BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 12 2008, pages 1456–1458
doi:10.1093/bioinformatics/btn212

Sequence analysis

nuScore: a web-interface for nucleosome positioning predictions

Michael Y. Tolstorukov[1,*], Vidhu Choudhary[2], Wilma K. Olson[3], Victor B. Zhurkin[4] and Peter J. Park[1,2,*]

[1]Harvard-Partners Center for Genetics and Genomics, Brigham and Women's Hospital, Boston, MA 02115, [2]Children Hospital Informatics Program, Boston, MA 02115, [3]Wright-Rieman Laboratories, Rutgers, The State University of New Jersey, Piscataway, NJ 08854 and [4]Laboratory of Cell Biology, National Cancer Institute, Bethesda, MD 20892, USA

Received on January 2, 2008; revised on April 25, 2008; accepted on April 26, 2008

Advance Access publication April 29, 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: Sequence-directed mapping of nucleosome positions is of major biological interest. Here, we present a web-interface for estimation of the affinity of the histone core to DNA and prediction of nucleosome arrangement on a given sequence. Our approach is based on assessment of the energy cost of imposing the deformations required to wrap DNA around the histone surface. The interface allows the user to specify a number of options such as selecting from several structural templates for threading calculations and adding random sequences to the analysis.

Availability: The nuScore interface is freely available for use at http://compbio.med.harvard.edu/nuScore.

Contact: peter_park@harvard.edu; tolstorukov@gmail.com

Supplementary information: The site contains user manual, description of the methodology and examples.

1 INTRODUCTION

Nucleosomes, the basic repeating units of chromatin, comprise 147 bp of DNA wrapped around the histone protein core (Richmond and Davey, 2003). Placement of nucleosomes at transcription start sites may significantly affect gene expression (Jenuwein and Allis, 2001; Kornberg and Lorch, 1999) by occluding access to DNA for regulatory proteins or recruiting them through the interactions with the histones. Therefore, prediction of nucleosome positioning on genomic sequences is of great biological interest.

The ability of DNA to form nucleosomes depends at least partially on the underlying sequence (Trifonov, 1980; Widom, 2001). Thus, the nucleosome arrangement for a DNA fragment can be predicted based on the sequence alone in the absence of other factors, e.g. chromatin-remodeling proteins. Most of the computational algorithms currently available for such predictions use scoring functions based on the distributions of dinucleotides in the sequences and are trained on the sets of sequences known to position nucleosomes (Ioshikhes *et al.*, 1999; Segal *et al.*, 2006; Yuan and Liu, 2007). Another approach is to use the sequence-dependent structural properties of DNA to identify sequences that wrap more (or less) readily around the histone core (Anselmi *et al.*, 2000; Mengeritsky and Trifonov, 1983; Sivolob and Khrapunov, 1995; Zhurkin, 1983).

*To whom correspondence should be addressed.

In line with the 'structural' approach, we estimate the energy cost of the deformations imposed by histones on DNA (Tolstorukov *et al.*, 2007). Such a cost is sequence-dependent (Olson *et al.*, 1998) and, since the sequence-specific interactions between the histones and DNA bases are essentially absent, is one of the main factors determining affinity of the histones to DNA. Here, we present a web-based application that implements this methodology for prediction of nucleosome positioning patterns on DNA sequences.

2 METHODOLOGY

The cost of DNA deformation is estimated by 'threading' the DNA fragment of a given sequence on the template comprising the trajectory of DNA observed in the structure of a nucleosome core particle determined experimentally. We use the representation of DNA structure at the dinucleotide level, i.e. with a set of collective variables specifying the relative positioning of neighboring base pairs (Twist, Tilt, Roll, Shift, Slide and Rise) (Dickerson *et al.*, 1989). The deformation energy score E of the threaded sequence can be expressed as (Olson *et al.*, 1998):

$$E = \sum_{n=1}^{L}\left(\frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6} f_{ij}(\mathrm{MN})\Delta\theta_i^n\Delta\theta_j^n\right), \quad (1)$$

where, $\Delta\theta_i^n = \theta_i^n - \theta_i^0(\mathrm{MN})$ is the imposed deviation of the i-th dinucleotide step parameter θ_i^n at the n-th step of the template from the rest-state value $\theta_i^0(\mathrm{MN})$ of the step MN. The $f_{ij}(\mathrm{MN})$ are DNA stiffness constants that depend on the sequence, and L is the number of base pair steps in the nucleosome template. The rest-state values and stiffness constants of the 16 dinucleotide steps were estimated from the analysis of a set of protein–DNA complexes (Olson *et al.*, 1998).

A strong positioning DNA sequence is expected to show a sharp dip in the energy score for the preferred template setting. The positioning score P can be defined by the relative deviation of the energy score calculated for a given position x of the template on the sequence with respect to the scores calculated for n neighboring positions of the template (Fig. 1):

$$P = \frac{E(x) - \langle E\rangle_{\pm n/2}}{\sigma_{\pm n/2}} \quad (2)$$

Large negative values of P denote nucleosome-attracting sequences (expected to favor nucleosomes) and large positive values point to nucleosome-refractory sequences (expected to repel nucleosomes). To avoid erroneous predictions in cases of very small values of $\sigma_{\pm n/2}$,

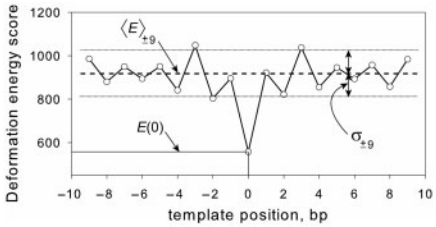

Fig. 1. Calculation of the nucleosome-positioning score P. Here, the 147 bp sequence from the currently best-resolved nucleosome structure (Richmond and Davey, 2003) is threaded on a template made up of the central 129 bp of this structure. The energy $E(x{=}0)$ at the position, corresponding to the experimentally observed setting with the nucleosome dyad positioned on the central base pair of the sequence, is compared with the mean energy $\langle E\rangle_{\pm 9}$ of the 18 neighboring settings of the template on the sequence (Equation 2). The positioning score equals −5 in the case shown.

the condition $|E(x) - \langle E\rangle_{\pm n/2})|/\langle E\rangle_{\pm n/2}\ E > 0.1$ is used in addition to Equation 2.

All numerical calculations are performed with FORTRAN code (available upon request). The interface was implemented using HTML and java script at the front end and CGI/Perl at the server end. A module written in R (http://www.R-project.org) is used to produce the output plots. Computational time depends linearly on the total number of the bases in the input sequences and the server can handle a large set of input sequences easily. The algorithm was tested on the sequences for which nucleosome positions were experimentally mapped to base pair resolution (see Supplementary material on the nuScore site).

3 OPTIONS AND CONTROLS

A general view of the nuScore interface is shown in Figure 2. The user can select from eight nucleosomal templates based on crystal structures solved to a resolution of 3Å or better. These structures contain native histones from different organisms and no additional ligands (see Methods section on the nuScore site for details). Since the threading templates are not symmetrical, two orientations of each template on the sequence are possible and the user has the option to select one. Also, the average or the lowest-energy score for the two template orientations may be reported in the output files. Alternatively, the user may use templates with the parameters symmetrized relative to the structural dyad.

A set of random sequences of the same dinucleotide composition can be generated for each input sequence. The number of random sequences in the set is specified by the user. The mean and standard deviation of the deformation and positioning scores for each random set are reported in the output files.

The user can select the size of the window used to calculate the nucleosome-positioning score, which specifies how many neighboring positions are used to compute the score at a given position (Equation 2). Considering 18 neighboring positions seems to be reasonable; using fewer neighbors may not be enough to produce statistically sound results and using more positions may interfere with known nucleosome-positioning signals, which often show 10-bp periodicities [e.g. AA:TT signal (Trifonov, 1980)].

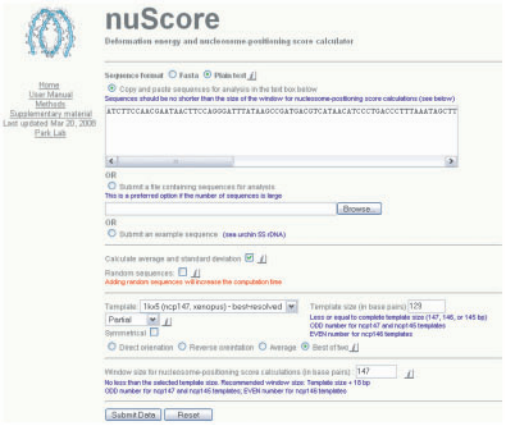

Fig. 2. A screen shot of the nuScore interface.

4 FUTURE DEVELOPMENTS

Other models of DNA flexibility, based on the analysis of other sets of DNA structures (Morozov *et al.*, 2005) or the molecular dynamics trajectories of DNA (Lankas *et al.*, 2003), can easily be implemented into the program. Incorporation of trimeric models is anticipated to increase the predictive power of our approach since such models combine descriptions of DNA deformability at the levels of dinucleotides and individual base pairs.

Finally, using threading to assess DNA deformation scores assumes that all sequences adopt exactly the same fold on the nucleosome. The structure of the nucleosomal DNA, however, may depend on the sequence. Optimization of the DNA trajectory for each given sequence will further advance our structural approach.

ACKNOWLEDGEMENTS

Funding: This work was supported in part by the Intramural Research Program of the NIH National Cancer Institute, Center for Cancer Research and USPHS grant GM20861.

Conflict of Interest: none declared.

REFERENCES

Anselmi,C. *et al.* (2000) A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability. *Biophys. J.*, **79**, 601–613.

Dickerson,R.E. *et al.* (1989) Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, **208**, 787–791.

Ioshikhes,I *et al.* (1999) Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc. Natl Acad. Sci. USA*, **96**, 2891–2895.

Jenuwein,T. and Allis,C.D. (2001) Translating the histone code. *Science*, **293**, 1074–1080.

Kornberg,R.D. and Lorch,Y. (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, **98**, 285–294.

Lankas,F. *et al.* (2003) DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.*, **85**, 2872–2883.

Mengeritsky,G. and Trifonov,E.N. (1983) Nucleotide sequence-directed mapping of the nucleosomes. *Nucleic Acids Res.*, **11**, 3833–3851.

Morozov,A.V. *et al.* (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.

Olson,W.K. *et al.* (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.

Richmond,T.J. and Davey,C.A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145–150.

Segal,E. *et al*. (2006) Genomes utilize a nucleosome positioning code to achieve biological function. *Nature*, **442**, 772–778.

Sivolob,A.V. and Khrapunov,S.N. (1995) Translational positioning of nucleosomes on DNA: the role of sequence-dependent isotropic DNA bending stiffness. *J. Mol. Biol.*, **247**, 918–931.

Tolstorukov,M.Y. *et al*. (2007) A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.*, **371**, 725–738.

Trifonov,E.N. (1980) Sequence-dependent deformational anisotropy of chromatin DNA. *Nucleic Acids Res.*, **8**, 4041–4053.

Widom,J. (2001) Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.*, **34**, 269–324.

Yuan,G.C. and Liu,J.S. (2008) Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.*, **4**, e13.

Zhurkin,V.B. (1983) Specific alignment of nucleosomes on DNA correlates with periodic distribution of purine-pyrimidine and pyrimidine-purine dimers. *FEBS Lett.*, **158**, 293–297.