**B-DNA-VIDEO: AN ACTIVE DATABASE FOR THE SIGNIFICANT B-DNA FEATURES OF TRANSCRIPTION FACTOR BINDING SITES.**

PONOMARENKO M.P., FROLOV A.S., PONOMARENKO J.V., VOROBIEV D.G., LEVITSKY V.G., PODKOLODNAYA O.A., OVERTON G.C.&, KOLCHANOV N.A.

Laboratory of Theoretical Genetics, Institute of Cytology and Genetics, (Siberian Branch of the Russian Academy of Sciences), 10 Lavrentieva ave., Novosibirsk, 630090 Russia

&Center for Bioinformatics, University of Pennsylvania, Philadelphia, USA

Keywords: 15 B-DNA, active database, transcription factor, binding sites, conformational features, physical and chemical features, nucleotide context, knowledge discovery system, knowledge base

The problem of creating methods allowing precise functional site recognition in nucleotide sequences is far from being solved. The increasing volume of experimental data demonstrating that the efficient performance of functional sites is to a large extent determined by their physical and chemical and conformational properties are being accumulated (Kim et al. 1996; Starr et al. 1995). Conformational featureies influence the stereochemical compatibility of the sites and the interacting proteins.

Numerous data obtained so far suggest the local nonuniformity of DNA conformational and physical and chemical features and their dependence on nucleotide context (Suzuki et al., 1997, el Hassan and Calladine, 1996; Yanagi et al., 1991). Numerous experimental and theoretical studies of B-DNA helix resulted in determination of the mean values of a considerable number of conformational and physical and chemical parameters of dinucleotides (Hogan and Austin, 1987; Gartenberg and Crothers, 1988; Sugimoto et al., 1996; Suzuki et al., 1996; Suzuki et al., 1997; Kabsch et al., 1982; Shpigelman et al., 1993; Suzuki and Yagi, 1995; Gorin et al., 1995).



Fig. 1. Schematic representation of the computer system B-DNA-VIDEO.

We suggest here a systemic approach to the problem of revealing significant DNA conformational and physical and chemical features of functional sites implemented in the computer system B-DNA-VIDEO. The computer system B-DNA-VIDEO (Fig. 1) comprises: (1) the database SAMPLE on sequenced DNA regions containing transcription factor binding sites; (2) the database PROPERTY on the values of context-dependent DNA conformational and physical and chemical parameters; (3) the knowledge discovery system on significant DNA site conformational and physical and chemical features; (4) the knowledge base B-DNA-FEATURES containing the knowledge about the significant DNA conformational and physical and chemical features; (5) the library of WWW-available programs for constructing the profiles of conformational and physical and chemical features in arbitrary nucleotide sequences and searching for the regions that are maximally similar in these features to the actual functional sites. The system B-DNA-VIDEO is accessible under the SRS query language (Etzold and Argos, 1993) at URL http://wwwmgs.bionet.nsc.ru/systems/BDNAVideo/. The system allows the access to the information on significant conformational and physical and chemical features of the sites accumulated in the knowledge base of this system. A number of modules developed earlier for the systems SITEVIDEO (Kel et al., 1993) and ACTIVITY (Kolchanov et al., 1998) were used in the system B-DNA-VIDEO. We have applied B-DNA-VIDEO to study the sets of various transcription factor binding sites and demonstrated that the binding sites of each transcription factor analyzed are characterized by a specific set of significant DNA conformational and physical and chemical features.

The significant results obtained through the procedure represented in (Ponomarenko et al., 1998) are stored in the knowledge base of functional sites. An entry of this database is exemplified by the description of transcription factor HNF1 binding site (Fig. 2). The name of the site in question (HNF1) is given in the field NM. The field DA indicates the link to the SAMPLE database, containing the sequences of the site in question used to reveal this B-DNA physical and chemical feature. The field PV contains the name of the B-DNA parameter studied (melting temperature); field DP, link to the B-DNA parameter database, where this parameter is described and its values listed. In this example, the HNF1 binding sites on the region [-21; 4] relative to the center of the site (between -1 and 1 positions) appeared to differ significantly for the random sequences in the mean values of the parameter "melting temperature". Utility of this feature is**U=**0.867. The mean melting temperature on the region [-21; 4] amounted to 67.82± 3.77° C for the site sequences; 73.54± 4.53° C; for the random sequences. Significance of this difference between site and random sequences is a <10**-13**. This information is contained in the fields AB - AL. The field FG indicates the link to the figure illustrating this difference. The distributions of the mean melting temperature calculated over the region [-21; 4] for the real and random sequences are compared in Fig. 3b. Note the left-shift of the distribution for the real sites compared to that for the random sequences. The field C-CODE contains the automatically generated C code of the computer program calculating the value of this parameter from a given DNA sequence. The field WW designed to start the executable program (contained in the field C-CODE) constructing the profiles of the significant conformational and physical and chemical features along an arbitrary DNA sequence and recognizing the regions of this sequence that are most similar to the site in question.

Table lists one most significant feature with the highest utility for some transcription factor binding sites. Location of the region within the site, the mean values of the corresponding parameter on this region of the site and in random sequences, the significance of the difference between these means, and the utility of this feature for discrimination between the site and random sequences are indicated in the table too. The histograms of the mean value of the most significant features for some transcription factor binding sites in comparison with random sequences are given in Fig. 3. The lengths of the significant regions vary in the range of 10 to 25 bp; this corresponds to 1-2.5 coils of B-DNA helix as well as to the mean length of the region of DNA interaction with transcription factors.



Fig. 2. The description of the significant conformational and physico-chemical features of transcription factor HNF1 binding sites in the knowledge base B-DNA-FEATURES.



Fig. 3. Histograms of the mean value for the transcription factor binding sites (black columns) and the random sequences (white columns).

Table. The best significant conformational and physical and chemical features of transcription factor binding sites.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Factor | parametr |   | region | utility | average mean for | Signif |
| Name | name | units | [a;b] | U | SITE | RANDOM | level |
| AP-1 | roll in protein-DNA complexes | degree | -5;17 | 0.664 | **2.97± 0.23** | 2.73± 0.29 | <10-10 |
| c-Fos | probability to nucleosome contact | probability | -10;13 | 0.601 | **0.12± 0.01** | 0.11± 0.01 | <10-3 |
| c-Jun | propeller twist for B-DNA's (NDB) | degree | -11; 6 | 0.614 | **-12.12± 1.00** | -12.58± 0.89 | <0.005 |
| NF-E2 | roll in protein-DNA complexes | degree | -4;19 | 0.918 | **3.25± 0.15** | 2.72± 0.29 | <10-8 |
| CRE-BP1 | propeller twist for B-DNA's (NDB) | degree | -9;12 | 0.609 | **-12.04± 0.60** | -12.57± 0.81 | <0.002 |
| CREB | roll in protein-DNA complexes | degree | -17; 8 | 0.635 | **2.98± 0.37** | 2.72± 0.27 | <10-8 |
| C/EBP | entalpy change | kcal/mol | -17; 7 | 0.651 | **-8.31± 0.46** | -8.63± 0.43 | <10-12 |
| ER | probability to nucleosome contact | probability | -6;17 | 0.819 | **0.13± 0.01** | 0.11± 0.01 | <10-9 |
| RAR | probability to nucleosome contact | probability | -14; 6 | 0.819 | **0.12± 0.01** | 0.11± 0.01 | <10-3 |
| RXR | roll in protein-DNA complexes | degree | -15; 8 | 0.831 | **3.17± 0.33** | 2.73± 0.28 | <10-8 |
| T3R | roll in protein-DNA complexes | degree | -4;10 | 0.916 | **3.29± 0.22** | 2.73± 0.39 | <10-10 |
| COUP | roll in protein-DNA complexes | degree | -9;11 | 0.802 | **3.01± 0.21** | 2.73± 0.31 | <10-3 |
| EN | roll in protein-DNA complexes | degree | -10; 2 | 0.989 | 2.19± 0.22 | **2.76± 0.42** | <10-5 |
| HNF1 | melting temperature | ° C | -21; 4 | 0.867 | 67.82± 3.77 | **73.54± 4.53** | <10-12 |
| OCT | bending towards minor groove | - | -7; 7 | 0.893 | 1.11± 0.03 | **1.14± 0.03** | <10-7 |
| HNF3 | bending towards major groove | - | -6; 6 | 0.881 | **1.09± 0.02** | 1.05± 0.02 | <10-8 |
| c-Myb | wedge angle in free DNA | degree | -5;21 | 0.612 | **4.69± 0.37** | 4.39± 0.45 | <0.005 |
| Ets | probability to nucleosome contact | probability | -11;13 | 0.772 | **0.12± 0.01** | 0.11± 0.01 | <0.002 |

The feature “roll in protein-DNA complex” is one of the most frequent features with the highest utilities (Table): it is typical of the binding sites of transcription factors NF-E2, CREB, EN (engrailed homeodomain), etc. Note that the increased roll of the CREB site is in agreement with the X-ray structure analysis data of GCN4-bZIP-ATF/CREB complex showing the increased DNA roll compared with the standard B-DNA, bending of the center of the DNA toward the major groove, and a widened minor groove as compared with straight B-DNA (Keller et al., 1995). The decreased roll demonstrated for EN site (Fig. 3a) is also consistent with the X-ray structure data on an engrailed homeodomain-DNA complex, demonstrating that the major groove is several angstroms wider than the normal in the region of homeodomain binding (Kissinger et al., 1990).

Thus, the results of this work demonstrate the utility of DNA conformational and physical and chemical properties of the functional sites for revealing a useful information on structure-function organization of these sites and the molecular mechanisms of their function. In addition, the approach described proved to be universal and applicable to analysis of such a qualitatively different objects as transcription factor binding sites, nucleosome binding sites, and homing mobile elements.

We are grateful to Ms. Galina Chirikova for help in translation. This work was supported by NIH (grant 2-R01-RR04026-08A2), Russian National Human Genome Project, Russian Ministry of Science and Technical Politics, Siberian Branch of Russian Academy of Sciences (grants IGSBRAS-97N13), and Russian Foundation for Basic Research.

**References**

1. el Hassan,H.A. and Calladine,C.R. (1996) Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA*. J. Mol. Biol.,***259**, 95-103.
2. Etzold,T. and Argos,P. (1993) SRS - an indexing and retrieval tool for flat file data libraries.*Comput. Applic. Biosci.*,**9,**49-57.
3. Gartenberg,M.R. and Crothers,D.M., (1988) DNA sequence determinants of CAP-induced bending and protein binding affinity.*Nature*,**333**, 824-829.
4. Gorin,A.A., Zhurkin,V.B. and Olson,W.K., (1995) B-DNA twisting correlates with base-pair morphology.*J. Mol. Biol.*,**247**, 34-48.
5. Hogan,M.E. and Austin,R.H., (1987) Importance of DNA stiffness in protein-DNA binding specificity.*Nature*,**329**, 263-266.
6. Kabsch,W., Sander,S. and Trifonov,E.N. (1982) The ten helical twist angles of B-DNA.*Nucleic Acids Res.***10**, 1097-1104.
7. Kel,A.E., Ponomarenko,M.P., et al., (1993) SITEVIDEO: a computer system for functional site analysis and recognition. Investigation of the human splice sites.*Comput. Applic. Biosci.,***9**, 617-627.
8. Keller,W., Konig,P. and Richmond,T.J. (1995) Crystal Structure of a bZIP/DNA Complex at 2.2 Angstroms: Determinants of DNA Specific Recognition.*J. Mol. Biol.,***254**, 657-667.
9. Kim J., de Haan G., Shapiro D.J. (1996) DNA bending between upstream activator sequences increases transcriptional synergy*. Biochemical and Biophisical Res. Communications,***226**, 638-644.
10. Kissinger,C.R., Liu,B., Martin-Blanco,E., Kornberg,T.B. and Pabo,C.O. (1990) Crystal Structure of an Engrailed Homeodomain-DNA Complex at 2.8 Angstroms Resolution: A Framework for Understanding Homeodomain-DNA Interactions.*Cell,***63**, 579-590.
11. Kolchanov N.A., Ponomarenko M.P., et al. (1998) Functional sites in pro- and eukaryotic genomes: computer models for predicting activity.*Mol. Biol. (Mosk),***32**, 266-267.
12. M.P. Ponomarenko, N.A. Kolchanov, J.V. Ponomarenko, A.S. Frolov, O.A. Podkolodnaya, D.V. Vorobiev, N.L. Podkolodny, G.C. Overton (1998) Revealing the conformational and physico-chemical dna properties applicable for predicting the activity of DNA functional sites. In this issue.
13. Shpigelman E.S., Trifonov E.N., Bolshoy A. (1993) CURVATURE: software for the analysis of curved DNA.*Comput. Appl. Biosci.,***9**, 435-140.
14. Starr, D.B., Hoopes, B.C., and Hawley, D.K. (1995) DNA bending is an important component of site-specific recognition by the TATA binding protein.*J. Mol. Biol.*,**250**, 434-446.
15. Sugimoto, N., Nakano, S., Yoneyama, M., and Honda, K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes.*Nucleic Acids Res.*,**24**, 4501-4505.
16. Suzuki M., Amano N., Kakinuma J., Tateno M. (1997) Use of a 3D structure data base for understanding sequence-dependent conformational aspects of DNA.*J. Mol. Biol*.,**274**, 421-435
17. Suzuki M., Yagi N., Finch J.T. (1996) Role of base-backbone and base-base interactions in alternating DNA conformations.*FEBS L.,***379**, 148-152.
18. Suzuki, M., and Yagi, N. (1995) Stereochemical basis of DNA bending by transcription factors.*Nucleic Acids Res.*,**23**, 2083-2091.
19. Yanagi,K., Prive,G.D., and Dickerson,R.E. (1991) An analysis of local helix geometry in three decamers and eight dodecamers.*J. Mol. Biol.,***217**, 201-214.