Open Sys. & Information Dyn. **10**: 321–333, 2003

Self-Organizing Approach for Automated Gene Identification

Andrey Yu. Zinovyev

Institut des Hautes Etudes Scientifiques, France
e-mail: zinovyev@ihes.fr

Alexander N. Gorban and Tatyana G. Popova

Institute of Computational Modeling of Russian Academy of Sciences
Akademgorodok, Krasnoyarsk, 660036 Russia
e-mail: gorban@icm.krasn.ru and tanya@icm.krasn.ru

(Received: December 6, 2002)

Abstract. Self-training technique for automated gene recognition both in entire genomes and in unassembled ones is proposed. It is based on a simple measure (namely, the vector of frequencies of non-overlapping triplets in sliding window), and needs neither predetermined information, nor preliminary learning. The sliding window length is the only one tuning parameter. It should be chosen close to the average exon length typical to the DNA text under investigation. An essential feature of the technique proposed is preliminary visualization of the set of vectors in the subspace of the first three principal components. It was shown, the distribution of DNA sites has the bullet-like structure with one central cluster (corresponding to non-coding sites) and three or six flank ones (corresponding to protein-coding sites). The bullet-like structure itself revealed in the distribution seems to be very interesting illustration of triplet usage in DNA sequence. The method was examined on several genomes (mitochondrion of *P.wickerhamii*, bacteria *C.crescentus* and primitive eukaryot *S.cerevisiae*). The percentage of truly predicted nucleotides exceeds 90%.

The paper is devoted to one of the most fundamental problem in DNA analysis, that is the problem of computational (automated) identification of coding (exons) and non-coding (junk and introns) regions in DNA sequence.

Current computational approaches to DNA coding regions identification have some limiting features [1]: the methods within the approaches need training sets of already known examples of coding vs. non-coding regions and they are able to recognize mainly protein-coding regions.

Some new approaches were declared to be free of these limitations. E. Yeramian [2, 3] considered the DNA sequence as a linear chain of strong (GC-bond) and weak (AT-bond) hydrogen bonds. According to this representation, the partition function was calculated and a thermal DNA stability map (a plot of probability of DNA basepair to be disrupted) was build up. Due to the proper temperature choice, the map in some cases shows reasonable correlation to the layout of coding

regions in DNA. The method was successfully applied for identification of coding regions in *Plasmodium falsiparum* in some non-standard for gene-finders situations.

Another promising approach is to introduce some measure that enables the DNA sequence partition into homogeneous subsequences corresponding to the "coding vs. non-coding" regions. The method of DNA segmentation proposed by Bernaola-Galvan et al. [4] is based on the entropy measure of codon composition. The basis hypothesis is that the codon composition for coding regions differs from that latter for non-coding ones due to the well-known fact of a bias in codon usage.

Audic and Claverie [5] and later Baldi [6] obtained the similar partition by expectation maximization algorithm applied to a mixture probabilistic model. In that case every homogeneous part corresponds to the definite Markov model.

It was found [7], that the inphase hexamers seem to be the most effective method separating DNA coding vs. non-coding regions, among the other methods with no implementation of external information into the separation procedure. Calculation of inphase hexamers consists in 1) division of given sequence into non-overlapping triplets, 2) calculation of dicodon occurences starting from the first, second and third position in triplets.

In this paper we introduce a method of identification of protein-coding regions in DNA that is based on the notion of *distinguished coding phase*. We consider the distribution of DNA sites in the multidimensional space of frequencies of non-overlapping triplets. The main feature of the distribution standing behind the method is provided by the distinguished coding phase. The same reason allows us to explain the effectiveness of some other measures using in some way the idea of coding phase (including such measures as inphase hexamers, assymetry, entropy etc., see [7, 8] for definitions). Implying the idea *explicitly*, we formulate the self-training procedure of identification of protein-coding regions both for entire genomes and unassembled genomes (short contigs of 200-300 bp long, where ORF extraction is impossible).

Consider an arbitrary DNA subsequence (a single strand). To get the vector of non-overlapping triplet frequencies called *triplet usage*, one divides the sequence into non-overlapping triplets of nucleotides and calculates frequencies of all observed triplets (64 ones). There are three different versions of the sequence triplet usage depending on the position of the start triplet. Starting with the first, the second and the third base in the sequence, one obtains 3 vectors or 3 distributions of triplet frequency called *triplet usage in three different phases*. Let's denote them $f^{(1)}_{ijk}$, $f^{(2)}_{ijk}$, $f^{(3)}_{ijk}$, $i, j, k \in \{A, C, G, T\}$. Consider also a mixed distribution $f^{(s)}_{ijk} = \frac{1}{3}(f^{(1)}_{ijk} + f^{(2)}_{ijk} + f^{(3)}_{ijk})$.

To begin with, suppose a given sequence is protein coding and homogeneous (without introns). Thus, one of the three triplet distributions (phases) is the codon distribution. The triplet usage that corresponds to the real DNA $\rightarrow$ protein translation is called *distinguished coding phase*. Distribution of triplets in protein coding region is strictly conserved and the coding phase has to be different from two others because of well known fact that frameshift mutations are lethal. It

means that three phases of triplet usage are not equivalent in protein coding region. Phase difference in protein coding regions is maintained in the process of evolution, insertion or deletion of one (or two) base makes all the coding region senseless.

Suppose now a given sequence to be non-coding. The situation becomes fairly opposite. Now the operations of a base deletion or insertion are allowed, because it makes no crucial changes in protein properties, as a rule. Thus, all three distributions are expected to be equivalent and proximal to the mixed one $f_{ijk}^{(s)}$.

One can determine whether a DNA site is protein coding through the detection of the occurrence of distinguished coding phase, whatever a hypothetical nature of genes (or exons) is. It provides a self-training technique for automated gene recognition. Here we propose probably the simplest method to do that.

It should be stressed that the method proposed is not similar to the methods based on the experimental fact of *codon bias*. Instead, we rely on three observations: 1) information in protein-coding regions is coded by triplets; 2) this information is conserved in evolutionary process but frameshift mutations are lethal; 3) junk regions are not sensitive to base deletion or insertion, as a rule. These assumptions yield a conclusion that protein-coding region must have three *different* triplet usages and the single distinguished coding phase; so, the presence of the coding phase can be revealed.

Note that triplet distributions $f_{ijk}^{(1)}$, $f_{ijk}^{(2)}$, $f_{ijk}^{(3)}$ are projections of pentamers distribution f_{ijklm}, $i,j,k,l,m \in \{A, C, G, T\}$, calculated at every third position starting from the first base. It means that information contained in pentamers distribution seems to be sufficient for a prediction of coding regions with the same accuracy as hexamers do (while requires shorter subsequence to evaluate frequencies).

GC-concentration pattern shows another interesting point. One can convert a set of DNA sites (coding or non-coding ones) characterized by triplet (or pentamer, hexamer etc.) frequencies into a set of points in multidimensional space of frequencies. GC-concentration in DNA sequence is a linear function of the frequencies of triplets in any phase. Majority of the known genomes have GC-rich coding regions in comparison to the non-coding ones. It means that the gradient of this function defines a distinguished direction in the multidimensional space of frequencies that may stand as a normal vector of coding/non-coding sites separation plane.

To realize this simple idea of distinguished phase as a procedure for identification of protein coding regions in DNA sequence let us look through some preliminary results.

Consider rather long DNA. Suppose every nucleotide in the sequence to be "coding" or "non-coding" one. Let this property of a nucleotide be characterized by a measure calculated over the whole window of the length W and centered in the nucleotide position. So, this measure separates coding subset from non-coding subset in the set of all nucleotides (positions) of the given DNA sequence.

To begin with, consider the results of investigation of effectiveness of two simple discrimination measures (see Fig. 1) in dependence on the window length W. Namely, these measures are (a) local GC-concentration and (b) "mixing entropy".

Several different genomes with known annotation have been examined for evaluation of the dependence.

The local GC-concentration is the percentage of G and C nucleotides in the window while the "mixing entropy" is calculated as follows:

$$S_M = \frac{1}{3}(3S - S^{(1)} - S^{(2)} - S^{(3)}),$$

where

$$S = -\sum_{ijk} f_{ijk}^{(s)} \ln f_{ijk}^{(s)}, \qquad S^{(m)} = -\sum_{ijk} f_{ijk}^{(m)} \ln f_{ijk}^{(m)}, \qquad m = 1, 2, 3.$$

The mixing entropy shows the difference of a triplet distribution observed in three various phases from the averaged one. The measure characterizes applicability of the method proposed immediately.

The measure effectiveness indicates the quality of a separation between coding and non-coding positions in DNA sequence. Let $A_W(i)$ be a measure calculated for the i-th position. Then the effectiveness Δ_W of this measure is

$$\Delta_W = \frac{\frac{1}{|\mathbf{G}|}\sum_{i\in\mathbf{G}} A_W(i) - \frac{1}{|\mathbf{J}|}\sum_{i\in\mathbf{J}} A_W(i)}{\sqrt{D_{A_W(i)}}},$$

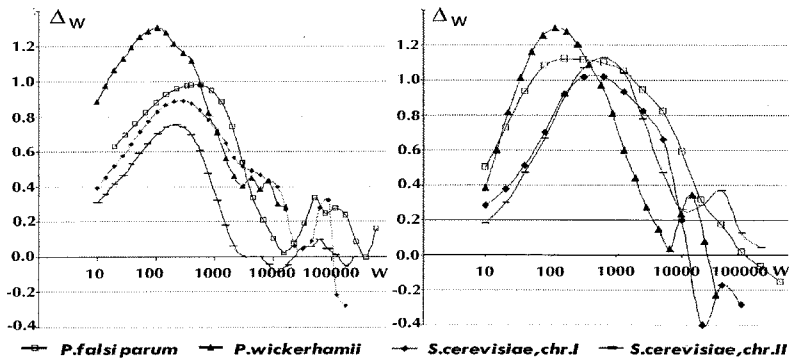
where $\mathbf{G}$, $\mathbf{J}$ are coding and non-coding subsets in the set of all nucleotides, $D_{A_W(i)}$ is the variance of $A_W(i)$ over the entire set $\{\mathbf{G}, \mathbf{J}\}$.

The window length (W) dependence of measure effectiveness (Δ_W) is shown in Fig. 1. The sets of junk $\mathbf{J}$ and genes $\mathbf{G}$ are separated for certain (a huge number of points yields the reliable difference between two mean values exceeding a standard deviation). It is clear (see Fig. 1) that the dependence is bimodal and maximal effectiveness is attained within a wide range of window length values. An optimal window length is about 400 bp for *S.cerevisiae* and *P.falsiparum* genomes and about 120 bp for shorter mitochondrial genome. So, the effectiveness of a measure based on calculation of GC-concentration and on the mixing entropy is the highest at the scales comparable to the average gene length within a genome while the mixing entropy seems to be more effective in comparison to GC-concentration.

Next point is devoted to the visual representation of DNA sites distribution in a multidimensional space of triplet frequencies illustrating the idea of the distinguished coding phase.

Consider rather long DNA sequence (entire genome, a chromosome or, probably, a set of short contigs in unassembled genome). One has to figure out an appropriate window length, as Fig. 1 shows. Each nucleotide in the sequence is characterized by 64-dimensional vector of non-overlapping triplet frequencies calculated in the window covering that latter. So, a DNA sequence is converted into a finite set of points in 64-dimensional space.

Preconditioning of the data set $X = \{x_i\}$, $i = 1, \ldots, N$ consists in standard centralization and normalization of each vector coordinate towards the zero mean



a) Local GC-concentration b) Mixing entropy

Fig. 1: Effectiveness of two measures (local GC-concentration (a), mixing entropy (b)) for several genomes. Bimodal character of graphs can be explained: first maximum is the difference of coding and non-coding regions themselves, second is statistical difference between vast regions.

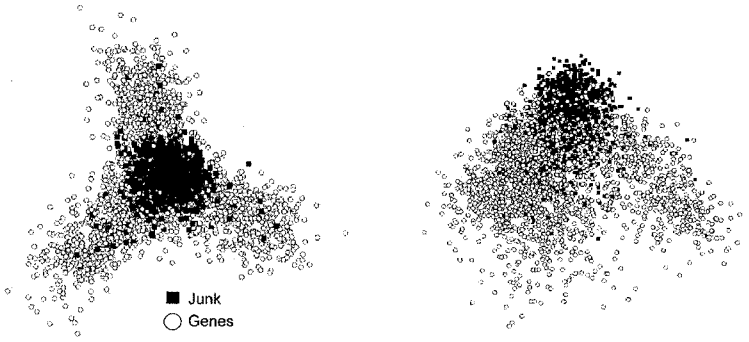
and the unit standard deviation,

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j},$$

where x_{ij} is the j-th coordinate of the i-th point and $\bar{x}_j$, σ_j are the mean value and the standard deviation of the j-th coordinate, respectively.

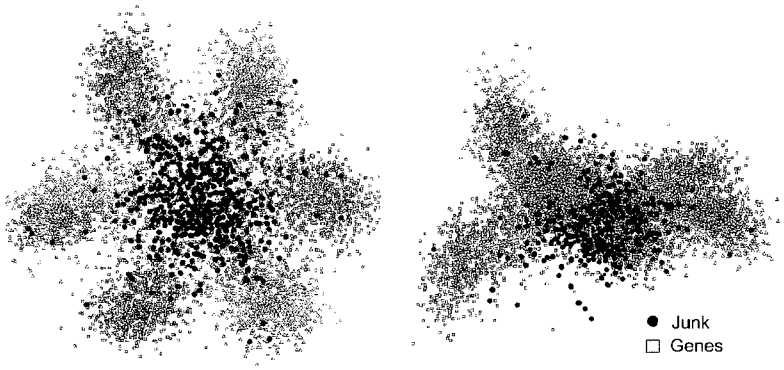
The set of normalized vectors $\tilde{x}_i$ was projected into the subspace spanned by the first three principal components of the distribution and visualized (see Figs. 2, 3, 4). The distribution has a bullet-like structure with the kernel corresponding to the non-coding regions and three (Figs. 2, 4) or six (Fig. 3) flank clusters corresponding to the protein coding regions. This structure reflects the difference between three phases in coding regions and their closeness observed in non-coding regions.

We studied three different genomes: 1) mitochondrion DNA of *Prototheca wickerhamii* (Fig. 2), 2) bacterial DNA of *Caulobacter crescentus* (Fig. 3) and 3) primitive eukaryot DNA of *Saccharomyces cerevisiae* (Fig. 4). All the examined entities have protein coding regions to be located over the both strands of DNA chain. This is the point revealing an occurrence of three or six flank clusters. Namely, in bacterial DNA (Fig. 2) three of six flank clusters correspond to protein coding sites of the forward strand (which is under consideration) and three other ones correspond to protein coding sites of complementary strand (triplets of that latter are complementary translated and read back to front). In the mitochondrion genome (Fig. 2), the images of coding sites of two strands turned out to be coincident. Six coding clusters of the yeast chromosome (Fig. 4) overlap each other and seem to yield three smeared clusters.



a) Top view b) Side view

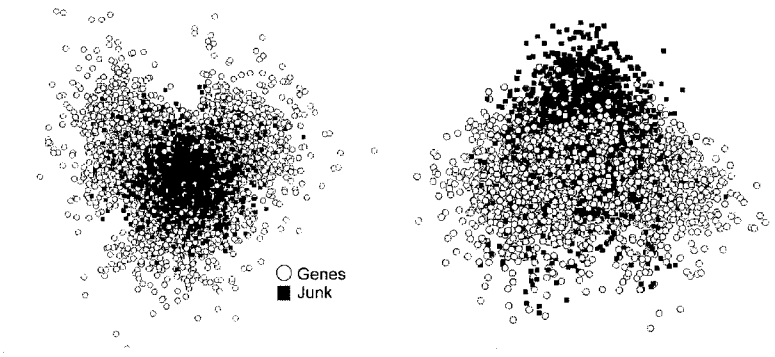
Fig. 2: Distribution of *P.wickerhamii* DNA sites in the subspace spanned by the first three principal components: a) projection on the plane perpendicular to the symmetry axis of the bullet-like structure; b) projection on a plane containing the axis.



a) Top view b) Side view

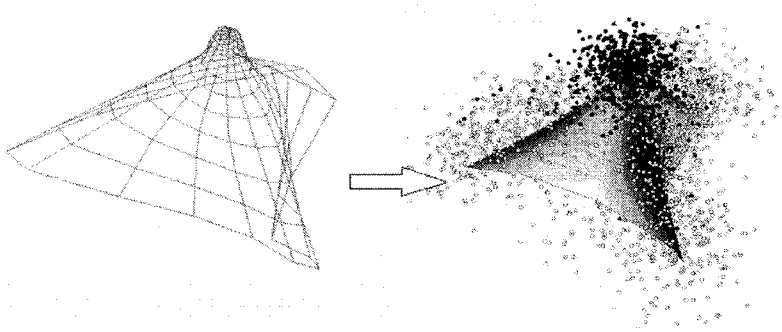
Fig. 3: Distribution of *C.crescentus* DNA sites in the subspace spanned by the first three principal components: a) projection on the plane perpendicular to the symmetry axis of the bullet-like structure; b) projection on a plane containing the axis.

A more advanced technology of data visualization was used for analysis of the distributions shown above. It is called the method of elastic maps (see [9, 10, 11, 12, 14]). The method of elastic maps develops a point approximation of 2D manifold of minimal energy placed in multidimensional data space, similar to the self-organizing maps (SOM) [15]. An elastic map is the non-linear two-dimensional principal surface (similar to the plain of two principal components) and works as a non-linear screen for data points projected on it. Briefly, elastic map is constructed



a) Top view b) Side view

Fig. 4: Distribution of *S.cerevisiae*, chr. III DNA sites in the subspace spanned by the first three principal components: a) projection on the plane perpendicular to the symmetry axis of the bullet-like structure; b) projection on a plane containing the axis.



a) Elastic net b) Elastic map in the data space

Fig. 5: Two-dimensional visualization of elastic map for *P.wickerhamii* dataset: a) form of the elastic net constructed; b) the elastic map position in the data space (projection on the first three principal components)

as following: initial 2D regular net of nodes with predetermined topology is placed into the multidimensional space; then, the net changes so that an elastic energy function reaches its minimum after applying optimization algorithm. This function is the sum of three terms characterizing (1) "node — data points" attraction, (2) energy of elastic stretching and (3) energy of elastic deformation of the net.

An elastic transformation of original net generates an elastic map as soon as one spans a piecewise-linear 2D-manifold over the deformed net. Then one may project piecewise-linearly the data points onto the map. This manifold (called

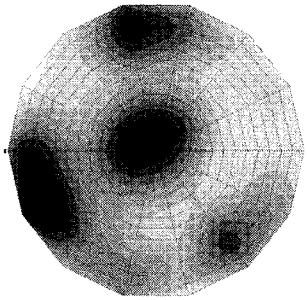


Fig. 6: Visual distribution density estimation by the elastic maps method.

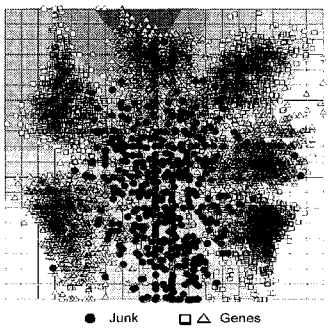


Fig. 7: The elastic map constructed for *C.crescentus* dataset (see Fig. 3)) together with data point projections and coloring by the value of start triplet frequency.

the elastic map) is used as a 2D screen for visualization of data points and any additional information, such as data density, distribution of internal and external data features and so on (see Figs. 6, 7).

Fig. 5 shows the elastic net for *P.wickerhamii* dataset. An initial net form was chosen to be the 2D-hemisphere. Optimization procedure makes the net take the form shown in the Fig. 5(a). The elastic net location in the data space (viewed in the projection on the first three principal components) is shown in the Fig. 5(b). A half-tone image was used to visualize the density (more precisely, its non-parametric estimation) of data points projections in Fig. 6; here the map is unfolded and represented in its internal coordinates. The distribution of data points has 4 clusters with central one corresponding to the non-coding regions and three flank clusters corresponding to protein-coding regions.

The elastic map constructed for *C.crescentus* is shown in Fig. 7 (in the internal coordinates of the map). In that case we chose simple rectangular net topology and initialized the net on the principal plane. The data points projections and the map coloring due to the value of start triplet (ATG) frequency are shown here. It is clear that the top middle cluster corresponds to the distinguished coding phase

of the forward DNA strand. Similarly, coloring the map according to CAT codon frequency reveals the right bottom cluster to correspond to distinguished coding phase of complementary strand. Another four side clusters correspond to coding regions, also, but in that case the triplet usages is shifted in phase due to the shift modulo 3 of the sliding window start position to the start nucleotide in the gene.

The cluster structure shown in Figs. 6, 7 is rather clear, so any clusterization algorithm can be implemented to classify the DNA positions into the central cluster (non-coding nucleotides), and flank clusters (coding nucleotides).

Thus, the procedure for unsupervised prediction of the occurrence of protein coding regions in DNA sequence is formulated as follows.

1) A window of the length W centered at position x is opened at every p-th position in the sequence. p is the sliding window step, which must be divisible by three in order to provide the same phase for the nucleotides from the same coding entity.
2) The subsequence isolated by the window is divided into $W/3$ non-overlapping triplets, and the frequencies of all observed triplets together with zero frequencies of other ones are arranged as the 64-dimensional vector. So, a DNA sequence is converted into a finite set of points in 64-dimensional space.
3) The set of 64-dimensional vectors runs through a preconditioning that consists in centering and normalization of each vector coordinate towards the zero mean and the unit standard deviation.
4) A visualization of data point projections over a 3-dimensional linear manifold spanned by the first three principal vectors of the distribution provides a researcher with the number of clusters and their compactness.
5) The data points observed in that 64-dimensional space must be divided into seven (or four) clusters due to some clusterisation algorithm.
6) Finally, the clusterisation procedure yields an attribution of each base at the position x to one of the clusters obtained previously. If this cluster is central one, then the base is likely to belong to non-coding region; otherwise, the base is suspected to belong to a coding region. One should expect the DNA site that consists of the bases belonging to the same cluster to correspond to the same protein-coding entity (same gene or exon).

A fragment of *P.wickerhamii* genome with the results of coding regions prediction is shown in Fig. 8. The parameters are as following: sliding window length $W = 120$, sliding window step $p = 3$, and the simplest method of K-means [16] for clusterisation into four clusters is implemented. The cluster number is ascribed to every point in the data set and to the corresponding nucleotide at the position x in DNA sequence, as well. The graph of cluster number of base alongside the given DNA sequence is shown in Fig. 8. Cluster number 0 corresponds to non-coding regions, clusters number 1, 2 and 3 correspond to the coding regions scanned in three different phases. The group of sequential bases with the same phase (except 0) belongs to the same gene (this mitochondrion genome has no introns). The genes occur in different phases over the chart (Fig. 8), since the junk sites differ in their length modulo 3.

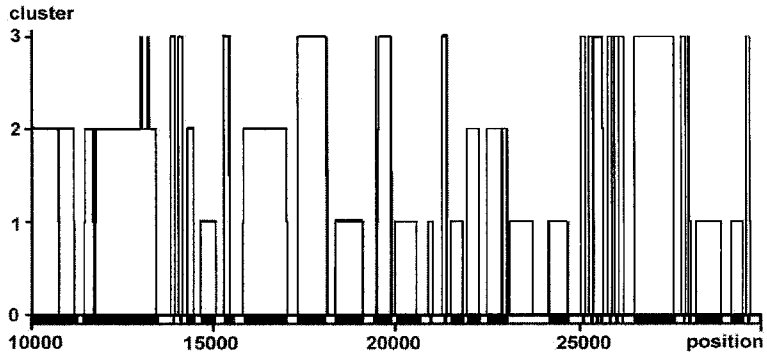


Fig. 8: Prediction of protein-coding regions for *P.wickerhamii* genome fragment. Cluster number 0 corresponds to non-coding sites. An abrupt jump of the cluster number in the chart predicts the start or the end of protein coding region. Solid lines indicate the known genes positions.

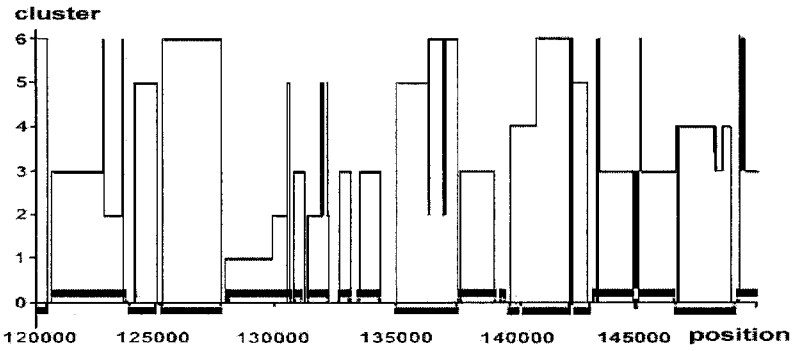


Fig. 9: Prediction of protein-coding regions for *C.crescentus* genome fragment. Solid line marks the known genes positions: top line marks forward strand genes, bottom line marks complementary strand genes

The pattern of predicted genes matches to real one (marked by solid lines) rather accurate. The genomes under consideration show the predicting accuracy of the method to run over 90% (a percentage of correctly predicted coding vs. non-coding nucleotides). This result is close to the exactness of gene-finders used in practice [1]; see [13] for more detailed comparative analysis.

The results of genes prediction for *C.crescentus* are shown in Fig. 9. The visual representation makes identify seven clusters here. As it was described above, the cluster 0 corresponds to non-coding regions. One can see the boundaries between coding and non-coding region to be correctly detected. The method fails to predict two exons, if they are located close each other, at the distance divisible by 3,

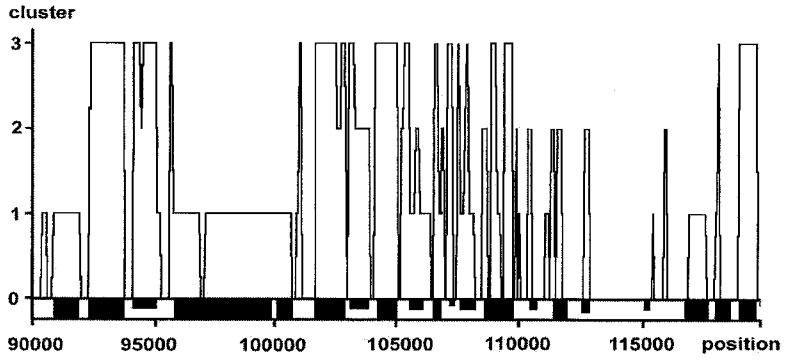


Fig. 10: Prediction of protein-coding regions for *S.cerevisiae*, chr. III. fragment by clustering into four clusters. Solid line marks positions of ORFs, the thickness of line corresponds to the confidence of gene presence (the thickest lines mark experimentally discovered genes.)

most likely they would be detected as a single exon. Nevertheless, the accuracy, calculated at the point level (i.e. a percentage of correctly detected coding and non-coding points) keeps rather high value of more than 90%.

The results of genes prediction for fragments of chr. III and IV of *S.cerevisiae* are shown in Figs. 10, 11. The cluster structure of the yeast dataset is not so distinct as bacterial one. So, we used a clusterisation into four and seven clusters for chr. III and IV datasets, respectively. The *S.cerevisiae* genome is not completely annotated yet. The genes not described precisely in genome annotation are shown in Fig. 10 by the variation in the thickness of solid line marking-up genes. An accuracy of prediction in the case of four clusters is found to be lower then that one in the case of seven cluster splitting. Yeast genome has the coding regions to be located over the both strands of DNA chain. So, the clusterisation into seven clusters is preferable, in similar situations.

The method proposed has rather high accuracy while detecting coding regions in the mitochondrion genomes, in the bacterial genomes and in the chromosomes of primitive eukaryot genomes where the coding regions density is high enough. Our algorithm allows to process unassembled genomes, since the dataset construction and clusterisation needs no genome assembly.

The technique for automated gene recognition proposed here is completely self-training. It needs neither predetermined information, nor the learning dataset. The sliding window length is the only parameter to be chosen by a researcher. However, its estimation makes no problem. The maximal effectiveness of the discrimination procedure is reached within a wide range of window length values close to an average exon length typical to the DNA under investigation.

An essential feature of the technique proposed is the preliminary visualization of the dataset in a subspace of the first three principal components of the distribution. As it was shown above, the distribution of DNA sites in the space of

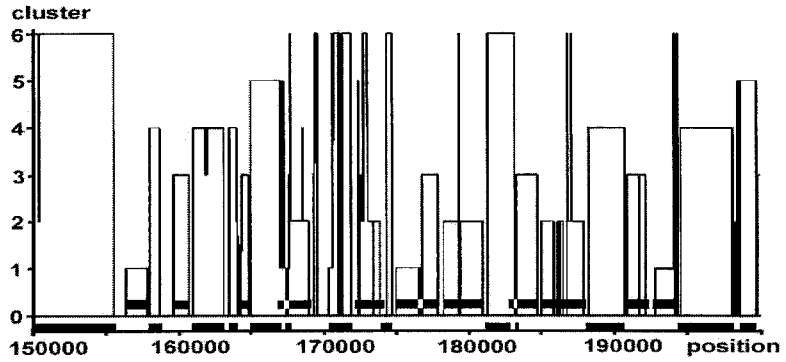


Fig. 11: Prediction of protein-coding regions for *S.cerevisiae*, chr. IV. Solid line marks the known genes positions: top line marks forward strand genes, bottom line marks backward strand genes

frequencies of non-overlapping triplets has bullet-like structure with one central cluster and three or six flank ones. Linear dimensions of this structure are determined by the amplitudes of two measures: local GC-concentration and mixing entropy. The bullet-like structure itself revealed in the distribution seems to be very interesting visual illustration of triplet usage in DNA sequence.

Besides the number of clusters provided, the visual representation is useful for estimation of the accuracy of coding regions identification. Clear and distinct cluster structure allows researches to expect high accuracy of prediction.

The main idea underlying the method is very simple: the distribution of non-overlapping triplet frequencies in three phases should be different in coding regions, while they should be similar in non-coding ones. The procedure for revealing the presence of distinguished coding phase is easy to realize, while the accuracy of prediction exceeds 90%. We believe, a genome annotation problem can benefit from the method proposed here.

Acknowledgements

Our efforts were inspired by Misha Gromov (IHES). We are thankful to Alessandra Carbone (IHES) and Iliya Karlin (ETH, Zurich) for stimulating discussion and help. We also thank Dr. Michael Sadovsky for the help in manuscript preparation.

Bibliography

[1] J.-M. Claverie, *Computational methods for the identification of genes in vertebrate genomic sequences*, Human Molec. Genetics **6**, 1735 (1997).

[2] E. Yeramian, *Genes and the physics of the DNA double-helix*, Gene **255**, 139 (2000).

[3] E. Yeramian, *The physics of DNA and the annotation of the Plasmodium falsiparum genome*, Gene **255**, 151 (2000).

[4] P. Bernaola-Galvan, I. Grosse, P. Carpena, and et.al., *Finding Borders between Coding and Noncoding DNA Regions by an Entropic Segmentation Method*, Phys. Rev. Lett. **85**(6), 1342 (2000).

[5] P. Audic, J.-M. Claverie, *Self-identification of protein-coding regions in microbial genomes*, Proc. Natl. Acad. Sci. USA **95**, 10026 (1998).

[6] P. Baldi, *On the convergence of a clustering algorithm for proteing coding regions i microbial genomes*, Bioinformatics **16**, 367 (2000).

[7] J. Fickett, *The Gene Identification Problem: An Overview For Developers*, Computers Chem. **20**, 103 (1996).

[8] A. Gorban, A. Zinovyev, and T. Popova, *Statistical approaches to automated gene identification without teacher*, Institut des Hautes Etudes Scientifiques Preprint, IHES/M/01/34 (2001).

[9] A. Gorban and A. Rossiev, *Neural Network Iterative Method of Principal Curves for Data with Gaps*, Journal of Computer and System Sciences International **38**, 825 (1999).

[10] A. Zinovyev, *Visualisaton of Multidimensional Data*, Krasnoyarsk State Technical University Press, Russia, 2000, 168pp.

[11] A. Gorban, A. Zinovyev, and A. Pitenko, *Data visualization by the method of elastic maps*, Informatsionnie technologii, Moscow. **6**, 26 (2000). (in Russian)

[12] A. Gorban and A. Zinovyev, *Visualization of data by method of elastic maps and its application in genomics, economics and sociology*, Institut des Hautes Etudes Scientifiques Preprint, IHES/M/01/36 (2001).

[13] A. Zinovyev, *Visualizing the spatial structure of triplet distributions in genetic texts* , Institut des Hautes Etudes Scientifiques Preprint, IHES/M/02/28 (2002).

[14] A. Gorban, A. Zinovyev, *Method of Elastic Maps and its Applications in Data Visualization and Data Modeling*, International Journal of Computing Anticipatory Systems, CHAOS. **12**, 353 (2001).

[15] T. Kohonen, *Self-Organizing Maps*, Berlin–Heidelberg, 1997, 420pp.

[16] J. Hartigan, *Clustering Algorithms*, John Wiley & Sons, Inc., New York, 1975.