2011 International Conference on Bioscience, Biochemistry and Bioinformatics
IPCBEE vol.5 (2011) © (2011) IACSIT Press, Singapore

The DNA Geometric Flexibility of Promoter in Model Organism Genomes

Yongchun Zuo , Qianzhong Li
Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot, 010021, China
Email: yczuo@imu.edu.cn

Abstract—**The general mechanism of transcription initiation is dependent on protein-DNA interaction. The promoter DNA has to possess the required 3D structure to allow DNA-binding events and accommodate alterations in activator protein positioning in order to allow the start of transcription. In this study, we introduced physical structural variables by defining three displacements (rise, shift and slide) and three rotation angles (twist, roll and tilt) for each dinucleotide to describe transcription initiation of four model organism genomes. The analysis results reveal that there are specific structural patterns in the core promoter region, and confirm that the existence of a hidden physical code which can modulates the gene transcription initiation. The results demonstrate that there is significant diversity in the geometric flexibility for different organizations. We believe the structure features originate from DNA 3D structures can describe the actual process of gene transcription initiation better than the curvature propensity or other general flexibility**

Keywords- Promoter sequences; DNA geometric flexibility; Molecular dynamics (MD) simulations; Transcription initiation regulation

I. INTRODUCTION

The first process of transcription initiation mechanism is usually dependent on protein-DNA interaction in which the different transcription factors (TFs) bind on the promoter region to enable its activity. An important step during transcription initiation is the open complex formation between RNAP and promoter sequence [1-3]. The transcription process takes place under conditions in which the two strands of DNA is separated by with the help of any external energy. So the DNA sequence-dependent three-dimensional structure is important for transcriptional regulation in the single binding sites and core promoter regions. Regulatory sequences such as core promoter regions not only contain specific sequence elements that serve as targets for interacting proteins, but also exhibit distinct structural properties of the sequence[4,5].

The function of promoter regions are assembling transcription factor protein, thus they usually have some distinct structural abilities to allow functional protein binding and transcription factor protein positioning. Hence it is necessary to investigate promoter sequences in higher dimensions than the sequence alone. The DNA 3D structure is characterized by the local angular parameters (twist, roll and tilt) and the translational parameters (shift, slide and rise) between two successive base-pair steps [6]. According to a recently developed force-field fitted to high level ab initio quantum mechanical calculations [7], the 16 DNA geometric flexibility parameters of all dinucleotides are calculated by long atomistic molecular dynamics (MD) simulations in water. Using the nearest-neighbor dinucleotide parameters derived from the structural models. In this paper, we converted the core-promoter sequences into a string of numerical values for describing the transcription start regions. The emerging results in this study show the existence of a hidden structural and deformation code conserved throughout the evolution. We believe that the physical patterns are useful for further elucidating the transcription regulation mechanism.

II. MATERIALS AND METHODS

A. Materials

Drosophila and Human promoter sequences: The Drosophila and Human Pol-II promoter sequences were extracted from Eukaryotic Promoter Database (EPD, Release 94)[8]. The basal promoter with 301bp are abstracted, starting from 250 bp upstream (position –250) and extending up to 50 bp downstream (position +50) of the TSS(position 0). The promoter sequences with base 'N' have been filtered out from the promoter datasets. At last, 1922 Drosophila promoters and 1862 Human Pol-II promoters are selected for building the promoter datasets.

Arabidopsis promoter sequences: The TSSs for Arabidopsis promoters were obtained from The Arabidopsis Information Resource (TAIR, http://arabidopsis.org). The TAIR database maintains a database of genetic and molecular biology data for the model higher plant Arabidopsis thaliana [9]. To obtain the sequences of the region of promoters spanning the first 200 bp upstream of the TSS [–200, –1] and the corresponding the 50 bp downstream of the TSS, the corresponding full-length cDNA datasets and the [–500, –1] regions datasets were downloaded from TAIR7_blastsets. Totally, we retrieved 27,775 non redundant promoter sequences of Arabidopsis with annotation TSSs.

E.coli promoter sequences: The core promoter sequences of *E.coli K-12* genome are abstracted from the RegulonDB database [10]. In prokaryotic cells, the type of promoter is defined by sigma factor of RNA polymerase. The 689 σ^{70} *E.coli* core promoter sequences with 81bp long regions (–60bp upstream and +20bp downstream flanking transcription start sites (TSS)) are abstracted from *E.coli K-12* dataset.

B. The dinucleotide flexibility parameters of DNA geometric freedom used in this paper

Recently several studies have reported that the sequence-dependent secondary properties are often involved in stability, curvature and bend ability of DNA in the promoter regions. Recently, according to the physical potentials derived from quantum chemical calculations, Goñi et al. developed the helical stiffness parameters to reveal the complexity of the deformation pattern of DNA. By combing with six geometric degree of freedom, they successfully determining promoter location by using a well-defined physical description of DNA deformability [11]. The results show that these dinucleotide flexibility parameters derived from long atomistic molecular dynamics (MD) simulations in water can define promoters as regions of unique deformation properties particular well, particularly near TSSs. In this study, six geometric degrees of freedom are used to describe the physical structure of transcription initiation for different species. The 16 different nearest-neighbor interactions values of each dinucleotide are shown in Table I.

TABLE I. THE DINUCLEOTIDE FLEXIBILITY PARAMETERS DERIVED FROM QUANTUM CHEMICAL CALCULATIONS

Dinucleotide	**Twist**	**Tilt**	**Roll**	**Shift**	**Slide**	**Rise**
AA/ TT	0.026	0.038	0.02	1.69	2.26	7.65
AC/ GT	0.036	0.038	0.023	1.32	3.03	8.93
AG/ CT	0.031	0.037	0.019	1.46	2.03	7.08
AT	0.033	0.036	0.022	1.03	3.83	9.07
CA/ TG	0.016	0.025	0.017	1.07	1.78	6.38
CC/GG	0.026	0.042	0.019	1.43	1.65	8.04
CG	0.014	0.026	0.016	1.08	2.00	6.23
GA/TC	0.025	0.038	0.020	1.32	1.93	8.56
GC	0.025	0.036	0.026	1.20	2.61	9.53
TA	0.017	0.018	0.016	0.72	1.20	6.23

Parameters values related to rotational parameters are in kcal/mol degree^2, while those related to translations are in kcal/mol Å ^{2}.

III. RESULT AND DISCUSSION

A. Structure profile of three displacements and three rotation angles in core promoter rigions

This rigid-body representation of base pairs reduces significantly the number of independent variables per chain molecule, thereby making it possible to study the normal modes of longer DNA fragments [12]. The use of the DNA geometric flexibility values allows us to define deformation properties of promoter regions with these in the genomic level, particularly near TSSs. The six variables show the relative change of angular and distance for neighbor base-pair in the three-dimensional space. Here we used a sliding window approach with a step of 1 and a window size of 1–15 nt to analyze the DNA geometric flexibility. The flexibility profiles based on the structural model are shown in Figure.1.

As the 3D structures of DNA possess a degree of anisotropic flexibility, depending on the sequence it might bend more easily in one plane than another. From the graphs shown in Figure.1, we can observe the six variables originate from DNA 3D structure can reveal the relative change of angle and distance for neighboring base pairs in the 3D space clearly. For promoter profiles of different organism genomes, both the local angular patterns and translational patterns show similar profiles. All the six 3D descriptors present a coherent shape of profiles, rising and falling nearly synchronously. All of them have two clear peaks/valley at position –35bp (TATA-binding protein location) and position 0bp (TSS). The first region, where the TBP is known to bind, is located around the 30 bp upstream of the TSS. This region is crucial for allow functional protein binding events and the assembly of the transcription machinery. The second region is located around the initiator itself, and needs to denature to allow transcription to start [13]. So it requires that the TATA boxes and initiator sequences have the distinct flexible and rigid sequences compared to other parts of promoters.

It is worth noting that all of the three organism genomes have the strikingly similar trend, but each type of promoters has the different values. For example, for slide profile the human promoters have the largest values and the *Arabidopsis* promoters have the lowest values. For the twist profile, the *Arabidopsis* promoters have the largest values and the human promoters have the lowest values. For one organism genome promoter, the six profiles of DNA geometric flexibility also show different patterns around the transcription initiation region. For most profiles of the DNA 3D descriptors, the patterns show two valleys at position –35 (TATA-binding protein location) and position 0 (TSS). But the slide profiles present two clearly peaks at the two regions. We can conclude that the structure features of DNA geometric flexibility have distinctive structural patterns along the sequences. These inherent structures may be specific to the interaction between proteins and DNA, and essential for transcription initiation. And these distinctive values of DNA geometric flexibility are also very useful for promoter identification.

B. The TFs binding regions have typical structural flexibility

In order to quantitative analyze the characterization of transcription factor binding regions associated to the structural parameters, the 16 flexibility values based on long atomistic MD simulations are applied to analyze the changes of stiffness and flexibility at the TATA-box and Initiator sites. Compared to those regions located far from the annotated TSSs, the structural patterns in these regulatory regions are quite complex. The results are shown in Table II.

As the results shown in Table 2, for the structural profile of rotational flexibility the twist profile of *Arabidopsis* promoter has the largest values in the –35bp and –0bp positions and the average stiffness values are 4.03 kcal/mol degree^2 and 8.50 kcal/mol degree^2, respectively. The human promoters show the largest tilt profile value the average stiffness values are 5.63 kcal/mol degree^2 and 12.00 kcal/mol degree^2, respectively. The structural profiles of translational flexibility perform the similar regularity with the rotational flexibility. The Shift pattern profile of *E.coli* promoter in the –10bp positions performs the lowest average stiffness with the value is 436.44 kcal/mol Å^2, but the Slide profile contains two peaks in the –35bp and –10bp positions and the average stiffness values are 352.70 kcal/mol Å^2 and 749.71 kcal/mol Å^2, respectively. The high flexibility near TSSs is required for some parameters, while rigidity is needed for others. This indicates that the local changes around the expected positions of TATA-box and TSS are higher informative than other regions. From the lower organisms (*E.coli*) to humans, all the structural patterns of DNA geometric flexibility measured in the core promoter regions are quite unique. Our use of DNA bases translation parameters and rotation parameters can be deeper, more complete describe the actual start of gene transcription, better general concepts like "curvature propensity" or "general flexibility" [14].

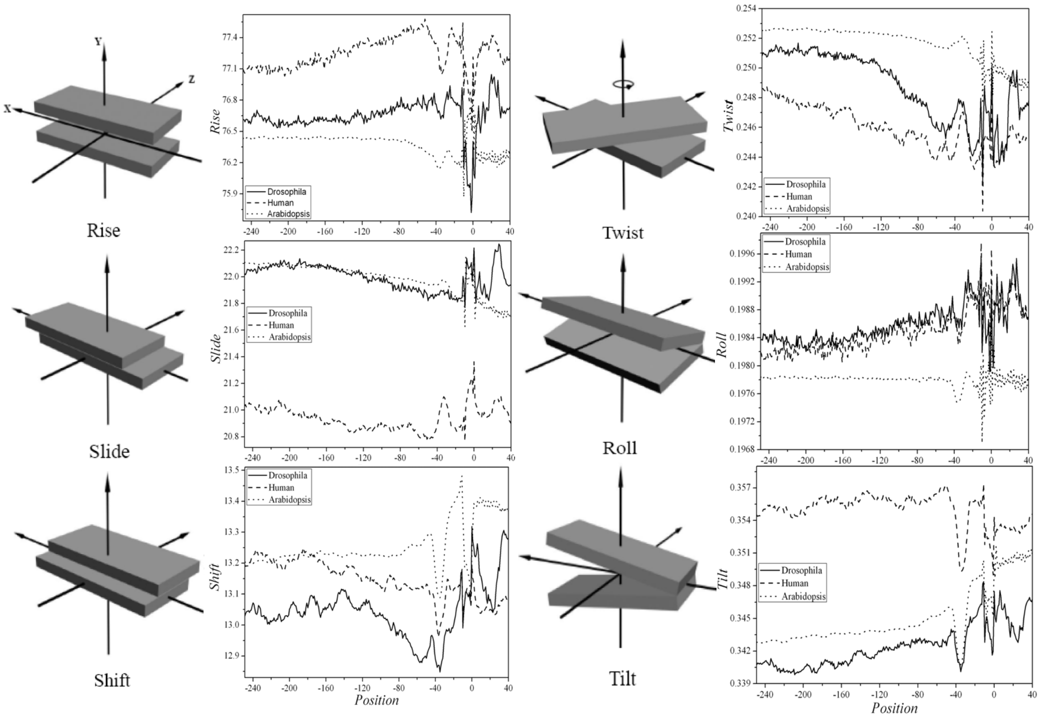

Figure 1. The profiles of DNA geometric flexibility based on structural model (with a step of 1 and a windows size of 10).

TABLE II. THE STIFFNESS VALUES OF DNA GEOMETRIC FEATURES IN THE TATA-BOX AND INR REGIONS (WITH A STEP OF 1 AND A WINDOWS SIZE OF 10).

Type	Drosophila		Human		Arabidopsis		E.coli	
	–40,–25	*–15,+2*	*–40,–25*	*–15,+2*	*–40,–25*	*–15,+2*	*–40,–25*	*–15,+2*
Twist	3.95	8.38	3.93	8.32	4.03	8.50	3.96	8.44
Tilt	5.48	11.72	5.63	12.00	5.50	11.82	5.49	11.54
Roll	3.18	6.75	3.18	6.76	3.16	6.72	3.18	6.75
Shift	206.67	446.34	208.77	445.86	211.09	451.32	209.20	436.44
Slide	350.03	747.87	336.03	716.38	351.17	745.43	352.70	749.71
Rise	1227.91	2592.43	1235.41	2618.85	1219.59	2599.72	1226.83	2606.14

Parameters values related to rotational parameters are in kcal/mol degree^2, while those related to translations are in kcal/mol Å ^{2}.

IV. CONCLUSION

During the past 20 years, the selection of right biological signals to recognize promoters remains obscure. Current findings indicate that protein–DNA binding specificity is also modulated by energetics [2,7,15]. However, the intrinsic limitations of the sequencing projects became clear when researchers realized that the mechanisms allowing the supramolecular organization of the genome and the control of its expression were not directly coded in the sequence, but depend on the chromatin structure and flexibility[11]. The conclusion findings in this paper indicate the existence of a hidden structural and deformation code, which has been conserved throughout the evolution and that helps the assembly of the transcription initiation complex.

ACKNOWLEDGMENT

We would like to thank Goñi J Ramon for flexibility parameters shown of dinucleotide in Table 1. This work was supported by the National Natural Science Foundation of China (No. 30560039), the Natural Science Foundation of Inner Mongolia Autonomous Region (No. 200607010101) and the "211 project" innovative talents training program of Inner Mongolia University.

REFERENCES

[1] X. Q. Cao, J. Zeng, H. Yan, "Structural property of regulatory elements in human promoters," Phys. Rev. E., vol.77, Apr. 2008, pp.1–7.

[2] A. Kanhere, M. Bansal, "Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes," Nucleic Acids Res., vol.33, Jun. 2005, pp.3165–3175.

[3] C. Yang, E. Bolotin, T. Jiang, F. M. Sladek, E. Martinez, "Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters," Gene, vol.389, Mar. 2007, pp.52–65.

[4] T. Abeel, Y. Saeys, E. Bonnet, P. Rouzé, de Peer. Y. Van, "Generic eukaryotic core promoter prediction using structural features of DNA," Genome Res., vol.18, Feb. 2008, pp.310–323.

[5] K. Florquin, Y. Saeys, S. Degroeve, P. Rouzé, de Peer. Y. Van, "Large-scale structural analysis of the core promoter in mammalian and plant genomes," Nucleic Acids Res., vol.33, Jul. 2005, pp.4255–4264.

[6] W. K. Olson, A. A. Gorin, X. J. Lu, L.M. Hock, V. B. Zhurkin, "DNA sequence-dependent deformability deduced from protein-DNA crystal complexes," Proc. Natl. Acad. Sci. USA. vol.95, Sep. 1998, pp.11163–11168.

[7] J. R. Goñi, A. Pérez, D. Torrents, M. Orozcon, "Determining promoter location based on DNA structure first-principles calculations," Genome Biol., vol.8, Dec. 2007, pp.R263.

[8] C. D. Schmid, R. Perier, V. Praz, P. Bucher, "EPD in its twentieth year: towards complete promoter coverage of selected model organisms," Nucleic Acids Res., vol.34, Jan. 2006, pp.D82–D85.

[9] D. Swarbreck, C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang, E. Huala, "The Arabidopsis Information Resource (TAIR): gene structure and function annotation," Nucleic Acids Res., vol.36, Jan. 2008, pp.D1009–D1014.

[10] S. Gama-Castro, V. Jiménez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Peñaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muñiz-Rascado, I. Martínez-Flores, H. Salgado, C. Bonavides-Martínez, C. Abreu-Goodger, C. Rodríguez-Penagos, J. Miranda-Ríos, E. Morett, E. Merino, A. M. Huerta, L. Treviño-Quintanilla, J. Collado-Vides, "RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation," Nucleic Acids Res., vol.36, Jan. 2008, pp.D120–D124.

[11] A. Pérez, F. Lankas, F. J. Luque, M. Orozco, "Towards a molecular dynamics consensus view of B-DNA flexibility," Nucleic Acids Res., vol.36, Apr. 2008, pp.2379–2394.

[12] M. J. Packer, M. P. Dauncey, C. A. Hunter, "Sequence-dependent DNA structure: dinucleotide conformational maps," J. Mol. Biol.. vol.295, Jan. 2000, pp.71–83.

[13] T. Abeel, Y. Saeys, P. Rouzé, de Peer. Y. Van, "ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles," Bioinformatics, vol.24, Jul. 2008, pp.i24–31.

[14] A. V. Morozov, K. Fortney, D. A. Gaykalova, V. M. Studitsky, J. Widom, E. D. Siggia, "Using DNA mechanics to predict in vitro nucleosome positions and formation energies," Nucleic Acids Res., vol.37, Aug. 2009, pp.707–722.

[15] P. Baldi, Y. Chauvin, S. Brunak, J. Gorodkin, A. G. Pedersen, "Computational applications of DNA structural scales," Proc. Int. Conf. Intell. Syst. Mol. Biol., vol.6, Oct. 1998, pp.35–42.