STATISTICAL APPROACHES TO AUTOMATED GENE IDENTIFICATION WITHOUT TEACHER *

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

[1] **Institut des Hautes Etudes Scientifiques, France**

[2] **Institute of Computational Modeling,**
Siberian Branch of Russian Academy of Science

Abstract

Overview of statistical methods of gene identification are made. Particular attention is given to the methods which need not a training set of already known genes. After analysis several statistical approaches are proposed for computational exon identification in whole genomes. For several genomes an optimal window length for averaging GC-content function and calculating codon frequencies has been found. Self-training procedure based on clustering in multidimensional codon frequencies space is proposed.

*Institut des Hautes Etudes Scientifiques Preprint. IHES M/01/34. Online-version: http://www.ihes.fr/PREPRINTS/M01/Resu/resu-M01-34.html, e-mail of corresponding author: gorban@icm.krasn.ru, auranic@online.ru

1 Introduction

More than 5 years elapsed since first genomes were fully decoded. Now the world data bank of decoded (and fully or partially annotated) sequences consist of 51 complete genomes available in public databases. Overall, nearly 13 billion nucleotides of sequence are contained in the GenBank database.

There is a kind of boom in molecular biology, and it is evident now that main expectations of mankind from science has been shifted to the area of biotechnologies allowing to interfere to the program of living cell and to construct in prospect artificial alive systems with programmable properties.

During the last decade much attention has been focused on the computational methods of gene identification. Though biochemical machinery in a live cell detects coding regions in DNA highly effectively, there is no simple way to predict them computationally with good accuracy. Nevertheless many programs now identify genes more or less effectively and some computational techniques proved to be useful. Most of the programs use for gene recognition some statistical methods such as linear discriminant analysis or its generalizations. Usually methods require preliminary training phase when weights of the decision function or neural network are calculated.

Generally, there are two basic directions in technologies of pattern recognition. First is the methods that use a training set to tune a classification rule. These methods are traditionally referred to as "learning with teacher" or supervised learning. The other group of methods are traditionally called "learning without teacher" or "self-training", "self-organizing" or unsupervised techniques.

This work considers methods of gene identification giving particular attention to the approaches where no training set and phase of learning on already known genes and junk of analogous sequences is necessary.

The main accent in the work has been made on the trying to undestand better some aspects of statistical approach in the problem of gene identification. The methods described are not ready practical instruments for gene-finding, but they may be applied in real practice using some additional tools (signal sensors, alignment etc.), because none of the modern gene-finders uses the only method.

ACKNOWLEDGMENTS. This work was initiated in the autumn of 2000 when one of the authors (Prof. A.Gorban) worked in IHES (Institut des Hautes Etudes Scientifiques, France). Our efforts were inspired by attention of Prof. M.Gromov to the work. We are thankful to Prof. A.Carbone and

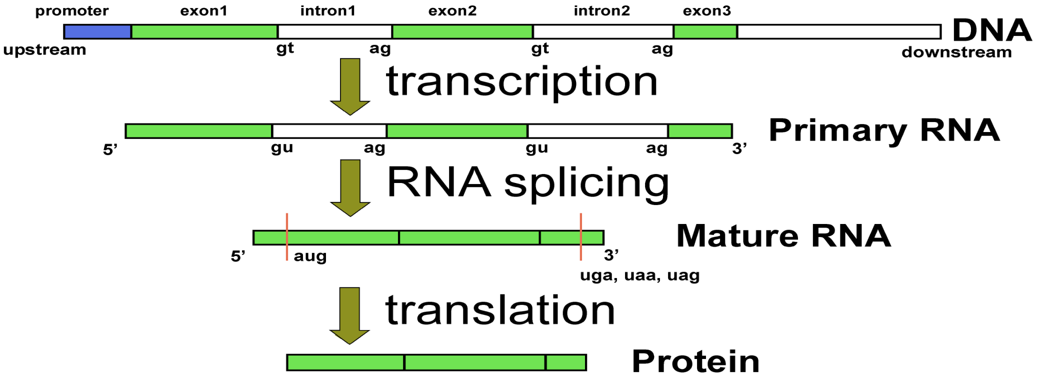

Figure 1: Scheme of processing of protein production by biochemical machinery of a living cell

Dr. R.Incitti (IHES) for stimulating discussion and help. The work was finished in the summer of 2001 in kind atmosphere of IHES, where two of authors were staying as visitors. The paper of A.Carbone and M.Gromov (2001) was a good inspiring example for us of how molecular biology can be thought from a mathematical angle.

2 Methods of automated gene identification

All information necessary to maintain cell life cycle is embedded in the sequence order of four nucleotides: A (Adenine), C(Cytosine), G(Guanine), T(Thymine) in the long DNA molecule. Receiving aminoacids from outside and using double DNA helix as a template, a cell produces all materials necessary for its life. This view on the cell functioning is usually referred to as "Crick's dogma" (named by the Nobel Prize winner in molecular biology), see Fig.1.

Physically DNA is a long molecule intricately packed in space and its structure is determined by the forces of two kinds.

Covalent bonds provide forming of *polynucleotides*. Molecule of each nucleotide A, C, G, T is built out of the sugar-phosphate group and the base attached to it. Sugar-phosphate groups are naturally polarized. They can bound with each other, forming molecules with hundreds of thousands nucleotides. One hundred bases long and shorter sequences are called *oligonu-*

cleotides (and they can be produced *in vitro* according to a given specification of letters).

Hydrogen bonds are weaker in the order of magnitude, and they provide DNA complementarity. So, DNA is two polinucleotides with equal length bound by hydrogen bonds , and in one of them every A letter is substituted by T in another, C replaced by G, and vice versa.

GC-bond is provided by three hydrogen bonds, AT-bond - by two, thus we may consider the two DNA strands as a binary sequence of strong (G-C) and weak (A-T) bonds.

A section of DNA in a gene, coding biological information, is called *exon*. Exons can be classified in four classes: ”starting” exon, ”inner” exon, ”terminal” exon and ”single” exon (in case when the gene has no introns). Replacment of one nucleotide in an exon for an other one (as well as operations of inserting ”superfluous” and deleting ”necessary” nucleotide) may change properties of coded protein radically, so exon compositions are practically identical for genes of organisms of the same species. Moreover, genomes of higher species contain in many genes almost the same base sets as their distant primitive ancestors.

Sections of DNA, that do not code information, may be *junk* or *introns*. Junk fills areas between genes. Junk function is forming the skeleton of DNA - its secondary space structure. It seems that small changes in junk composition don't lead to considerable modifications in DNA properties. Introns are areas dividing exons in a gene. In translation process introns are cut off and the information coded in them is not present in the resulting protein.

The problem of *automated* (not experimental) *genes identification* may be formulated as following:

A sequence of letters A, C, G, T, corresponding to the order of nucleotides in genome, is given at the input of computer program. At the output we need to have a list of identified genes (biologically active sections of DNA) with indicated start, end and gene structure - division into the *exons* and *introns*, and to depict the scheme of translation of mature RNA to protein with possible assumptions for the structure and the function of resulting product.

This problem definition can be specified (and simplified) in a way and it determines the method of evaluating identification accuracy. For example, it is possible to consider following definitions:

1) Evaluate the confidence of nucleotide in the definite position to belong to the biochemically active section of genome (exon of a gene). It is *identification of the nucleotide level*. As a result of the work of identification program,

every nucleotide could be classified on the four groups: a) "predicted" and already "known"; b) "not predicted" and "known"; c) "predicted" and "unknown"; d) "not predicted" and "unknown". Comparing with known full annotation to evalute the accuracy of the method, we should count coding nucleotides predicted to be coding (true positives), non-coding nucleotides predicted to be non-coding (true negatives), and the errors of two type: coding nucleotides predicted to be non-coding (false negatives) and non-coding nucleotides predicted to be coding (false positives).

2) Identification on the *exon level*. In this definition exons are identified (without considering their gene membership). After work a program gives a set of exons and each of them belongs to one of the three classes: a) "predicted" and "known"; b) "predicted" and "unknown"; c) "not predicted" and "known". An exon may be predictected exactly (with both of its borders), partially (only one border exactly predicted), and in the sense of overlapping (the predicted exon only overlaps the real one).

3) Gene identification. Start of gene, end of gene and structure of gene are predicted.

There are two ways of DNA analysis. One is to consider both complementary strands separately, and to identify separately W-(Watson) and C-(Crick) genes. The other way is to analyze only W-strand (taking into account complementarity). On the second way the problem of genes *overlapping* appears, when two different genes in the upper (W) and lower (C) strands correspond to the same base pair position.

Possibility of alternative translation is an additional complexity for identification procedures, when different divisions of the gene into exons and introns correspond to different products of translation.

It is possible to distinguish three different approaches in the methods of gene identification. They could be called *similarity search*, *content search* and *signal search*.

Similarity search is one of the first group of methods that were applied to identificate genes in new genomes. It is based on the statement that the function of a gene defines to some extent its nucleotide composition, and if two genes code similar products then the corresponding sites of DNA will be similar.

One of the early attempts to evaluate the possibilities of similarity search in a new genome using already known analogs in a database was made by Seely (1990). Rather big collection of genetic sequencies (Genbank release 56) was arbitrarily divided into two halfs. Then genes from one part of the collection were searched with use of the other part as a database. The result was almost 75% correctly identified genes. But when applied to the real new experimentally annotated genomes the method gave only 20-25% of identified genes. Now it is stated that in new decoded genomes up to 50% of all genes may be identified by methods of similarity search.

Content search is based on the fact that statistical characteristics, calculated in DNA analysis, differ considerably in coding and non-coding regions. For more than 15 years the whole "zoo" of such features was formed. All these features appeared from observation of structure of nucleotide compositions in genes and junk. The earliest features - frequencies of codon (triplets) usage, some types of Fourier-transform were thoroughly investigated and their ability for gene identification was systematically tested (see, for example, Fickett, 1996).

The earliest attempts to undertake content search looked for a discriminate function (linear, as a rule) in multidimensional space of the features. This approach yielded quite good results and some methods proposed were included in computer programs (for example, HEXON, GRAIL) that became real instruments for primary investigation of new decoded sequences. These programs usually use discriminating rule that is trained on the known analogous samples.

A recent kind of fashion in this field is to develop such features and rules separating genes and junk, that do not require preliminary training. As a rule, these methods use one or two integral measures (for example, nonuniformity of codon usage expressed in entropy terms, patchiness of junk, expressed in difference of the means calculated in small and large windows, probability of DNA melting etc.).

It is worth noticing that methods of content search and similarity search share common ideological premise which can be called "comparison with sample". In case of similarity search such comparison is made at the level of alphabet, and in case of content search some integral characteristics are compared. On one end of an imaginary scale there is a method of reducing all information about the site under consideration to one value. On the other end - a case when features are such expressions as "on the first position letter b_1, on the second - b_2" and so on, and as a result we have highly multidimensional space where set of samples (sequences from database) are disposed.

The third principle of genes identification - signal search - is based on the hypotheses about physical and chemical processes, initiating transcription.

The molecule that initiates the start of transcription ”recognizes” it by the presence of active sites - signals, that are short sequences with a definite structure. There is no clear concept of what are the factors that cause some sites of DNA to serve as signals. Dictionaries of signals - initiators and terminators of transcription - are known, but all these sequences may occur in DNA without initiating any process.

At the early stages of using signal search there were hopes that it would be possible to construct one or more consensus signal sequences and to measure the distance from DNA site to the consensus (using alignment). The following approach was supposed: the first letter of consensus sequence is the most frequent first letter in all already known signals, the second is the most frequent second letter and so on. Though this approach turned out to be too primitive, at present one of its generalization is successfully applied (when all four letters are used rather than one with calculated probabilities, and resulting consensus is a probability matrix).

At present tens of programs and algorithms realize automated gene identification, recent excelent overview of perfomance of part of them is given in the paper of Rogic, Mackworth, Ouellette (2001). The most effective programs use several approaches simultaneously. Unfortunately to choose one best program is not a trivial task as well as to compare in a reasonable way results of their analysis. First, different algorithms show different results on different databases of annotated genomes. Second, so far there is no single opinion how to compare one program with an other (especially it concerns comparing predicted gene structures). Nevertheless computer prognosis of genes allocation is now the necessary stage in experimenter’s work.

3 Features used in content search

In this section we make a short overview (using Fickett, 1996) of the features that help to do content search of genes (exons).

1. *Codon Usage.* Sequence under investigation are divided into successive non-overlapping nucleotide triplets (test codons), and all their frequencies are calculated. As a result we have vector with $64 = 4^3$ components.

2. *Hexamer-n, n=0,1,2 or Inphase Hexamers.* Frequencies of hexamers usage (there are $4096 = 4^6$ possible hexamers) offset by n are calculated. The Hexamer-0 measure gives dicodon frequencies. The Hexamer-1,2 measures give dicodon frequencies offset by 1,2 from Hexamer-0.

3. *Hexamer Usage.* Sequence under investigation are divided into successive non-overlapping hexamers, and all their frequencies are calculated.

4. *Open Reading Frame.* Length of the longest site in the window starting from start-codon and ending with stop-codon. Possibly start-codon really initiate transcription, but it may occur ”accidentally”.

5. *Amino Acid Usage Measure.* The 21-vector obtained by translating the sample window of sequence, beginning with the first base, according to the appropriate genetic code, and counting the frequencies of the 20 amino-acids and ”stop”.

6. *Diamino Acid Usage Measure.* The 441-vector given by translating the window and counting all the (overlapping) dipeptides (including ”stop” as an ”aminoacid”).

7. *Stability at Hydrophobicity Measure.* First define the information value of a codon as $\sum_{j=1,3}[\sum_{j=1,n_j}(p_i \times d_{ij})]/n_j$, where n_j is the number of sense mutations of the codon, p_i is the probability of the $i-$th mutation, and d_{ij} is the difference in hydrophobicity caused by the mutation. The information value of a window, which we take as the Stability of Hydrophobicity Measure, is then the average information value of the test-codons in that window.

8. *Composition Measure.* $f(b,i)$, where for each base b = A, C, G, T and each test-codon position i = 1, 2, 3; $f(b,i)$ is the frequency of b in position i.

9. *Position Asymmetry Measure.*
Define $\mu(b) = \sum_i [f(b,i)]/3$ and $asymm(b) = \sum_i [f(b,i) - \mu(b)]^2$. Then define the position asymmetry measure to be [assym(A), asymm(C), asymm(G), asymm(T)].

10. *Entropy Measure.*
Given $f(b,i)$ as above, define $entropy(i) = \sum_b f(b,i) \ln[f(b,i)]$. If the three values of $entropy(i)$ are significantly different a coding region is predicted, and the one with the largest difference from random is predicted to be third codon position.
We define the Entropy Measure to be $[entropy(1), entropy(2), entropy(3)]$.

11. *Autocorrelation Measure.* Let $auto(b,i)$ be the number of pairs of base b with i intervening bases. For the measure we correct for the number of such pairs expected on the basis of base composition alone, giving the matrix $[auto(b,i)/(windowlength - i - 1)(frequencyofb)^2]$, where b=A,C,G,T, and $i = 0, 1, ..., 9$.

12. *Fourier Measure.* Let the window be $2M$ long. Let $EQ(x,y)$ be the function which is 1 if $x = y$ and 0 otherwise.
Define the $n-$th Fourier coefficient (dropping the constant $1/4M^2$ for sim-

plicity) by: $FC(n) = \sum_p \{\sum_m [EQ(basem, basem - p)]\} \exp(\pi inp/M)$. Then define the Fourier Measure to be $[FC(2M/2), FC(2M/3), ..., FC(2M/9)]$ (i.e. the Fourier coefficients of the autocorrelation function for periods 2-9).

13. *Word Measure.* Divide the window into successive. non-overlapping words of length 2 and also into words of length 3. The measure is the pair of chi-squared values comparing the frequency distributions of these words with the uniform distribution.

14. *Run Measure.* Lets $S_1, S_2, S_3, ..., S_{14}$ be the non-trivial subsets of the set A, C, G, T. For each S_i construct a new sequence by replacing each base in S_i with 1 and replacing each base not In S_i, with 0. Using this sequence define r_{ij} to be the number of runs of 1 of length j, for $j = 1, 2, 3, 4, 5$ and let r_{i6} be the number of runs of 1 of length greater than 5. The run measure will be the set of values $[r_{ij}]$.

15. *Dinucleotide Bias Measure.* Let $f(w)$, for any possible word w, be the frequency of w in the sample window. Now for each dinucleotide ab let $bias(ab) = f(ab) - f(a)f(b)]/f(a)f(b)$. The Dinucleotide Bias Measure will be the bias values for the 16 dinucleotides.

Fickett made an attempt to benchmark these measures for their ability to distinguish between coding and non-coding sequences. There were homogeneous (fully coding or fully non-coding) sequences taken from the database. This set was divided into training and testing subsets. Linear division rules were learnt on the trainig set, and then it was tested on the second set. The average accuracy on the coding and non-coding parts of the test set was taken as the overall accuracy of the measure.

Derivable from each other measures turned out to give similar results. Maximum of the accuracy with using only one measure was 76%. Use of several measures yielded the accuracy of 82-88% in different experiments.

The most effective measure was Inphase Hexamer Frequencies which seems to embody little biological understanding.

Another result was that some measures give considerably different linear rules (their weight values) for sequences with different GC-concentration. Several authors tuned linear rules for different groups of sequences which differ by GC-bonds concentration and then defined procedures of linear interpolation of the weights.

Additionally we take a note of the following regularities which were discovered in content search:

a) introns differ considerably from exons and from junk (have their own characteristics such as the two-base periodicity in the occurrence of certain oligonucleotides)

b) intergenic DNA has statistical properties very different from gene flanking sequences

c) not all components of multidimensional measure vectors are equally informative. Possibly "signal-to-noise" ratio of the measure could be improved by pruning out the less informative variables.

In the end it is possible to conclude that the epoch of invention new statistical measures come to the end. We think that the future generation of gene identification programs will use methods which do not require a training set for learning and the methods will be applicable for whole genomes.

For example, in the work of Bernaola et.al., 2000 the following procedure of finding borders between junk and exons has been proposed: a pointer slides along the genome and divides it into two subsequences in each of them a measure analogous to the entropy introduced above is calculated. The measure of current position is the value characterizing the degree of heterogeneity of two parts as compared to the whole sequence (similar to the mixing entropy). As far as the measure exceeds the value calculated for a random sequence with a definite significance level, the sequence is cut at this point. Otherwise the sequence remains undivided. The procedure continues recursively for each of the two resulting subsequences created by each cut. Before a new cut is accepted, a check that the subsequences formed by the cut remain significantly different from their neighbors is perfomed. The process stops when none of the possible cutting points has a significance level exceeding threshold s. Then the authors say that a sequence is segmented at "significance level" s.

In the next section we look at an approach when DNA is considered as a two linear complimentary chains of nucleotides, that is stapled by strong and week hydrogen bonds.

4 DNA Thermal Stability Maps

Thinking of a DNA molecule as a linear chain of strong and week hydrogen bonds between nucleotides, one may model statistically the process of melting of DNA (thermal disruption of complementary bonds with temperature growth).

Every microscopic state of the chain is described by a binary sequence of closed (1) and opened (0) complementary bonds. Three factors determine

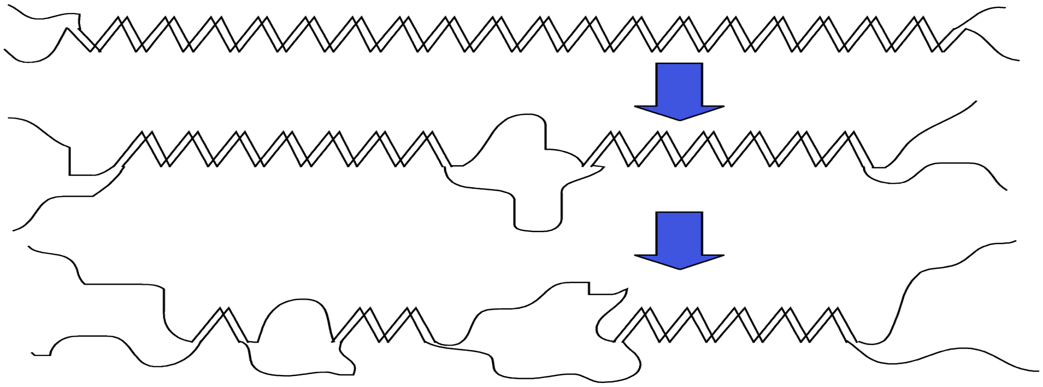

Figure 2: DNA melting

the energy of the state: a) open ends of the chain; b) consecutive bonded units; c) loops of length j: they are j consequtive nonbonded units that are both preceeding and followed by at least one bonded unit. A symbolic picture of the melting process is shown in fig.2.

It is natural to assign to the open end of the chain statistical weight 1 (it corresponds to the zero energy). Every bonded unit gives contribution s_i (it may have two values: s_{GC} and s_{AT}). Statistical weight of a loop is described by function $\omega(j)$, where j is the length of the loop.

Form of $\omega(j)$ dependence is an essentially unknown characteristic of the model. On the other hand, this dependence is responsible for the presence of long-range correlations and cooperative effects like zipping. For the linear chain and in the model of random walk of the loop one can get dependence $w(j) \sim j^{-\alpha}$, where α is a number in range from 1,5 to 2 (see Wada, Jacobson, 1980).

Full enumeration of all possible binary combinations for calculation of the partition function demands the number of computational operations proportional to 2^N, where $N \sim 10^6$ is length of sequence. Since it is unrealistic to solve the task with the present state of computational powers (and is improbable to be ever possible), it is necessary to optimize the enumeration taking into account specificity of the problem.

It turns out that good organization of enumeration and optimal grouping of the states reduce the number of operations to proportional to N^2. So the overall number of operations needed to calculate the probability of every bond to be opened is proportional to N^3. This number is still too great to make operative calculations (it is mentioned in the work of Yeraminan, 2000 that when using Sun Ultra2 workstation 200MHz it took 5 days to make calculation for the sequence of 48kbp length).

The number of operations proportional to N^2 is necessary because to calculate the partition function one should take into account the contribution of loops of different lengths (from $j = 1$ to N). The idea suggested in the work of Frank-Kamenetskii,1969 based on a simple statement that if $w(j)$ meets condition $w(i + j) = w(i)w(j)$, then a long loop could be considered as a sequence of loops with lengths 1 and 2. This is possible only when $w(j) \sim \exp(kj)$. Unfortunately such a form of function w has no physical meaning.

Nevertheless, if one approximates function $w(j)$ with a sum of exponents, i.e.

$$w(j) \sim \sum_{i=1}^{K} A_i \exp(k_i j),$$

then the process of calculating the partition function could be separated into several threads then in each of them the contribution of the loop is described by exponent. As a result the number of computational operations reduces to $K \times N$, where K is number of exponents in approximation of $w(j)$.

In the calculation that was made by Yeramian 1990,2000, for approximation of the function by exponential series, Pade-Laplace transform method (see Yeramian E., Claverie P., 1987) was applied. The approximation error of function $w(j) = j^{-\alpha}$ on interval $j \in [0..2000]$ was 0.04%. Applying the approximation to the calculation of DNA stability gives resulting error less then 1%.

The map of gene allocations was superposed with the picture of probability of DNA disruption. For a definite temperature correlation between genome annotation and probability graph becomes clear. As a rule genes are more "refractory" then junk areas. The author explains it by the hypothesis about genome origin as a result of chaining ancient genes or RNA in one long molecule. With growth of temperature first junk areas are disrupted, and genes after them.

5 Local Binding Energy. Optimal Window for GC-averaging.

Unfortunately even though the above described calculations claim to have direct relationship to the real DNA double-helix they use some simplified approximations. First, it is the replacement of the intricate spatial structure of the DNA in cell by a linear chain. Second, it is unknown *realistic* statistical weight of a loop. Third, the picture of correlation differs considerably with changing temperature, so the temperature at which the method has predictive power is one of the parameters to be fitted.

Calculation of the partition function is rather complicated procedure though it seems to us that comparable results can be obtained by calculating incomparably simpler functions. In the next sections we will show several possible approaches to analyze genome structure and all of them have to some extent predictive power for gene identification.

Calculation of DNA stability map and comparing it with genome annotation shows that probably a considerable part of the information about allocations of coding regions in genome is contained in the simple and physically clear value of locally evaluated energy of binding of two complementary strands of DNA.

Of course, the optimal form of the kernel to average GC-function is unknown. In our work we consider two simple variants of step-function (this section) and an exponent (next section). Detailed analysis of suitable basis for the optimal expansion of the kernel is an opened direction.

Presentation of DNA as a chain of strong and weak bonds corresponds to projection $S(\{A, C, G, T\} \to \{0, 1\}) = (\{G, C\} \to 1,\ \{A, T\} \to 0)$, that leads to loss of a part of information, contained in the DNA word (one bit is needed to encode every nucleotide but not two). But, one can expect that in such a coding an essential part of information, needed for distinguishing genes and junk, is nevertheless preserved.

Let's analyze statistical characteristics of the binary sequence.

Averaging the sequence with sliding window of different widths we get graphs of changing local binding energy vs position of DNA. Let's determine optimal window width. Averaging with this window we hope to get the most contrast picture of correlation of binding energy with allocations of coding regions.

Divide all genome for the regions of genes (in one of two complementary strands) and regions of junk.

We will average the binary sequence with sliding window of width W, which for simplicity we make an even number. The mean value is denoted by

$$A_W(i) = \frac{1}{W} \sum_{j=i-W/2}^{i+W/2} C(j),$$

where $C(j) = 1$, if a GC-pair is in the j-th position of DNA and $C(j) = 0$ if otherwise.

Let a genome contain N_G exactly identified genes, each of them occupies a region $G_k, k = 1 \ldots N_G$. Let's separate in genome N_J regions in the areas between ORF's (where nonoccurence of genes is guaranteed). Each of them occupies region $J_k, k = 1 \ldots N_J$. Value

$$A_G(W) = \frac{\sum_{k=1 \ldots N_G} \sum_{i \in G_k} A_W(i)}{\sum_{k=1 \ldots N_G} \sum_{i \in G_k} 1} \quad (1)$$

is the local binding energy, calculated with sliding window of W width, averaging for the genes. So

$$A_J(W) = \frac{\sum_{k=1 \ldots N_J} \sum_{i \in J_k} A_W(i)}{\sum_{k=1 \ldots N_J} \sum_{i \in J_k} 1} \quad (2)$$

is an analogous value, but averaged for junk areas. Let's choose window W such that

$$\Delta(W) = \frac{A_G(W) - A_J(W)}{\sqrt{DA_W}} \to max, \quad (3)$$

where D is dispersion.

We had in experiments several sequences: Prototheca wickerhamii (GenBank U02970), Plasmodium falciparum (Chromosome II) AE001362, Yeast genome (Chromosome I,II,III,IV,VIII). We chose these sequences wishing to compare the results with calculations of probability of DNA melting (see previous section). Specificity of the Yeast genome is that in corresponding annotation the genome has already separated into the ORF's which are characterized by definite value of "reliability": how confident a presence of biologically active site is detected. ORF with reliability "1" corresponds to the genes found experimentally. Reliability "6" corresponds to the ORF, where a gene is either not detected or the resulting function of the protein

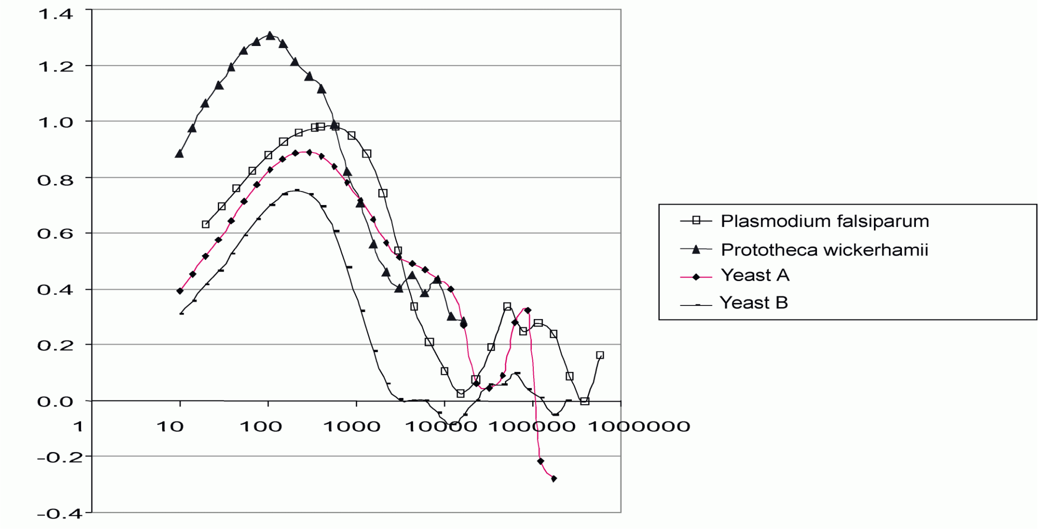

Figure 3: Normalized on standard deviation difference between genes and junk vs window width

is unknown. For calculations we take ORFs with reliability ”1” (gene in any case).

Graphs of $\Delta(W)$ dependence for several genomes are shown in fig.3. For genomes of Yeast and Plasmodium falciparum the optimal window width is $W_{opt} \sim 400$, for a short (50 kbp) mitochondrial genome Prototheca wickerhamii $W_{opt} \sim 100$. It is interesting that the dependence has bimodal character. One more local maximum corresponds to $W \sim 100000$. Bimodality $\Delta(W)$ can be explained: first, we have statistical difference of genes and junk themselves, and also we have regions in a genome with length about 100000, where one gets more genes (or junk) as the average.

Graphs of local binding energy $A_W(i)$ that obtained by optimal averaging are shown in fig. 4. It is apparent that setting an acceptable threshold (this value may be optimal for a ”training” set of gene and junk regions or evaluated from reasonable considerations: the simplest way to choose EA_W, where E - averaging operator), one can get quite a contrast picture of correlation values of the local binding energy and allocations of coding regions. The separation accuracy calculated as a number of correctly classified nucleotides was 60-70 %.

Let’s characterize every homogenous (fully coding or fully non-coding) region S_k of a genome by two measures: the value of local binding energy,

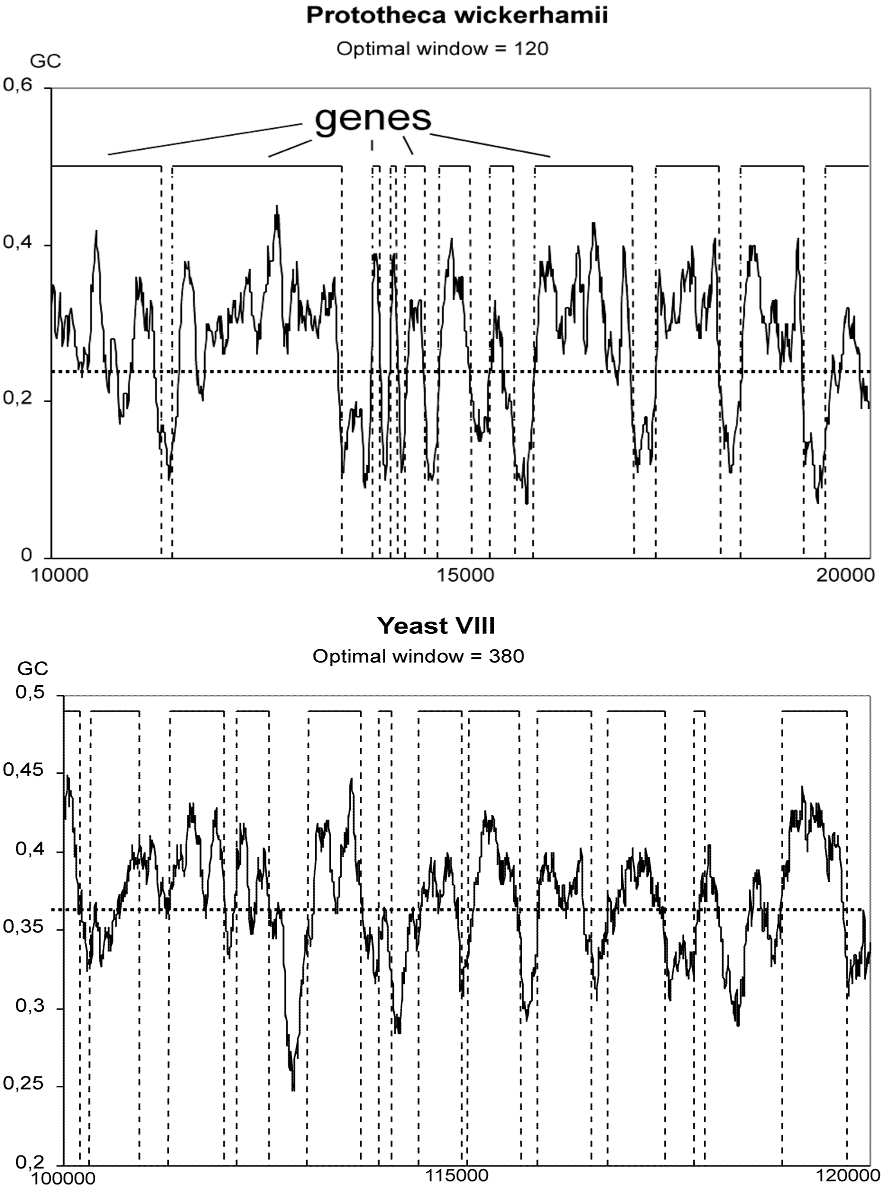

Figure 4: GC-concentration vs DNA nucleotde position

calculated by averaging on S_k, i.e. $A(S_k) = \frac{\sum_{i \in S_k} C(i)}{\sum_{i \in S_k} 1}$ and average value on the region of local binding energy that was calculated with optimal window width: $A_W(S_k) = \frac{\sum_{i \in S_k} A_{W_{opt}}(i)}{\sum_{i \in S_k} 1}$. Let's visualize then all S_k on the $A_W \times A$ plane (see fig. 5). It is obvious that the homogeneous regions of genes and junk are separated with confidence: genes are allocated in the area of large A and A_W, besides for them $A_W < A$, and the junk regions are allocated in the area of small A and A_W, and for them $A_W > A$, as a rule.

Finer tuning of the two-dimensional linear discriminate function gives the following separation errors: Prototheca wickerhamii - 0%, Plasmodium falciparum - 1,5%, Yeast Chromosome VIII - 5%.

It is worth noticing that for Prototheca wickerhamii genome it is possible to see the grouping of the points, corresponding to the regions that have a definite function. So, a separate "cloud" of points corresponds to the comparatively short ($\sim 70b$) tRNK-genes.

It is easy to understand the character of points separation in fig.5 if observe (see fig.4) that the genes are represented by the Λ-like form of graph, and the junk - V-like. Calculation of $A_W(S_k)$ included not only information about the region itself but also from flanking regions of the gene (with length about $W/2$). So, for a gene value A_W becomes smaller then $A(S_k)$, calculated for the region S_k only.

Other Measures. Reconstructed Frequencies.

For the window in binary sequence it is possible to calculate different integral characteristics. For example, one can consider the binary sequence to consist of the words made of units or zeros only. For example, for sequence $\{00011010001111\}$ the dictionary consist of words $\{\{0\}, \{00\}, \{000\},$
$\{1\}, \{11\}, \{111\}, \{1111\}\}$ (a three-zeros word may be considered as a separate word or a concatenation of shorter words).

The frequency of occurrence of $\{1\}$ word is the value of GC-concentration in the window. In addition we can choose as measures of the window frequencies of all words encountered with a definite length, *entropy* of such frequencies distribution, average length of the zero-words or unit-words and so on.

A more systematic approach is based on the reconstruction of the word frequencies using frequencies of other, shorter subwords.

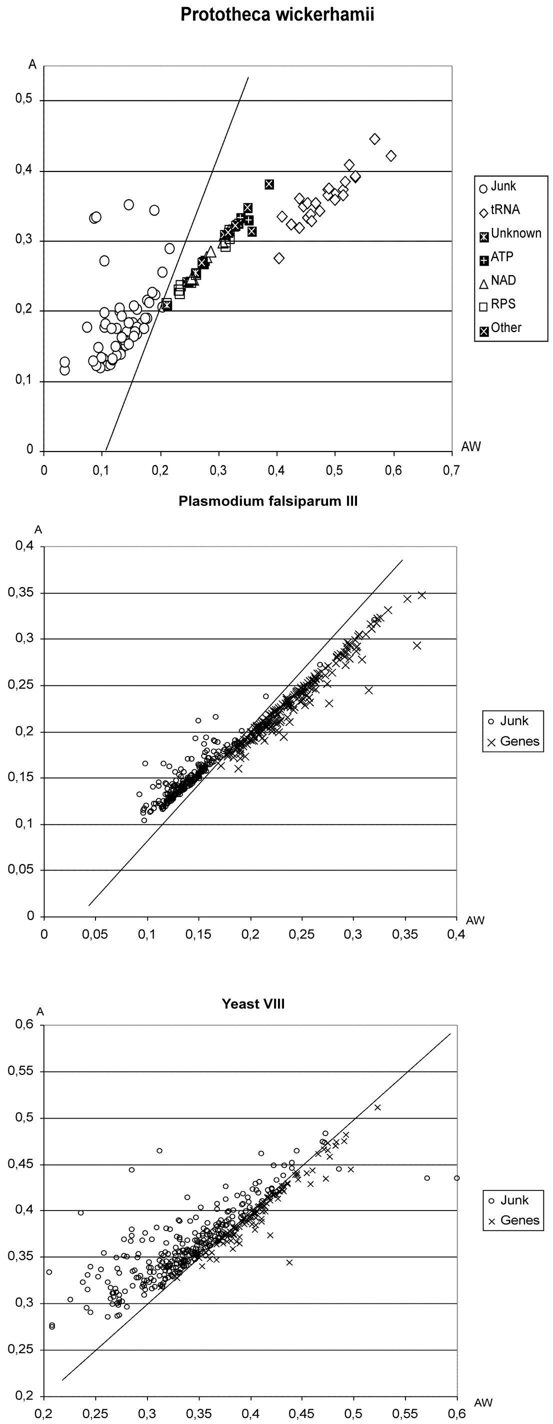

Figure 5: Analysis of the regions of junk and genes on the $A \times A_W$ plane

If the occurrence of the units is an independent random event then the frequency of the word $\{11\}$ should be equal to $\tilde{c}_{11} = (c_1)^2$. We denote by c_{11} the real value of frequency of the word $\{11\}$. Reconstructed frequency of the word c_{111} is equal to $\frac{(c_{11})^2}{c_1}$, that corresponds to reconstructed dictionary with maximal entropy (see Gorban, Popova, Sadovsky). For measures we can take relations between real and reconstructed frequencies - "reconstruction error", i.e.

$$d_k = c_k/\tilde{c}_k - 1,$$

where

$$d_{21} = c_1^2/c_{11} - 1, d_{20} = c_0^2/c_{00} - 1,$$

$$d_{n1} = \frac{c_{\underbrace{11..1}_{n}}}{\tilde{c}_{\underbrace{11..1}_{n}}} - 1 = \frac{c_{\underbrace{11..1}_{n}} c_{\underbrace{11..1}_{n-2}}}{(c_{\underbrace{11..1}_{n-1}})^2} - 1,$$

$$d_{n0} = \frac{c_{\underbrace{00..0}_{n}}}{\tilde{c}_{\underbrace{00..0}_{n}}} - 1 = \frac{c_{\underbrace{00..0}_{n-2}} c_{\underbrace{00..0}_{n}}}{(c_{\underbrace{00..0}_{n-1}})^2} - 1,$$

A set of such features for $n = 1..N$ describes differences between the given distribution of zeros and units and a random sequence making by multiple interchanges of the parts of initial sequence.

Unfortunately, possibilities of such a consideration are rather poor. The simplest consideration shows that statistical "noise" expected while calculating d_k is about

$$\Delta d_k = \frac{1}{\sqrt{W}} \sqrt{\frac{1 - \tilde{c}_k}{\tilde{c}_k}},$$

where W is the window length, where d_k is calculated, $\tilde{c}_k$ is the corresponding reconstructed frequency. For the window length of 1000 basepairs and $c_k \sim 0.4$ we get $\Delta d_k \sim 0.04$ and this value grows rapidly when $\tilde{c}_k$ decreases. So to reconstruct the frequencies of the words $\{111,000,1111,0000\}$ and longer, the window length about several thousands basepairs is not sufficient. Calculations made on several genomes showed that no significant differences of reconstructed frequencies from real were detected. More precisely, values of d_k do not exceed significantly the values of Δd_k.

6 Laplace Transform of GC-content Function

By averaging GC-content in a window with length W, we mean that the properties of the current position of DNA are determined by the whole region with length $W/2$ surrounding the point. In the case of simple averaging all basepairs in the window make equal contributions to the resulting sum.

In this section we will consider nonlocality by using integral Laplace transform of GC-content function $C(i)$. Recall that it is equal to 1 if in the i-th position of DNA we have GC-bond and 0 otherwise.

Let's introduce forward and backward Laplace transforms of $C(i)$:

$$f_i(p) = \sum_{j=i}^{N} C(j) \exp{(i-j)p},$$

$$b_i(p) = \sum_{j=1}^{i} C(j) \exp{(j-i)p},$$

where N is the length of the whole sequence. Parameter p defines effective length ($\sim 1/p$ basepairs) of the region that has influence on the properties of DNA in the i-th position.

Let's consider Taylor series of $b_i(p)$ and $f_i(p)$ functions at point $p = p_0$:

$$f_i(p) \approx f_i(p_0) + \sum_{k=1}^{n_t} f_i^{(k)}(p_0)(p-p_0)^k,$$

$$b_i(p) \approx f_i(p_0) + \sum_{k=1}^{n_t} b_i^{(k)}(p_0)(p-p_0)^k,$$

where

$$f_i^{(k)}(p_0) = \frac{1}{k!} \sum_{j=i}^{N} (i-j)^k C(j) \exp{(i-j)p_0},$$

$$b_i^{(k)}(p_0) = \frac{1}{k!} \sum_{j=1}^{i} (j-i)^k C(j) \exp{(j-i)p_0}.$$

If we calculate values of $f_i(p_0), f_i^{(k)}(p_0), b_i(p_0), b_i^{(k)}(p_0)$ in every position of sequence, then the following recurrent formulas are useful (they make the task of calculating the functions N-linear by the number of computer operations):

$$f_{i-1}(p) = \exp{(-p)} f_i(p) + C(i-1),$$
$$b_{i+1}(p) = \exp{(-p)} b_i(p) + C(i+1),$$
$$f_{i-1}^{(k)}(p) = \exp(-p) \sum_{j=0}^{k} C_j^k f_i^{(k)} (-1)^j,$$
$$b_{i+1}^{(k)}(p) = \exp(-p) \sum_{j=0}^{k} C_j^k b_i^{(k)} (-1)^j,$$

where C_j^k are binomial coefficients. In these formulas the forward function is calculated in the left direction to avoid roundoff errors accumulation.

Graphs of equal scale adjusted values of $f_i^{(k)}$ and $b_i^{(k)}$ with $p_0 = 0.01$, calculated for a fragment of Prototheca wickerhamii sequence, are shown in fig. 6.

It is apperent that Taylor coefficients reveal strong correlation with each other and dependencies of i become smoother for higher k and "shift" to the left for the forward function and to the right for the backward.

We can consider the values of $f_i(p_0), f_i^{(k)}(p_0), b_i(p_0), b_i^{(k)}(p_0)$ as multidimensional coordinates of the i-th DNA position. Since the values do not change considerably with every step by i then in the multidimensional space of the Taylor coefficients we have almost continuous trajectory parametrized by i. We can get visual presentation of the curve by viewing it projected in the three-dimensional subspace of principal vectors of distribution of points along the curve. The resulting pictures are shown in fig.7. The trajectories are rather tangled spiral-like curves. It points out to the presence of some periodicity in averaged GC-content.

7 Mixing Entropy

In section 3 more than twenty different characteristics were described that are used in statistical DNA analysis for content genes search. It is worth noticing that not all of these features have clear physical meaning.

Note one more time, that almost always the decision rules to be constructed on the features are learned on some training set of coding and non-coding sequences and then applied to the new ones.

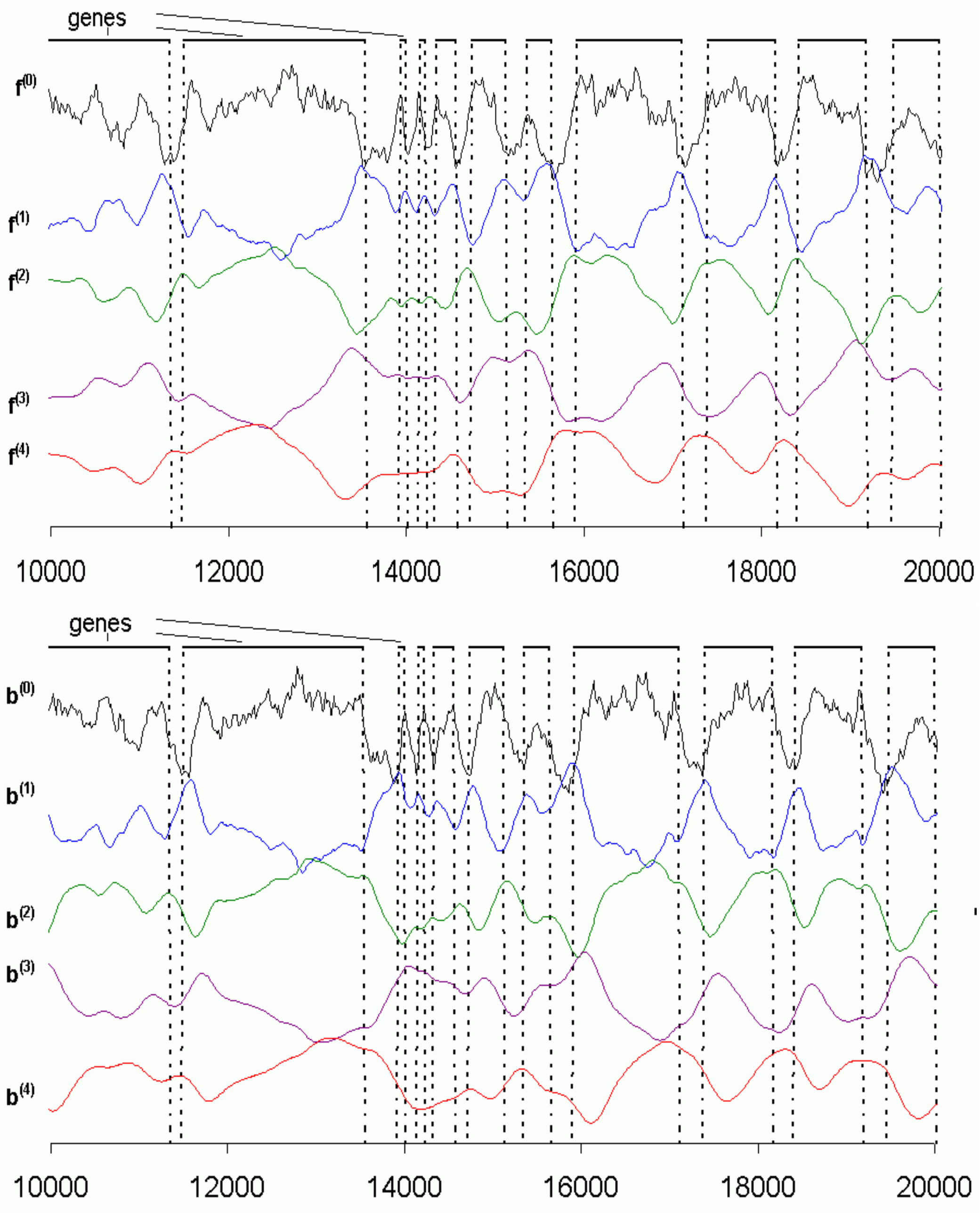

Figure 6: Coefficients $f_i^{(k)}(p_0)$ and $b_i^{(k)}(p_0)$ for fragment of Prototheca wickerhamii genome

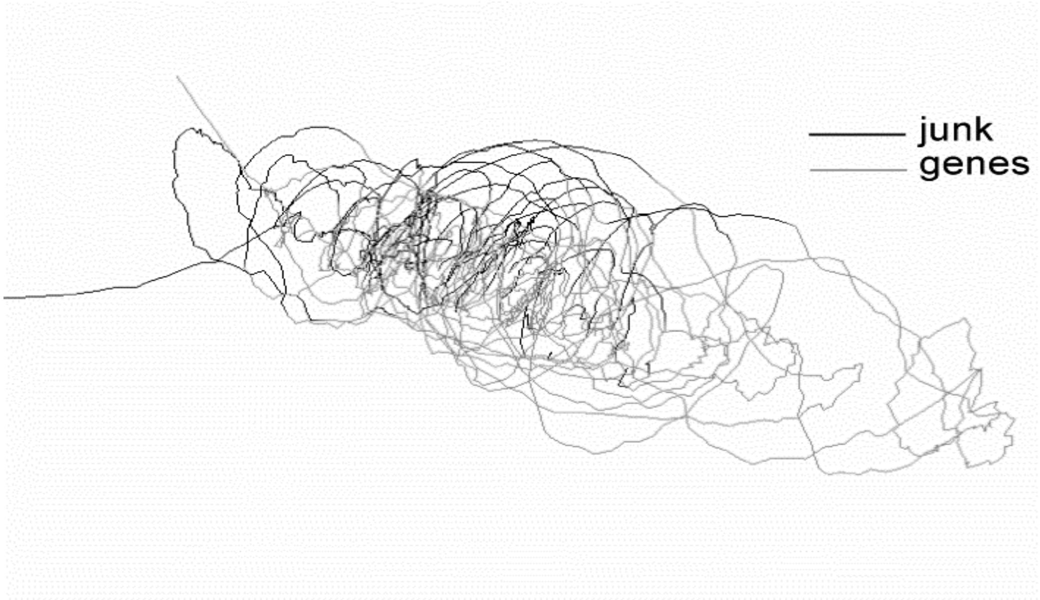

Figure 7: Trajectory of DNA in multidimensional space of Taylor coefficients of GC-content Laplace transform

In this section we will consider one more feature that we called *mixing entropy*. This feature has clear interpretation and holds promise for possibility to build gene identification procedures that need no training phase.

The information that defines the order of aminoacids in protein are coded in DNA by codons - triplets of nucleotides. Note that this coding is excessive: from 64 possible codons 61 may code 20 aminoacids (three definite codons are stop-signals for translation process). Some aminoacids (methionine, tryptophan) are coded by a single codon, other can be coded by one of the two, three and even six codons.

It was noticed long ago that there is a kind of discrimination in using codons. Some codons are used more frequently that their synonyms in aminoacids coding.

Several authors (see Beranola) used this fact to construct procedures of finding borders of exons by comparing entropies of distribution of codons on the left and on the right side of the border. Actually the whole sequence in these works was divided on the homogeneous parts.

We introduce here a formally similar approach but it seems that to use the concept of codons discrimination is not necessary for protein coding exons identification in DNA. Actually we make an accent on the idea of *distinguished phase* in distribution of codons.

If we take an arbitrary window of coding sequence (without introns) and divide it into successive non-overlapping triplets, starting from the first base pair in window, then this decomposition and arrangement of the real codons may not be *in phase*. We can divide the window into triplets in three ways, shifting every time on one base pair from the beginning. So we have three triplet distributions and one of them coincides with the real codons distribution. So the protein coding region are characterized by the presence of distinguished phase.

Junk evidently has no such property because, as it was mentioned above, inserting and deleting the base pair in junk do not change properties of DNA considerably, thus this kind of mutations is allowed in the process of evolution. But every such mutation breaks the phase, so we can expect than distributions of triplets in junk will be similar for all three phases.

Fig. 8 demonstrates that it is really true. In this figure distributions of triplets in three phases are compared for the sequence of randomly selected gene *nad4* of Prototheca wickerhamii (GenBank U02970) and for a junk sector from the same genome. All possible triplets are enumerated and the number of the triplet is X-coordinate with Y is its frequency.

It is worth noticing that the distribution of frequencies is quite non-uniform. There are eight sharp peaks in the junk‘s distribution which correspond to the triplets whithout G and C letters. It is because of the high AT-richness in this piece of junk.

So if we mix all three distributions together then the summary triplet distribution in junk will be similar to all of three distributions. But in case of protein coding region we will get more uniform distribution with greater entropy.

Let’s characterize the sliding window by value of mixing entropy:

$$ME = \frac{1}{3}(3S - S^{(0)} - S^{(1)} - S^{(2)}),$$

where $S^{(k)} = -\sum_i (f_i^{(k)} \ln(f_i^{(k)}))$ is entropy for the triplet distribution with phase k, S is the entropy of the mixed distribution, i iterates through all possible codons.

To calculate the optimal window size where the triplet frequencies are calculated we scanned through window lengths just as we did it to find optimal window of GC-averaging. The resulting graphs are shown in fig. 9.

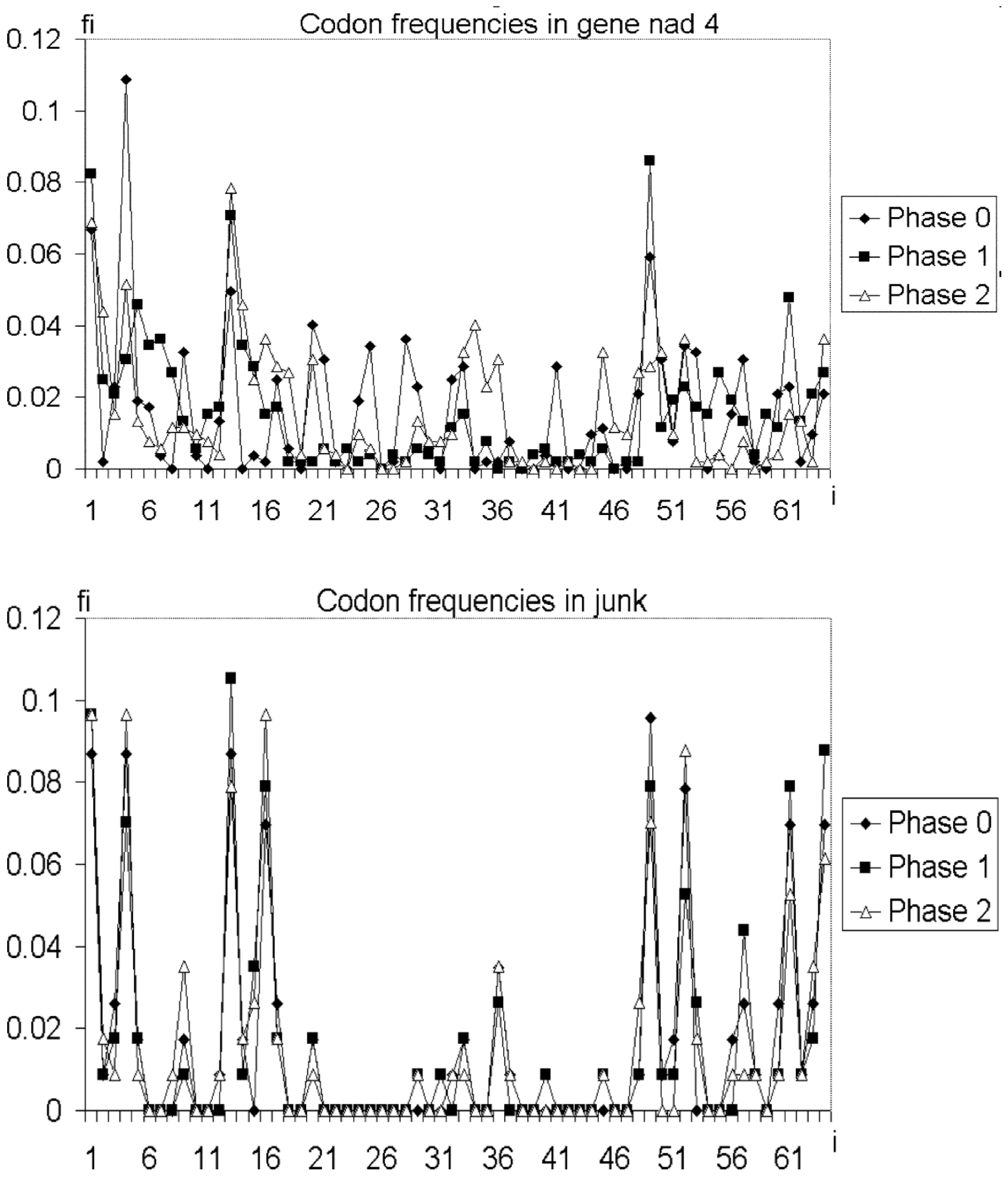

Figure 8: Comparison of triplet distributions for junk and gene regions. All codons are enumerated and X-axis is the codon number, Y is its frequency.

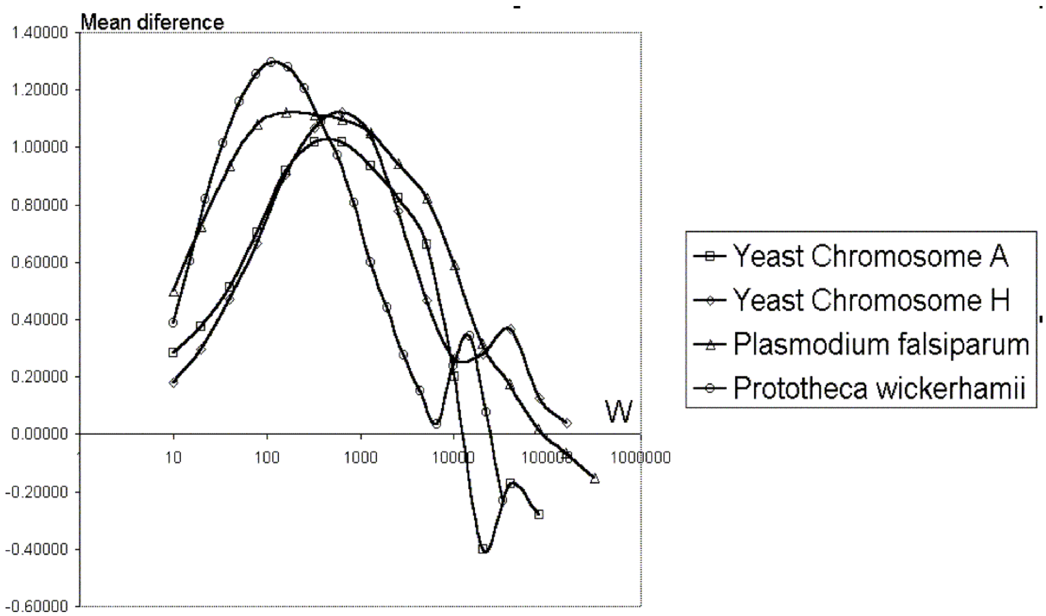

Figure 9: Normalized mean difference in mixing entropy values for junk and genes vs window width

It turns out that the value of ME is more suitable for separating genes and junk compared to the value of GC-content.

Graphs of the ME values along a sequence are shown in fig. 10.

8 Visualizing Triplet Distribution and Self-Training Procedure of Exon Identification

In this section we investigate distribution of frequencies of using triplets in a window sliding along the whole sequence. Visualizing of the distribution of frequencies in multidimensional space will allow us to formulate the procedure of gene identification without knowing anything about a new sequence.

We analyzed DNA as a single strand (without distinguishing W- and C-strand separately). The sliding window was divided into successive non-overlapping triplets, starting from the first base pair in the window (triplets in 0-phase) and frequencies of all triplets were calculated. So, every base pair is characterized by a 64-dimensional vector of frequencies. For our experiments we took every 12-th base pair in the case of short mitochondrial DNA and

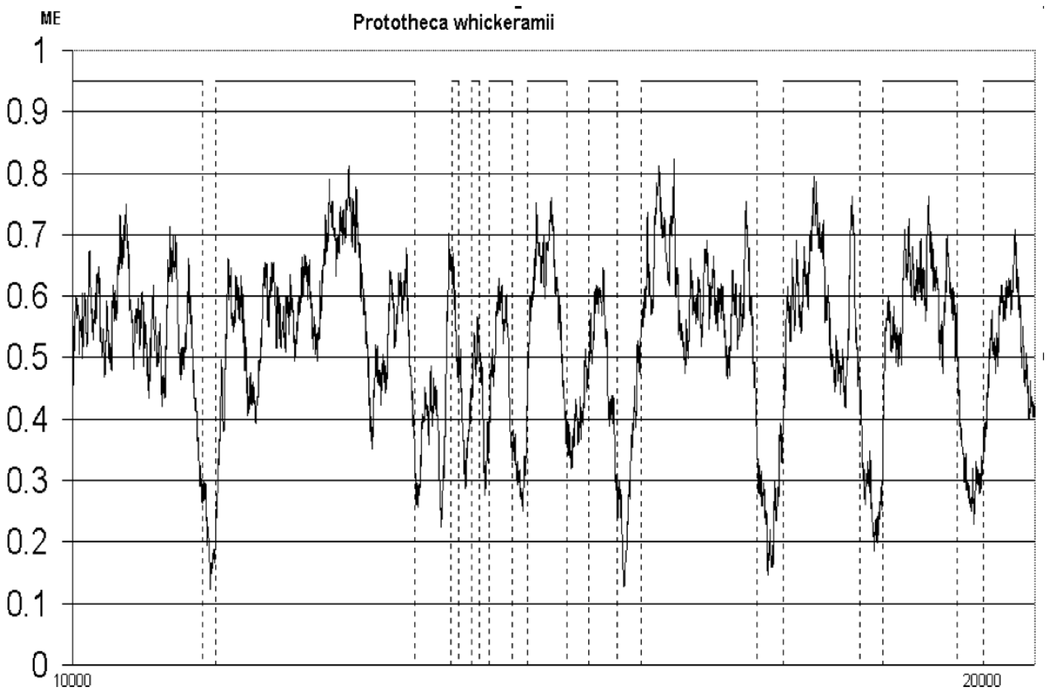

Figure 10: Mixing Entropy vs DNA nucleotde position

every 60-th base pair in the case of longer sequences of chromosomes of the Yeast genome.

As a result we have a set of multidimensional data points $x_i, i = 1...N$ in the space of frequencies. The data were centered and normalized on the unit standard deviation of every coordinate:

$$\widetilde{x}_i^k = \frac{x_i^k - \bar{x}^k}{\sigma_k},$$

where x_i^k is value of the k-th frequency characterizing the i-th base pair, $\widetilde{x}_i^k$ is normalized value, σ_k and $\bar{x}^k$ are standard deviation and mean value of the k-th coordinate.

To visualize the set of data points three principal vectors were calculated. Principal vectors are eigen vectors of covariance matrix of data distribution. In these directions dispersions of the cloud of data points reach their maximums. After orthogonal projecting in the subspace of the three principal vectors we can visually represent the cloud of data points.

The resulting pictures of data (plan view and side view) are shown in figures 11,12. It is evident that the distribution has 4 clusters. Central cluster is junk distribution and other three are distributions of protein coding regions in three different phases.

Taking into account all mentioned above, the pictures are quite understandable. Since the junk has no distinguished phase, it is represented by almost normal distribution situated in the center of the data point cloud. Protein coding regions with different phases forms three wings on the sides of the junk kernel.

Using this representation we may formulate the procedure of determining whether the base pair belongs to a protein coding region or not.

Assume that we know nothing about exon allocations. Nevertheless we may construct the data point cloud in the space of frequencies and perform clustering on 4 clusters. We do not consider here different methods of clustering. In our work we used simplest clustering using iterative algorithm of dynamical centroids with the distance from a point to a cluster as a distance to its geometric center. Four clusters correspond to the four types of homogeneous in a certain sense regions of genome.

The resulting pictures of clustering are shown in fig. 13.

Then we determine whether the base pair belongs to the junk or coding region in the definite phase just by calculating the distance of the point in

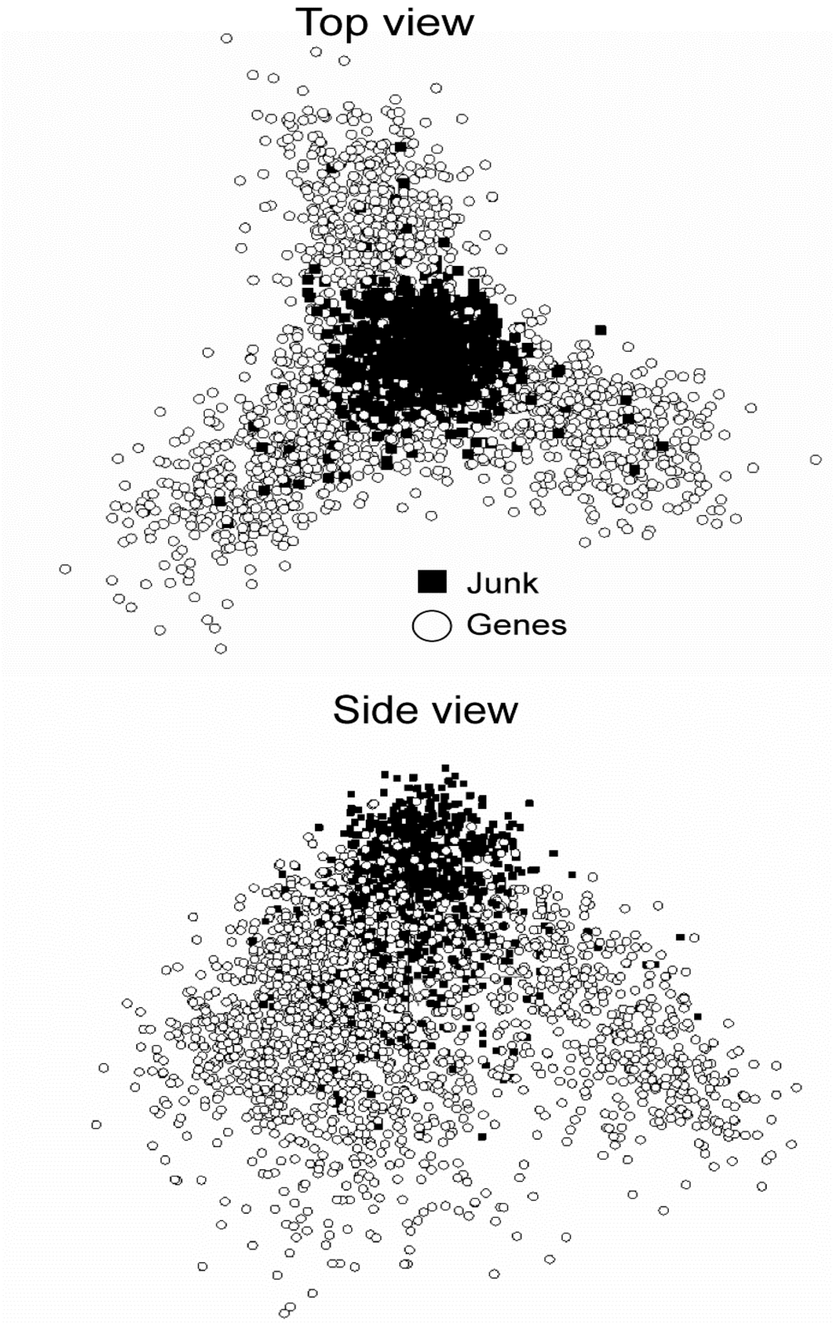

Figure 11: Visualizing Triplet Distributions For Prototheca wickerhamii genome

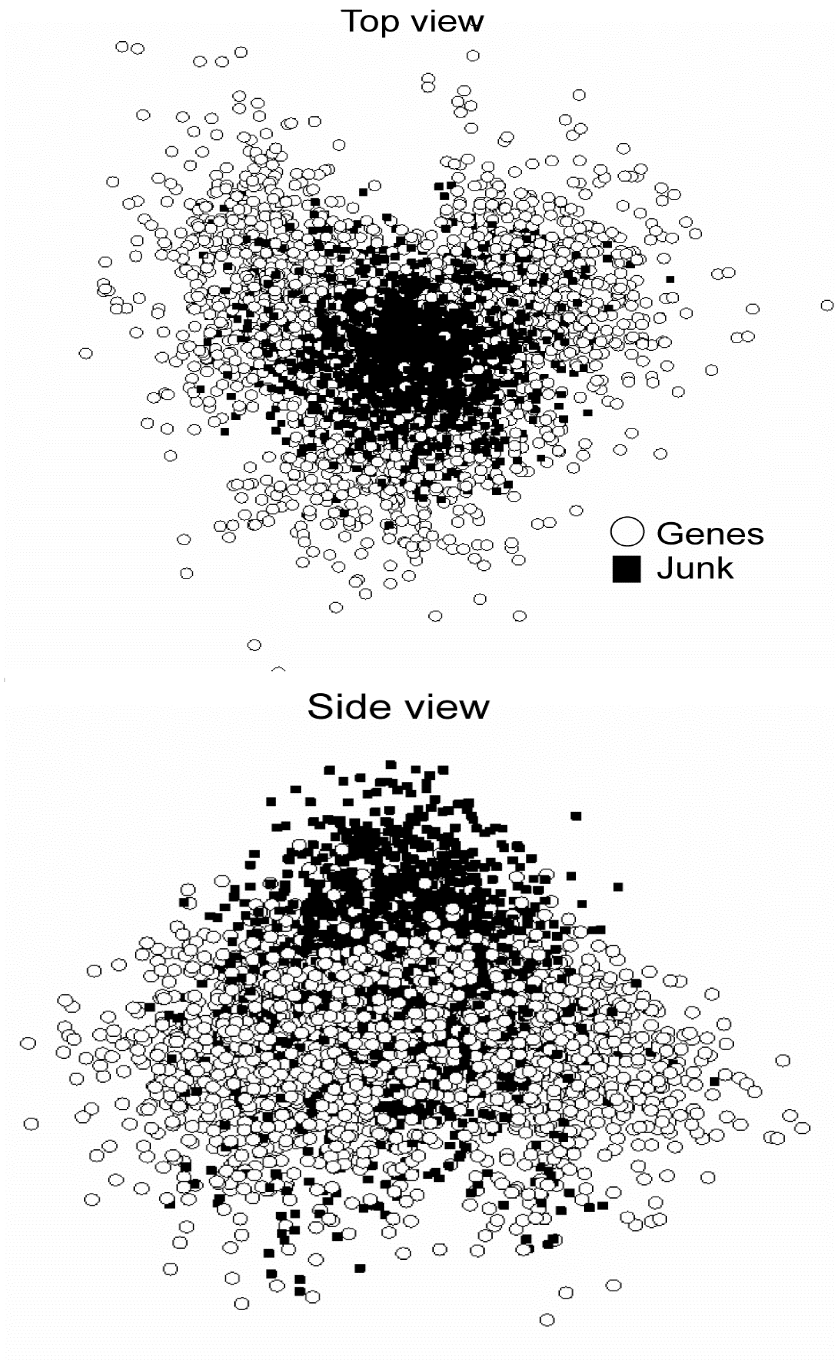

Figure 12: Visualizing Triplet Distributions For Yeast Chromosome III

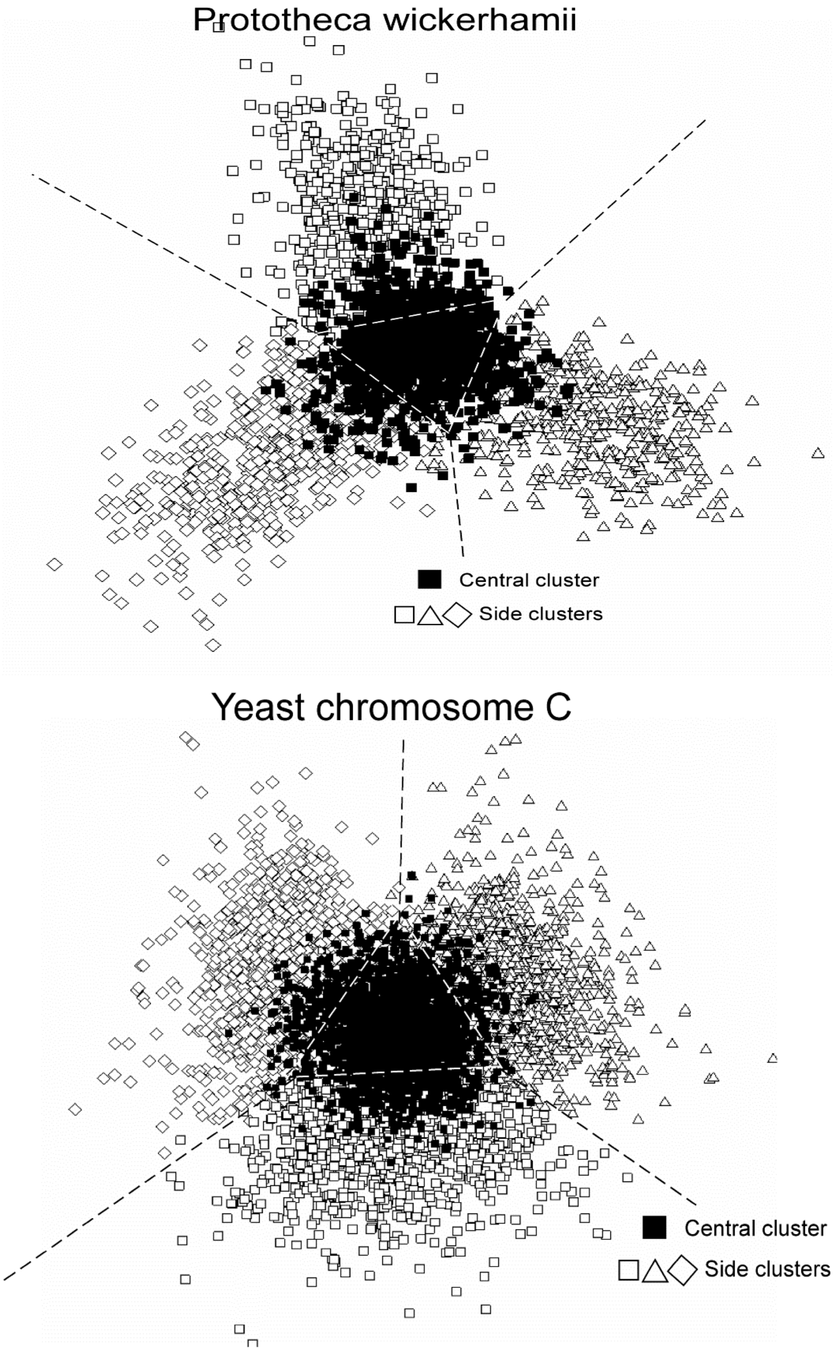

Figure 13: Clustering Triplet Distributions

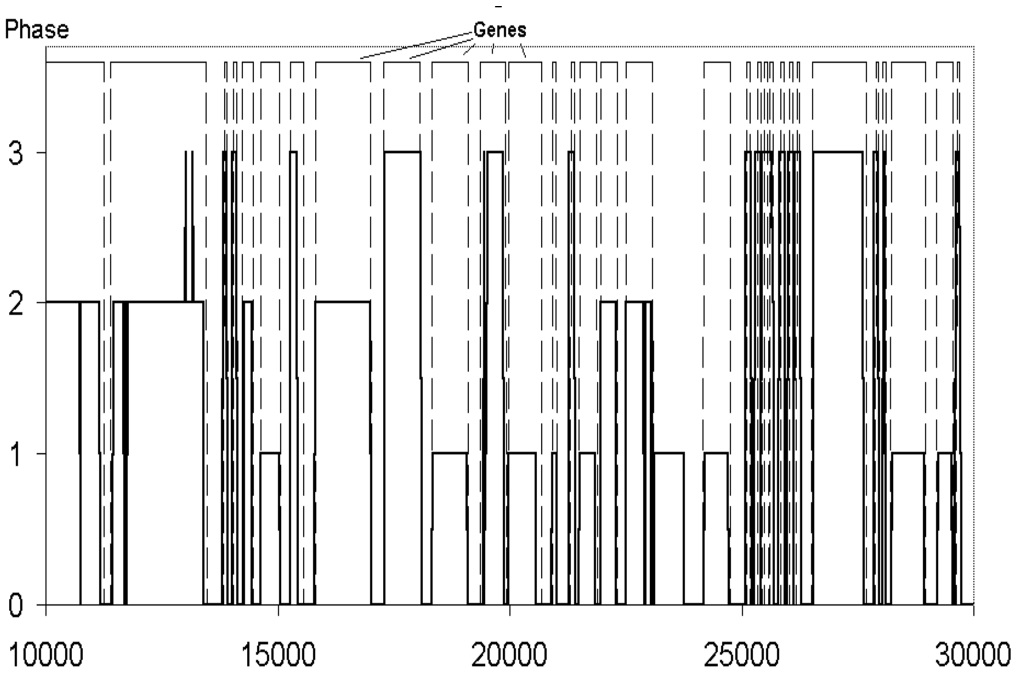

Figure 14: Exon predictions for Prototheca wickerhamii genome

frequencies space to all four clusters and choosing the closest one. We did it for analyzed sequences and compared with real exon allocations. Accuracy of point determination turns out to vary from 60% of all base pairs to 82% in case of "well-clustered" Prototheca wickerhamii genome. Note here that this is accuracy of classification of points (does the basepair in the position belong to an exon or not).

Fragments of resulting graphs are shown in fig.14,15. In these figures dashed line denotes borders between real genes given in the annotation. Besides in the case of Yeast Chromosome III different heights of the bars correspond to the different reliabilities of the gene presence. The highest bars correspond to the surely reliable genes. The solid line is the graph of phase (cluster number) of triplet distribution in the window centered in the current position. It is clear that base pairs, which belong to the same exon, tend to have equal phase in the window.

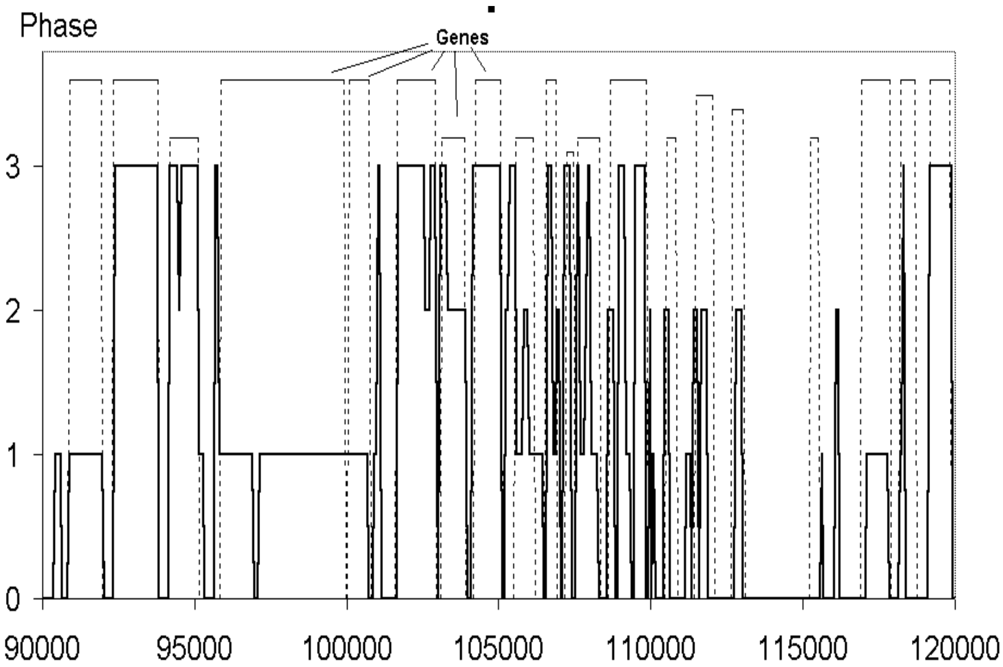

Figure 15: Exon predictions for Yeast Chromosome III

9 Discussion: Codon Usage, GC-content, Hexamers and other

Lets try to make a summary of the results of our simple experiments.

First consider the ancient codon usage measure. Methods of pattern recognition applied to the gene identification is based on the belief that in coding regions there is some specific codon composition compared to junk regions due to the phenomenon of codon discrimination in coding amino-acids.

But we could see that the form of distribution of 64-dimensional vectors, each of them corresponds to the codon usage measure calculated in a winndow of DNA, has amazing bullet-like structure. The structure can be explained by the notion of existence of distinguished phase in the triplet distribution in coding regions. In this space points which correspond to junk and exons can be separated by a linear function, but this separation is not the most perfect one.

From other hand, the level of GC-content (or local binding energy) evidently has influence on the codon distribution (when GC-content is low, the codons AAA, AAT, ATA, ATT, TAA, TAT, TTA, TTT have greater frequency than others). Actually, GC-content function is linear functional in the 64-dimensional space of codon frequencies and its gradient is a distinguished direction in the space along what the "exon-junk" separation is quite good:

$$GC\% = \sum_{i=1}^{64} \alpha_i f_i,$$

where i enumerates all codons, f_i is frequency of $i-$th codon, α_i is the fixed weight of the codon (for example, $\alpha_{AAA} = 0, \alpha_{GAT} = \frac{1}{3}, \alpha_{GCA} = \frac{2}{3}, \alpha_{GCG} = 1$ and so on). It seems that linear discriminant function separating junk windows and exon windows in 64-dimensional space of codon frequencies should have similar weights.

Second, it is known that inphase hexamers measure (described in section 3) is the best single measure for separating coding and non-coding regions. It is interesting that the idea of using inphase measures is very close to the idea of using mixing entropy. Actually, many other measures (Entropy, Assymetry etc.) use the difference in distribution of triplet decomposition with three different phases. But it seems that none of them use the idea in the explicit form, except inphase hexamers.

Actually it is easy to see that inphase codon usage measure can be derived from hexamer usage measure. Let‘s take distributon $f^{(0)}$ of triplets in 0-phase, and the distribution $h^{(0)}$ of pairs of these triplets. Then frequencies of triplets in 1-phase and 2-phase equal

$$f^{(1)}_{XYZ} = \sum_{I,J,K} h^{(0)}_{IXYZJK},$$

$$f^{(2)}_{XYZ} = \sum_{I,J,K} h^{(0)}_{IJXYZK},$$

$$X, Y, Z, I, J, K \in \{A, C, G, T\}.$$

So, distributions of triplets in any phase can be derived from the distribution of pairs of triplets in one phase.

Another important note is that codon usage in the one DNA strand evidently defines codon usage in the opposite strand. The notion of distinguished phase does not allow to determine in what strand predicted exon is situated. It is interesting that most programs analyse two strands separately, though the measures in the complementary regions are not independent. We think that this question needs in more detailed consideration.

The last note concerns normalizing codon usage in our experiments on the unity standart deviation of every frequency. It means that we take into account not the absolute value of the variances but the relative ones. Our experiments shows that it is crucial for existence of bullet-like structure in the space of codon frequencies.

In the overview made by J.-M. Claverie (1997) three general problems of current computational gene identification methods were underlined: A) the most of the methods detect only protein coding exons; B) most of the methods work with a piece of sequence containing only one gene; C) most of the programs use methods of pattern recognition with learning with teacher - they need a traning set for tuning their parameters. We hope that the next generation of gene-finders will overcome the last two problems, may be using similar ideas as described in this paper. We hope that such physically clear characteristics as local binding energy and related measures (see, for example, Yeramian E., 2000) can help to solve the A problem although.

10 Resume

In the paper we tried to touch upon the problem of automated gene identification considering some recent statistical approaches such as calculations of DNA stability map and entropy segmentation method. These approaches are rather specific because 1) they do not require the preliminary stage of learning on the known sample database, 2) the methods are applicable mainly to the whole DNA sequence rather than its separate fragments.

Considering calculations of DNA thermal stability we believe that the complicated procedure of calculation of the partition function gives results which can be compared by efficiency for gene finding with the much simpler value of local binding energy, calculated by averaging GC-concentration with the window of some optimal width. The width was evaluated ($\sim$ 400bp for long genomes, and $\sim$ 120bp for the short mitochondrial one) and this revealed some interesting regularities. Providing the borders of hypothetical genes are known (say, using analysis of start and stop codons), more than 90% of genes may be identified using only two simple features. Several approaches for further investigations have been proposed.

Discussing such ancient sequence measures as Codon Usage, Assymetry, Entropy, Inphase Hexamers and the recent method of entropy segmentation we thought that there is one common principle that underlies all of them and the principle is simply the fact that the genes are carriers of biological information coded by codons. The phase of the coding stands out among all possible partitions of a gene onto consecutive non-overlapping triplets and the phase should be strictly conserved in the process of evolution. The junk lacks this feature because even if there was long time ago such a phase then it was broken in the process of evolution because mutations that did it were allowable.

Visual confirmation of the hypothesis is the picture of multidimensional distribution of windows of DNA in the triplet frequencies space. Viewing the picture we formulated the procedure of exon identification which has the same specificity that two mentioned methods have. The procedure is self-training (uses learning without teacher): it does not require any additional training set and is based on the good DNA clustering in the triplet frequencies space due to the presence of the distinguished coding phase. The measured quality of the clustering is simultaneously an evaluation of the method accuracy. The pictures of superposition of genetc map and graph of coding phase are rather convincing.

References

[1] *Almirantis Y.*. A Standard Deviation Based Quantification Differentiates Coding from Non-coding DNA Sequences and gives Insight to their Evolutionary History. J. Theor. Biol.(1999), V.196. pp.297-308.

[2] *Bernaola-Galvan P., Grosse I., Carpena P. and others.* Finding Borders between Coding and Noncoding DNA Regions by an Entropic Segmentation Method. Phys.Rev.Letters (2000), V.85, N.6.

[3] *Burge C.B., Karlin S.*. Finding the genes in genomic DNA. Current Opinion in Structural Biology, 1998. No.8. pp.346-354.

[4] *Carbone A., Gromov M.*. Mathematical Slices Of Molecular Biology. IHES Preprint, IHES/M/01/03, 2001.

[5] *Claverie J.-M.* Computational methods for the identification of genes in vertebrate genomic sequences. Human Molec. Genetics 6. pp. 1735-1744 (1997).

[6] *Fickett J.W.*. The Gene Identification Problem: An Overview For Developers. Computers Chem.,1996. Vol.20, No.1, pp.103-118.

[7] *Frank-Kamenetskii M.D., Frank-Kamenetskii A.D.*. (1969) Mol. Biol. 3, pp. 295-301.

[8] *Gorban A.N., Popova T.G., Sadovsky M.G.* Classification of symbol sequences over their frequency dictionaries: towards the connection between structure and natural taxonomy. Open Sys.& Information Dyn. 7, p.1-17,(2000).

[9] *Jacobson H., Stockmayer W.*. J. Chem. Phys. 18, 1600. (1950).

[10] *Rogic S., Mackworth A.K., Ouellette F.B.* Evaluation of Gene-Finding Programs on Mammalian Sequences. Genome Research. Vol. 11, Issue 5. pp. 817-832 (2001).

[11] *Searls D.B.*. Bioinformatics Tools For Whole Genomes. Annu.Rev.Genomics Hum.Genet, 2000. No. 01. pp. 251-279.

[12] *Seely O.Jr., Feng D.-F., Smith D.W., Sulzbach D., Doolittle R.*. (1990) Genomics 8, 71.

[13] *Wada A., Yabuki S., Husumi Y.*. CRC Crit. Rev. Biochem. 9, (1980). pp. 97-144,

[14] *Yeramian E.*. Genes and the physics of the DNA double-helix. Gene 255 (2000). pp. 139-150.

[15] *Yeramian E.*. The physics of DNA and the annotation of the Plasmodium falsiparum genome. Gene 255 (2000). pp. 151-168.

[16] *Yeramian E., Schaeffer F., Caudron B., Claverie P., Buc H.*. An optimal formulation of the matrix method in statistical mechanics of one-dimensional interacting units: efficient iterative algorithmic procedures. Biopolymers, Vol.30. 1990. pp.481-497.

[17] *Yeramian E., Claverie P.* (1987) Nature 326, pp. 169-174. Biol. 3, pp. 295-301.

[18] *Zhang M.Q.*. Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc. Natl. Acad. Sci. USA, 1997. Vol. 94, pp. 565-568.

[19] *Zhang M.Q.*. Statistical features of human exons and their flanking regions. Human Molecular Genetics, 1998, Vol. 7, No. 5. pp. 919-932.