NUCLEOSOME POSITIONING SIGNAL ANALYSIS

Orlov Yu.L.*, Levitsky V.G.

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

Keywords: *nucleosome, gene expression regulation, context signals*

Summary

Motivation: Computer analysis of context features in nucleosome formation sites is of a great importance. The aim was to examine the variable memory Markov model method for estimation of nucleosome formation potential (tendency to bind histone octamer and nucleosome forming) of arbitrary nucleotide sequences.

Results: The variable memory Markov models method was applied for estimation of the nucleosome formation ability and showed quite similar results on genomic DNA as previously applied method of discriminant analysis of oligonucleotide frequencies RECON. An analysis of phased nucleotide sequences containing nucleosome formation sites revealed periodic signals in local text complexity.

Availability: Software is available by request to the corresponding author.

Introduction

Chromatin from various genomic regions is, as a rule, represented by regular arrays of nucleosomes (Aalfs, Kingston, 2000; Becker, 2002). The neighboring nucleosomes are connected by linker DNA with a length ranging from 20 to 80 bp. The sequence-directed nucleosome positioning plays an important functional role in providing a proper interaction of DNA functional sites with non-histone proteins. The mechanisms of sequence-directed nucleosome positioning have been studied in numerous experiments both *in vivo* and *in vitro*; the results obtained suggest the existence of a specialized chromatin (nucleosome) code determining such positioning through multiple histone-DNA interactions (Trifonov, 1997).

Various research teams have succeeded in discovering a number of periodic contextual and conformational signals and rules regulating nucleosome positioning (Trifonov, 1997; Kiyama, Trifonov, 2002; Levitsky *et al.*, 2004). Although this field is intensely studied, the mechanisms underlying nucleosome positioning are yet far from being clearly understood. Computer analysis of nucleosome positioning code needs different unrelated methods, such as estimation of sequence linguistic complexity, search for periodic signals and Markov models. Indeed, context nucleosome positioning code (Trifonov, 1997) assumes degeneracy (quite different DNA sequences are able to interact with histone octamer and nucleosome formation), weakness of context signals, and absence of clear signal localization. Markov model (Orlov *et al.*, 2002) of text generation is based on preceding symbols, thus as a statistical model Markov model corresponds to theoretical assumptions about nucleosome code.

New prediction method is based on Variable memory Markov model and allows detect the prediction tendency to bind histone octamer for arbitrary nucleotide sequences. The software developed allows to reveal context features such as periodicity and low complexity region distribution in DNA sequences containing nucleosome formation sites.

It was compared with the software developed previously, which allows predicting nucleosome formation potential, by dinucleotide frequencies (Levitsky *et al.*, 2001; Levitsky *et al.*, 2004).

Methods

To develop programs for detecting and recognizing nucleosomal context and to evaluate their prognostic capability, several samples of NFS – DNA sequences with experimentally demonstrated ability to form nucleosomes – are used. They include 141 sequences extracted from the GenBank/ EMBL databank according to the accession numbers and positions indicated in the database Nucleosomal DNA (Ioshikhes and Trifonov, 1993). These samples were used for designing software packages (Levitsky *et al.*, 2001), intended for analysis of contextual NFS properties and prediction of the ability of arbitrary DNA sequences to form nucleosomes.

Analysis of text complexity of NFS was done using "LowComplexity" software for estimation of DNA sequence complexity by several methods including modified Lempel-Ziv method and entropy estimations (Orlov *et al.*, 2004).

Prediction of nucleosome formation potential was developed using extended Markov model trained on the database. We consider Variable Memory Markov Models for the generation of symbols based on a stationary source. The local preceding context (1–9 bp) defines the current state of the Markov model independent of the position in the text. Such model is based on suffix tree. Suffix trees as a basis for text generation were studied for proteins as an alternative to the HMM in the form of a sub-class of probabilistic finite automata.

Results and Discussion

Prediction of nucleosome formation potential. Prediction of nucleosome formation potential was preformed by estimation of probability of arbitrary sequence correspond to the Markov model trained on the database. Fit function was constructed as minus logarithm of probability to obtain a sequence S in sliding window by fixed length n: -Prob(S_n). Testing the software on control sample of experimentally defined NFS showed correct results. Nucleosome formation is especially important in regulatory regions of genes, in particular, promoters (Aalfs, Kingston, 2000). We analyzed set of gene promoters from TRRD database (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd). Averaged profiles are presented at Figure 1.

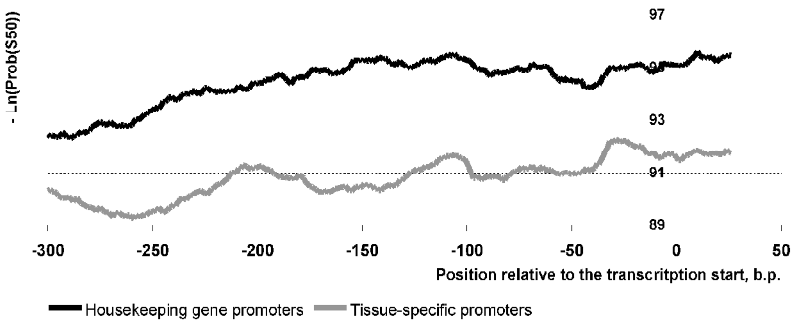

Fig. 1. Nucleosome formation potential profiles in sliding window 50 bp for sets of eukaryotic gene promoters phased relative to the transcription start [-300;+50]. Lower level means greater similarity to nucleosome formation sites. Averaged profile for the set of promoters of housekeeping genes (high expression level) and tissue-specific genes (lower expression level) are designated by solid black and light gray lines correspondingly.

Comparison of nucleosome potential function based on Variable Memory Markov Models and function based on dinucleotide frequencies and discriminant analysis (Levitsky *et al.*, 2001) revealed the similar results for nucleosome formation site sequences and promoter sequences.

Promoters of housekeeping genes have less preference to contain NFS than tissue-specific genes (Levitsky *et al.*, 2001). Thus the research suggested that high expressed genes in nucleus should be free of chromatin packing than rare expressed (tissue-specific) genes.

The result is in good accordance with our previous analysis (Levitsky *et al.*, 2001). Moreover significant correlation of nucleosome potential estimation functions was shown on genomic DNA. In addition our analysis shown than introns and 5'-non-translated gene regions have less nucleosome formation potential than exons and regulatory regions of gene expression.

Low complexity regions in nucleosome formation sites. Let us consider local complexity profiles by sliding window for phased set of NFS. Profile value for every sequence reflects number of copying operation (direct repeats) to construct the sequence. It is an integer value; we averaged all this values for every position of the phased sequence set. The DNA sequences by length 400 bp were phased relative to the center of NFS [-200;+200].

Middle part [-73;+73] corresponds to 147-bp DNA wrapped around histone octamer (gray bar in Fig. 2). Linker DNA sequences (approximately 50 bp) are designated by stripped bars to left and to right from the center. Since we assume existence of ordered nucleosome array in genomic DNA we designate far left and right flanks of sequence by gray bars as part of neighboring nucleosome sites.

Figure 2 presents only averaged complexity profile by modified Lempel-Ziv method in sliding window 27 bp for whole sequence set. One can see, that complexity values are minimal at the flanks of core nucleosome site (site position designated by gray bar). Linker DNA corresponds to higher level of text complexity. Moreover, local trend of increase and decrease of complexity from NFS center exist (dotted line). At whole we see symmetrical picture of complexity values distribution (Fig. 2) corresponding to biophysical assumptions about molecular mechanisms of DNA packing. In addition, profile complexity shows periodical distribution of local complexity minima (Fig. 2, vertical arrows). These minima have period 10–11 bp corresponding to previous estimations (Kiyama, Trifonov, 2002; Ioshikhes, Trifonov, 1993). Such preference in local minima distribution could be connected with periodic distribution of simple sequence repeats, even as short as 2–3 nucleotides.

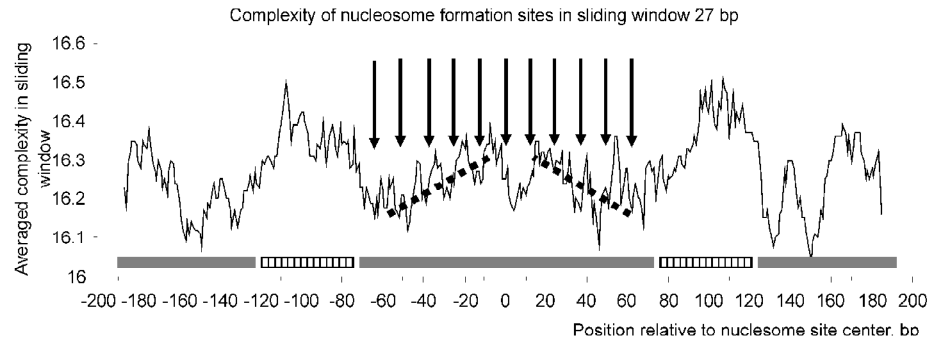

Fig. 2. Averaged complexity profile by modified Lemple-Ziv method for phased set of nucleosome formation sites [-200;+200]. Gray bars indicate 146 bp core histone binding sequences, stripped bars correspond to linker DNA. Profile trends are indicated by straight dotted lines. Arrows show periodic (10–11 bp) local complexity minima.

We plan further developing of an integrated information system Nucleosomal DNA Organization, comprising Nucleosome Positioning Region Database and software packages for nucleosome formation sites (NFS) recognition (Levitsky *et al.*, 2001).

Acknowledgements

The authors are grateful to Yu.D. Oshchepkov, A.V. Katokhin, O.A. Podkolodnaya and N.A. Kolchanov for valuable discussion. This work was supported in part by the RFBR (02-07-90355, 03-04-48555), Ministry of Education (E 02-6.0-250), SB RAS (Integration Projects 119), NATO (LST.CLG 979815), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), Project No. 10.4 of the RAS Presidium Program "Molecular and Cellular Biology".

References

Aalfs J.D., Kingston R.E. What does 'chromatin remodeling' mean? // Trends Biochem Sci. 2000. V. 25. P. 548–55.

Becker P.B. Nucleosome sliding: facts and fiction // EMBO J. 2002. V. 21. P. 4749–53.

Ioshikhes I., Trifonov E.N. Nucleosomal DNA sequence database // Nucleic Acids Res. 1993. V. 21. P. 4857–9.

Kiyama R., Trifonov E.N. What positions nucleosomes? – A model // FEBS Lett. 2002. V. 523. P. 7–11.

Levitsky V.G., Katokhin A.V., Podkolodnaya O.A., Furman D.P. Nucleosomal DNA organization: an integrated information system // Bioinformatics of Genome Regulation and Structure / Eds. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston/Dordrecht/London, 2004. P. 3–12.

Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis // Bioinformatics. 2001. V. 17. P. 998–1010.

Orlov Yu.L., Potapov V.N., Filippov V.P. Recognizing functional DNA sites and segmenting genomes using the program Complexity // Proc. of the BGRS'2002. Novosibirsk: IC&G, 2002. V. 3. P. 243-246.

Orlov Yu.L., Potapov V.N. Complexity: Internet-resource for analysis of DNA sequence complexity // Nucleic Acids Res. 2004. (web-server issue 2004). In press.

Trifonov E.N. Genetic level of DNA sequences is determined by superposition of many codes // Mol. Biol. (Mosk). 1997. V. 31(4). P. 759-67.