

Gorban A.N.[1,2], *Popova T.G.**[1], *Zinovyev A.Yu.*[3]

[1] Institute of Computational Modeling SB RAS, Krasnoyarsk, Russia; [2] Institute of polymer physics, ETH, Zurich, Switzerland; [3] Institut des Hautes Études Scientifiques, Bures-sur-Yvette, France
* Corresponding author: e-mail: tanya@icm.krasn.ru

Keywords: *triplet frequency, visualization, gene recognition, unsupervised learning*

Summary

Motivation: The effectiveness of most unsupervised gene-detection algorithms follows from a cluster structure in oligomer distributions. Existence of this structure is implicitly known but it was never visualized and studied in terms of data exploration strategies. Visual representation of the structure allows deeper understanding of its properties and can serve to display and analyze characteristics of existing gene-finders.

Results: The cluster structure of genome fragments distribution in the space of their triplet frequencies was revealed by pure data exploration strategy. Several complete genomic sequences were analyzed, using visualization of distribution of 64-dimensional vectors of triplet frequencies in a sliding window. The structure of distribution was found to consist of seven clusters, corresponding to protein-coding genome fragments in three possible phases in each of the two complementary strands and to the non-coding regions with high accuracy. The self-training technique for automated gene recognition both in entire genomes and in unassembled ones is proposed.

Availability: http://www.ihes.fr/~zinovyev/bullet/

Introduction

In Fig. 1a, b one can see two projections of the 3D data scatters. Each point represents a genome fragment clipped by the sliding window and presented by its non-overlapping triplet frequencies. One see the seven cluster structure of the distribution. The central cluster (Fig. 1a, c) corresponds to non-coding genome fragments while side clusters correspond to protein-coding fragments related to three possible phases (reading frames at translation) in each of the two complementary strands with higher than 90 % accuracy on nucleotide level.

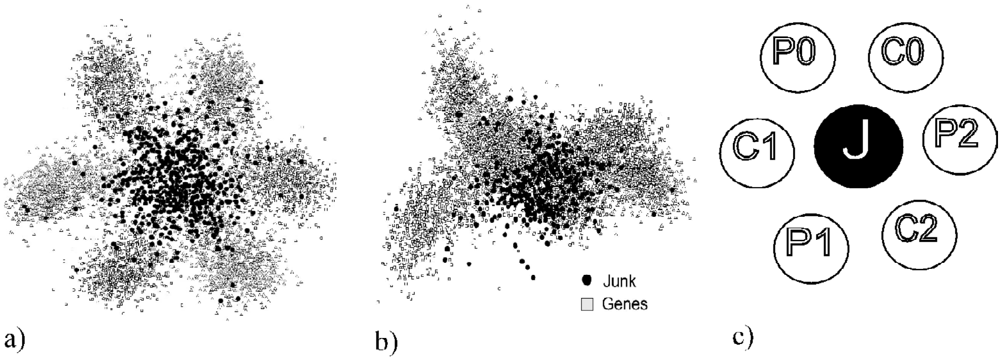


Fig. 1. Visualisation of *C.crescentus* (GenBank NC_002696) genome fragments distribution (a, b) and general seven clusters structure (c).

Seven clusters structure of genome fragments distribution plays important role in ability of modern gene-finders for unsupervised (and, to lesser extent, also for supervised) learning in prokaryotic genomes (Audic *et al.*, 1998; Baldi, 2000). Actually existence of the structure makes the prokaryotic gene-finding so efficient. While using seven hidden states for hidden Markov model in gene-prediction was introduced long ago (see, for example, Borodovsky *et al.*, 1993) and being widely exploited so far, this structure was never visually presented and analyzed by pure data exploratory strategy.

Here we introduce (1) the seven cluster structure of genome fragments distribution in the space of non-overlapping triplet frequencies and as illustration a simple unsupervised procedure for detecting coding regions; (2) some features of coding regions and gene-finders that become evident after the seven clusters structure was revealed.

Model

Seven clusters structure for compact genomes. Consider we have DNA sequence – genome fragments or complete genome. It is converted into the set of 64-dimensional vectors of triplet frequencies as follows.

A sliding window of the length W (W is to be about average exon size – 200–400 in our studies) and centered at position i is characterized by non-overlapping triplet frequencies calculated throughout the window: starting from the first nucleotide and up to the end. So, each data point vector $X_i=\{x_{is}\}$ corresponds to i-th window and has 64 coordinates which are frequencies of all possible triplets s = 1,…,64.

The standard centering and normalization on unit dispersion procedure is then applied, i.e., $\tilde{x}_{is} = \frac{x_{is} - m_s}{\sigma_s}$, where m_s and σ_s is the mean value and standard deviation of the s-th triplet frequency in the dataset.

Visualization of this 64-dimensional dataset in projection onto the 3-dimensional linear manifold spanned by the first three principal vectors of the distribution gives the well-detected seven clusters structure (Fig. 1).

Analysis of the distribution shows that the central cluster (Fig. 1a) is formed by the points X_i taken from the non-coding genome regions while side clusters are formed by the points of protein-coding regions. More specifically (see Fig. 1c), cluster P0 corresponds to the case when countered triplets are codons of the direct strand genes, C0 – codons of complementary strand genes, but in complementary translation and read from back to front (“shadow” genes, because only direct strand is considered), clusters P1, P2, C1, C2 contain points from coding regions too but read with 1 and 2 nucleotides shift relative to gene start.

Simple unsupervised procedure for detecting coding regions. Scanning the sequence with the sliding window step divisible by three and applying some clustering algorithm (K-Means clustering in the 64-dimensional space in our case) one obtains homogeneous with respect to the cluster label regions within the sequence. The J cluster gives non-coding regions but other clusters mark out protein coding regions. In more detail the gene predicting algorithm can be found in (Gorban *et al.*, 2003).

To evaluate the ability of the procedure to detect genes the base-level sensitivity and specificity were calculated, which are commonly used in this case:

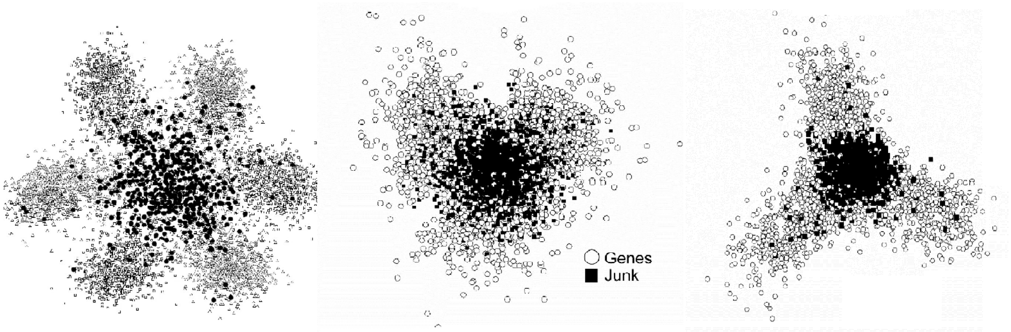
$$Sn = \frac{TP}{TP + FN}, \quad Sp = \frac{TP}{TP + FP},$$

where TP (true positives) is the number of coding bases predicted to be coding; FP (false positives) is the number of non-coding bases predicted to be coding, and FN (false negatives) is the number of coding bases predicted to be non-coding.

Results and Discussion

In this section we consider some results and observations obtained while analysing the seven cluster structure of some genomes and comparing it to well known facts and gene-finders.

The seven cluster structure and estimation of gene prediction accuracy. The cluster structure for some genomes is presented in Figure 2. The distribution in Fig. 2a shows very clear separation on seven clusters; no surprise that in this case unsupervised gene-prediction gives both high specificity and sensitivity.



a) Sn = 0.93; Sp = 0.97 b) Sn = 0.89; Sp = 0.91 c) Sn = 0.82; Sp = 0.93

Fig. 2. a) *C. crescentus* (GenBank NC_002696); b) *S. cerevisiae* chr.IV (GenBank NC_001136); c) *P. wickerhamii* (GenBank NC_001613).

The distribution of triples in *S. cerevisiae*, chr. IV (Fig. 2b) forms seven clusters as well; though they are not clearly seen on 2D-pictures, because two "phase triangles" P0-P1-P2 and C0-C1-C2, projected into the principal subspace are positioned on two parallel planes, perpendicular to the projection plane. Nevertheless simple clustering algorithm yields good prediction. The situation is worse in case of *P. wickerhhamii* mitochondrion genome. In this case distributions of triplets in the direct and reverse strands indeed overlap: P0 cluster overlaps with C1 cluster and so on. One can predict in this case that gene recognition procedures will often mix genes in the direct and reverse strands, though ORF-strategies can probably resolve this conflict.

So, visualization of datasets is useful to evaluate how reliable gene prediction could be.

Gene identification accuracy of our CLUSTER method and GLIMMER gene-finder. Choosing GLIMMER (version 2.02) (Salzberg *et al*., 1998; Delcher *et al*., 1999) for comparison was dictated by our desire to take a gene-predictor that uses no additional learning information, except one that can be extracted from the genetic sequence itself.

Table.

Genome	CLUSTER		GLIMMER	
	Sn	Sp	Sn	Sp
Helicobacter pylori	0.94	**0.95**	**0.96**	0.78
Haemophilus influenza	0.93	**0.88**	**0.96**	0.84
Escherichia coli	0.91	**0.87**	**0.96**	0.76
Bacillus subtilis	0.89	**0.95**	**0.97**	0.79
Caulobacter crescentus	0.89	**0.76**	**0.94**	0.60

The results of this comparison are shown in the Table. As one can see, the sensitivity of our method is lower in all cases, on the other hand specificity of our method is significantly better.

We have analysed why the GLIMMER produced a lot of false-positive predictions using visualization tool and seven cluster structure. Our analysis shows that 80 % of false positives for *C. crescentus* in the 64-dimensional space of triplet frequencies are closer to the C0 centroid, while only 2 % of true-positive predictions for *C. crescentus* are close to the C0-centroid. Such discrepancy cannot be explained simply by "presence of unknown genes" but it is due to some effect of this HMM-based predictor, which often takes "shadow" genes as positive predictions.

We have analysed why our clustering method produced more false-negatives than GLIMMER did for *E. coli* genome. It became clear from genome annotation: a half of missed genes are noted as predicted only by computational methods, other significant groups are ribosomal genes and transposases. It is known that ribosomal genes, some other highly expressed genes as well as horizontally transfered genes can have rather different (with respect to the average) codon usage, that is why our simple procedure based on triplet frequencies failed to predict them.

Some observations and conclusions

1. Our study shows relatively high performance of using only triplets for gene prediction in compact genomes. Using hexamer frequencies (that is common practice in modern gene-finders) can be more sensitive, but also can lead to certain undesirable "overfitting" effects and worse specificity.
2. The structure of codon usage over all genes in a genome is known to be inhomogeneous, especially in fast-growing organisms as *E. coli* and *B. subtilis*, however, the cluster structure shows a less within group dispersion than between groups dispersion. Thus the gene-prediction is possible even without preliminary genes classification.
3. Frequency normalization plays an important role in cluster structure formation. It indicates the importance in distinguishing coding and non-coding regions of those codons that may not have high frequency values but considerably change their frequencies after reading frame shift.
4. The proposed procedure for detecting genes is fully automated and requires no prior learning on known genes. The method can be applied for the rough annotation of unassembled genomes, since it does not require preliminary extraction of ORFs.
5. The cluster structure may be very useful in solving the problem of choosing a "good" learning dataset as it is not very well defined yet (see, for example (Mathe *et al.*, 2002)).

References

Audic S., Claverie J.-M. Self-identification of protein-coding regions in microbial genomes // Proc. Natl Acad. Sci. USA. 1998. V. 95. P. 10026–10031.

Baldi P. On the convergence of a clustering algorithm for protein-coding regions in microbial genomes // Bioinformatics. 2000.V. 16. P. 367–371.

Borodovsky M., McIninch J. GENMARK: parallel gene recognition for both DNA strands // Comp. Chem. 1993. V. 17. P. 123–133.

Gorban A.N., Zinovyev A.Yu., Popova T.G. Seven clusters in genomic triplet distributions // In Silico Biology. 2003. V. 3. 0039. </isb/2003/03/0039>

Salzberg S.L., Delcher A.L., Kasif S., White O. Microbial gene identification using interpolated Markov Models // Nuc. Acids Res. 1998. V. 26(2). P. 544–548.

Delcher A.L., Harmon D., Kasif S., White O., Salzberg, S.L. Improved microbial gene identification with GLIMMER // Nuc. Acids Res. 1999. V. 27(23). P. 4636–4641.

Mathe C., Sagot M.F., Schiex T., Rouze P. Current methods of gene prediction, their strengths and weaknesses // Nucleic Acids Res. 2002. V. 30(19). P. 4103–4117.

Подготовлено к печати
в редакционно-издательском отделе ИЦиГ СО РАН

Подписано к печати 23.06.2004
Формат бумаги 70x108/16. Усл.-печ.л. 38,22. Уч.-изд.л. 32,2
Тираж 250. Заказ 335
Отпечатано в типографии СО РАН
630090, Новосибирск, Морской просп., 2