# RECOGNIZING FUNCTIONAL DNA SITES AND SEGMENTING GENOMES USING THE PROGRAM "COMPLEXITY"

*[1*] Orlov Yu.L., [2] Potapov V.N., [1] Filippov V.P.*

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: orlov@bionet.nsc.ru
[*]Corresponding author

***Key words:*** *sequence analysis, complexity, data compression, complete genomes, suffix tree visualization, variable memory Markov model*

## Resume

*Motivation:* Large-scale sequencing of genomes opens new opportunities for analysis of long nucleotide sequences. Of special interest is the detection of common contextual properties that are stable to evolutionary changes in terms of the genome. The next problem is segmentation of genome sequences on the basis of these context properties.

*Results:* A program that constructs Markov models with variable memory for generating genetic texts was developed. Using these variable memory Markov models, a method for recognizing functional DNA sites was developed and tested on promoters containing TATA-box sites. A method for segmentation of genomic sequences was developed using an estimated probability of observing a certain region taking into account its local contexts.

*Availability:* http://wwwmgs.bionet.nsc.ru/programs/complexity/.

## Introduction

Large-scale projects of sequencing of whole microbial and eukaryotic genomes open promising opportunities for comparison of DNA sequences in different organisms. Of special interest is the analysis of most general characteristics of genome sequences. Information measures of symbol sequences exemplify such general characteristics (Haring, Kypr, 1999).

The algorithm proposed, basing on methods of the data compression theory, allows construction of the variable memory Markov model to generate sequences (Orlov, Potapov, 2000; Orlov et al., 2002). The model of text generation is represented unambiguously by suffix trees (Ron et al., 1996).

"Complexity", an Internet-accessible software tool (http://wwwmgs.bionet.nsc.ru/programs/complexity/), generates probabilistic suffix trees (PST) for a specified nucleotide sequence. The representation of the model as a generation tree source in a GIF format facilitates visual comparison of the inner structure of the texts.

## Methods and Algorithms

Let us consider a stationary stochastic grammar model of text generation. The model $T$ generates the sequence $X^n = X_1X_2...X_n$ with a probability: $P(X^n) = P(X_1|S_1)P(X_2|S_2)...P(X_n|S_n)$. The probability $P(X_n|S_n)$, which is independent of the position of $n$ in the sequence, is determined only by the preceding context $S_n$. The probability $P(D_i|S_j)$ of occurrence of a certain letter $D_i \in \{A,T,G,C\}$ $(i = 1,2,3,4)$ in each of the contexts $S_j$ is determined as follows by the corresponding parameters of the distribution $\theta$:

$P(D_i|S_j)=\theta^i_j$,

where $\sum_{i=1}^{4} \theta^i_j = 1$, $j = 1,2,...|T|$, $|T|$ is the total number of contexts in the model $T$ (number of leaves in the tree), and $4 \le |T| \le 4^k$ ($k$ is the maximum context length).

Equivalent states of the Markov model corresponding to contexts of various lengths can be integrated using the algorithm developed for constructing contextual probabilistic tree sources (Orlov, Potapov, 2002; Orlov et al., 2002), which employs algorithms previously developed within the source coding and data compression theories (Barron et al., 1998).

A context tree can be represented graphically; the contexts might vary in length, and none of them is the end of another (Fig. 1).

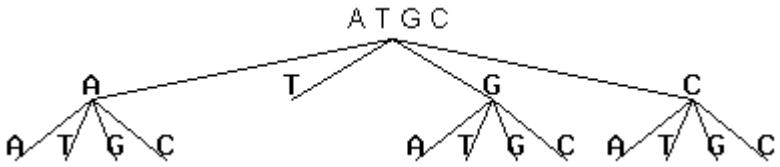


**Fig. 1.** Example of a generating tree source. The tree is constructed using the program Complexity for nucleotide sequences of the AP-1 transcription factor binding site (http://wwwmgs.bionet.nsc.ru/mgs/dbases/nsamples/). Each link from the leaves to the root corresponds to a context in a DNA sequence and has its own set of probabilities of generating the next symbol.

For example, in the tree $T$ shown in Fig. 1, there are twelve contexts with a length of two nucleotides (AA, TA, GA, AG, TG, GG, CG, AC, TC, GC, and CC) and one context with a length of one nucleotide (T). Totally, there are 13 preceding contexts, i.e., $|T| = 13$. Each context specifies four numbers: the probabilities of generating symbols located to the right of this context. To determine these numbers, we use frequencies of the corresponding oligonucleotides that are by one nucleotide longer: totally, we have $4 \times 13 = 52$ numbers.

## Implementation and Results

Construction of generating models for extended DNA sequences

Earlier, samples of nucleotide sequences of several functional classes were analyzed using the method proposed (Orlov, Potapov, 2000). Analysis of the nucleotide sequences of various functional classes (coding, noncoding, and regulatory regions) shows that DNA sequences have tree sources of various structures. The models differ in the order of the Markov chain, the tree structure, and number of branches (Orlov et al., 2002) (see Figs. 1 and 2). Figure 2 shows that models can be more complex than a model shown in Fig. 1.

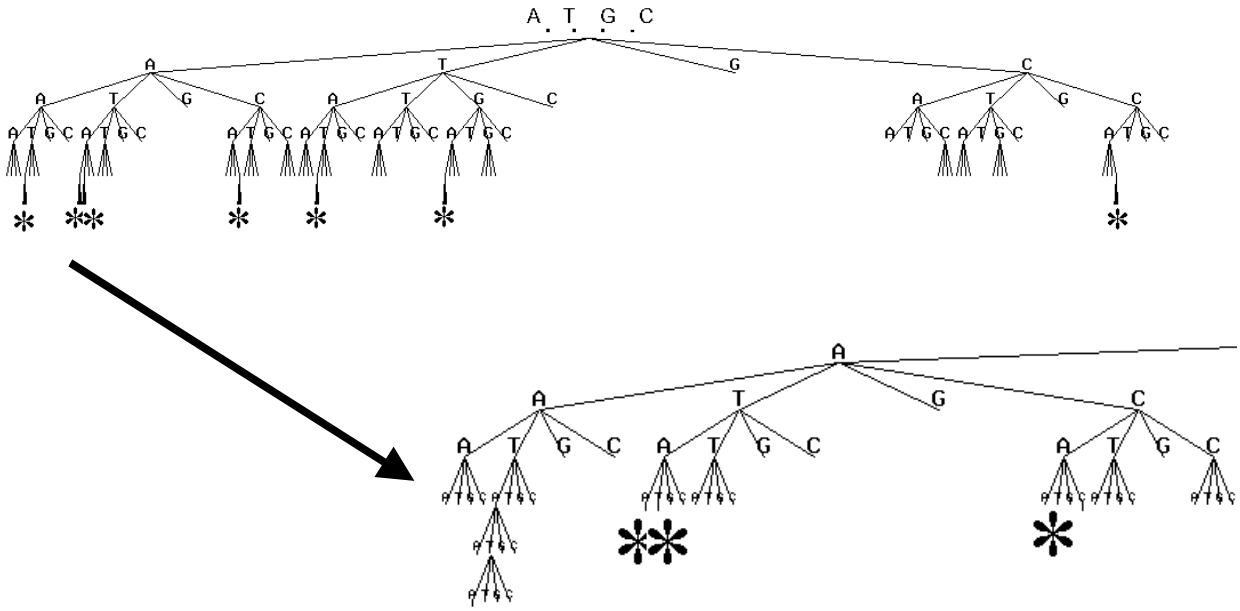


**Fig. 2.** Context tree source for nucleotide sequences containing TATA box. (534 sequences were extracted from the TRRD database, the sequence length was 20 bp). The letters for preceding contexts with a length of more than 3 bp are not shown because of a limited space in the scheme; an asterisk (*) denotes contexts more than 5 bp in length. Below, we give an enlarged fragment of the same tree, which is bordered with a gray line. Long contexts marked with the first asterisk from the left are shown in full: ATATAA, TTATAA, GTATAA, and CTATAA.

There is a technical problem of a simultaneous graphic representation of all contexts. It is impossible to arrange contexts of a length of more than 3–4 nucleotides (totally, 64–256 possible contexts) on a standard page or on a computer screen; therefore, an image is given iteratively.

The model developed allows the assessment of the degree to which certain sequences are close to a specified sequence in terms of their local contexts. These assessments can be used to (1) recognize relatively short regions in long sequences and (2) segment long DNA sequences into regions that differ in their contextual compositions.

Recognition of functional sites

Let us consider the problem of recognition of functional regions in extended sequences. In this case, the model is constructed not for one sequence but for a sample of functional regions. For a sample of sequences, a probabilistic tree source with a set of probabilities of generating symbols from the DNA alphabet is calculated taking into account that several first leftward symbols forming the first context are missing. A tree source was constructed for the sample of TATA-box sequences from the TRRD database (see Fig. 2). Then, promoter sequences from the TRRD database were analyzed. The probability of obtaining this region was calculated by a sliding window of 20 bp and the logarithmic profile for such a probability with the minus sign was constructed (Fig. 3). The minimal value of the profile is in the region with TATA boxes that are most typical of this model. The minimal profile value for the metallothionein-I gene promoter (AC EMBL: J00605) corresponds to an actual TATA box indexed in the TRRD database at a distance of –28 to –23 bp before the transcription start (denoted by the arrow in Fig. 3).

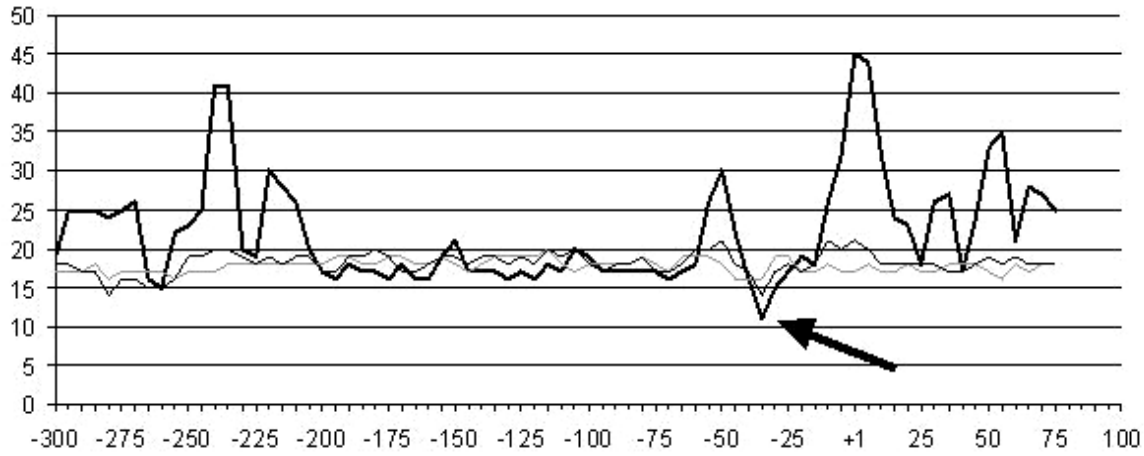


**Fig. 3.** Profile of the TATA box recognizing function in the promoter region [-300;+100] of the metallothionein-I gene. The solid curve refers to recognition in the tree source model. The gray and thin black curves refer to recognition by the nucleotide and dinucleotide frequencies in the same sliding window.

As is evident, the profile of the recognition function in the variable memory Markov model shows a more prominent TATA-box region in comparison with that recognized by the standard Makov models using only nucleotide and dinucleotide frequencies.

In constructing a model of a generating tree source for a corresponding sample of functional sites, it is unnecessary to align preliminarily the sites or even determine their boundaries. Positioning of symbols is not necessary in contrast to the weight matrix. Local contexts substitute positioning. Thus, this method of simulation is an appropriate alternative to the weight matrix method.

Genome segmentation

Let us consider the problem of segmentation of long sequences. The variable memory model allows recognition of typical and atypical regions of the entire sequence, i.e., the latter are the regions with a low probability of being accidentally observed. The probability for a region is calculated as a probability of observation of all the symbols taking into account the previous local context. In this case, the training involves the sequences of the entire genome.

The probability profile constructed by a sliding window of a length of 1 kbp shows notable regions that are statistically atypical of the entire genome. Figure 4 gives an example of a profile in a sliding window for the *Bacillus subtilis* genomic sequence. The arrows show the regions that are most atypical of the genome. These are regions of ribosomal protein genes (14 kbp), ribosomal and transport RNAs (98–100 kbp, 165–170 kbp, 635 kbp, 946–950 kbp, and 3,172 kbp, respectively). Thus, the regions that belong to the translation system—the most ancient and conservative system in living organisms—differ mostly in the local context composition.

The position (bp) is plotted on the X axis, and a sum of logarithms of the observation probability for nucleotides (with the minus sign) in a 1 kbp window of the tree source model is plotted on the Y axis.

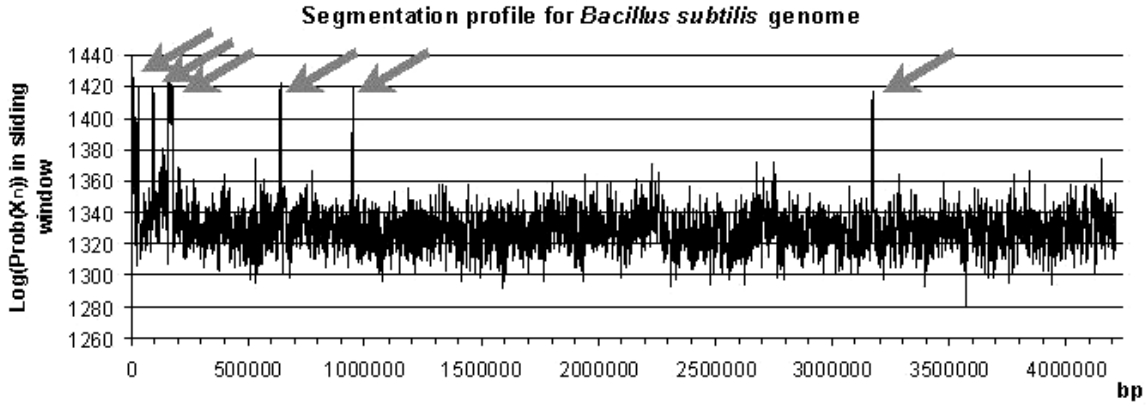


**Fig. 4.** Profile of compliance with the general model (statistical proximity) of local regions of *Bacillus subtilis* genome in a sliding 1 kbp window.

## Discussion

The program Complexity allows us to construct a model of generation of a genetic text and determine the complexity of this text. The presented version of the program is designed for analysis of long DNA sequences of any size.

An important characteristic of genomic sequences is their oligonucleotide composition, which reflects evolutionary interactions between organisms (Karlin, Ladunga, 1994; Scherer et al., 1994). Examples of distribution of short oligonucleotides and clustering of significant oligonucleotides were studied by Haring and Kypr (1999). However, a unified method for the representation of oligonucleotides specific of a genomic DNA has not been developed. A graphic representation in the form of a tree structure is rather illustrative. The program developed presents a tree structure of oligonucleotides based on the selection by the minimum description length principle (Barron et al., 1998). This distinguishes our model from a graphic representation of statistically under-represented and over-represented oligonucleotides constructed by the program Verbumculus (http://wwwdbl.dei.unipd.it/Verbumculus/) (Apostolico et al., 2000).

The recognition of functional regions using variable memory Markov models is a promising method. Currently, we work on segmentation of all sequenced genomic sequences.

## Acknowledgements

The authors thank V.A.Likhoshvai, M.A.Pozdnyakov, and N.A.Kolchanov for their helpful discussions. The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 00-07-90337, 02-07-90355, 00-04-49229, and 02-01-00939); Ministry of Industry, Science, and Technologies of the Russian Federation (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); and the INTAS foundation (grant № YSF 00-178).

## References

1. Apostolico A., Bock M.E., Lonardi S., Xu X. (2000). Efficient detection of unusual words. J. Comput. Biol. 7(1/2):71–94.
2. Barron A., Rissanen J., Yu B. (1998). The minimum description length principle in coding and modeling. IEEE Trans. Inform. Theory. 44:2743–2760.
3. Haring D., Kypr J. (1999). Variations of the mononucleotide and short oligonucleotide distributions in the genomes of various organisms. J. Theor. Biol. 201:141–156.
4. Karlin S., Ladunga I. (1994). Comparisons of eukaryotic genomic sequences. Proc. Natl Acad. Sci. USA. 91:12832–12836.
5. Orlov Yu.L., Filippov V.P., Potapov V.N., Kolchanov N.A. (2002). Construction of stochastic context trees for genetic texts. (Bioinformation Systems e.V.) In Silico Biology, 2(0022) <http://www.bioinfo.de/isb/2002/02/0022/>
6. Orlov Yu.L., Potapov V.N. (2000). Estimation of stochastic complexity of genetic texts. Computational Technologies (Novosibirsk), 5:5–15.
7. Ron D., Singer Y., Tishby N. (1996). The power of amnesia: learning probabilistic automata with variable memory length. Machine Learning. 25:117–149.
8. Scherer S., McPeek M.S., Speed T.P. (1994). Atypical regions in large genomic DNA sequences. Proc. Natl Acad. Sci. USA. 91:7134–7138.