BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 8 2006, pages 1007–1009
doi:10.1093/bioinformatics/btl059

Structural bioinformatics

Structural bioinformatics of DNA: a web-based tool for the analysis of molecular dynamics results and structure prediction

Surjit B. Dixit* and David L. Beveridge

Department of Chemistry and Molecular Biophysics Program Hall-Atwater Laboratories, Wesleyan University, Middletown CT 06459, USA

Received on December 19, 2005; revised on January 25, 2006; accepted on February 14, 2006
Advance Access publication February 17, 2006
Associate Editor: Martin Bishop

ABSTRACT

Summary: We report here the release of a web-based tool (MDDNA) to study and model the fine structural details of DNA on the basis of data extracted from a set of molecular dynamics (MD) trajectories of DNA sequences involving all the unique tetranucleotides. The dynamic web interface can be employed to analyze the first neighbor sequence context effects on the 10 unique dinucleotide steps of DNA. Functionality is included to build all atom models of any user-defined sequence based on the MD results. The backend of this interface is a relational database storing the conformational details of DNA obtained in 39 different MD simulation trajectories comprising all the 136 unique tetranucleotide steps. Examples of the use of this data to predict DNA structures are included.

Availability: http://humphry.chem.wesleyan.edu:8080/MDDNA

Contact: sdixit@wesleyan.edu

Supplementary information: Supplementary data including color figures are available at *Bioinformatics* online.

Understanding the details of DNA structure on the basis of sequence composition is an important step in being able to predict the structure of longer biologically relevant sequences. Important regulatory control achieved by protein–DNA recognition often rely on subtle intrinsic and protein induced sequence dependent structural deformability of DNA, referred to as the indirect readout (Dickerson, 1983; Travers, 1993). The linear sequence and regular structure of DNA, often depicted as a long linear helix devoid of long-range self interactions, coupled with the presence of only four standard nucleotide building blocks, has suggested that the DNA structure prediction problem is far simpler than the protein folding problem.

Despite this, in the 50 years since the structure of DNA in fiber was solved, understanding and predicting the fine details of DNA structure and the role of sequence has been a challenge (Neidle, 1999). The anomalous migration of certain sequences in gel mobility studies have raised some of the most interesting debates in the field about the origin and nature of the intrinsically bent DNA (Beveridge *et al*., 2004b; Crothers *et al*., 1990; Hagerman, 1990). High resolution X-ray crystallography has been the main experimental resource for detailed structural data of DNA molecules (Berman *et al*., 1996) and quality solution state results using the NMR residual dipolar coupling technique have only recently begun to appear (MacDonald and Lu, 2002). The simulation studies of nucleic acids including solvent at ionic strengths relevant to *in vitro* experiments and *in vivo* phenomena is now well into the second generation and provides useful adjunct to the structural analysis of DNA at atomic resolution (Beveridge *et al*., 2004b). The results of MD on DNA are much improved, and they provide a reasonable description of nucleic acid dynamical structures in solution and their complexes with proteins as shown in the diverse applications of the method (Cheatham, 2004; Giudice and Lavery, 2002; MacKerell, 2004; Norberg and Nilsson, 2002; Orozco *et al*., 2003). Another theoretical approach for DNA structure prediction from sequence information has been proposed recently (Farwer *et al*., 2006).

A major line of investigation of DNA sequence effects has focused around understanding the oligomeric DNA structure in terms of sequence subunits (Yanagi *et al*., 1991). The minimum structural unit that carries information on the 3D structure of DNA is the dinucleotide base pair step, 5′-XpY-3′ where X and Y may be A, T, G or C. The four alternatives lead to 16 XpY permutations of which 10 are unique. Structural studies of DNA have revealed that individual XpY step may clearly be subject to sequence context presented by the nearest neighboring base pairs. Thus the minimum monomeric unit necessary to describe the details of DNA structure would be the tetranucleotide steps, of which there are 136 unique permutations. The availability of such data in the crystallographic database has been sparse although there have been recent attempts in this direction (Hays *et al*., 2005). In an effort to obtain the structural information of all these 136 unique permutations of DNA tetranucleotide in a consistent manner, a consortium of researchers [the *Ascona B-DNA Consortium* (ABC)] performed 15 ns of simulation on 39 different 15mer DNA sequences, which comprised multiple copies of all the 136 tetranucletoides. Details of the simulation and analysis are presented elsewhere (Beveridge *et al*., 2004a; Dixit *et al*., 2005). It should be stressed that this represents a considerable computational task, corresponding to a total of roughly 0.6 μs of simulation for systems containing ∼24 000 atoms. The resulting trajectories involve ∼600 000 coordinate sets and roughly 400 GB of data and the organized analysis of this large dataset calls for novel methods of data management.

A two-tier approach has been adopted for the dissemination of such large datasets. Apart from the data comprising the Cartesian coordinates of the molecular trajectory, the set of intra- and inter-base pair helicoidal parameters, together with the conformational parameters of the sugar–phosphate backbone of DNA (Dickerson *et al*., 1989; Olson *et al*., 2001) provides a smaller but complete set of descriptors to define the fine structural details

*To whom correspondence should be addressed.

Downloaded from bioinformatics.oxfordjournals.org at Periodical Department on July 7, 2011

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

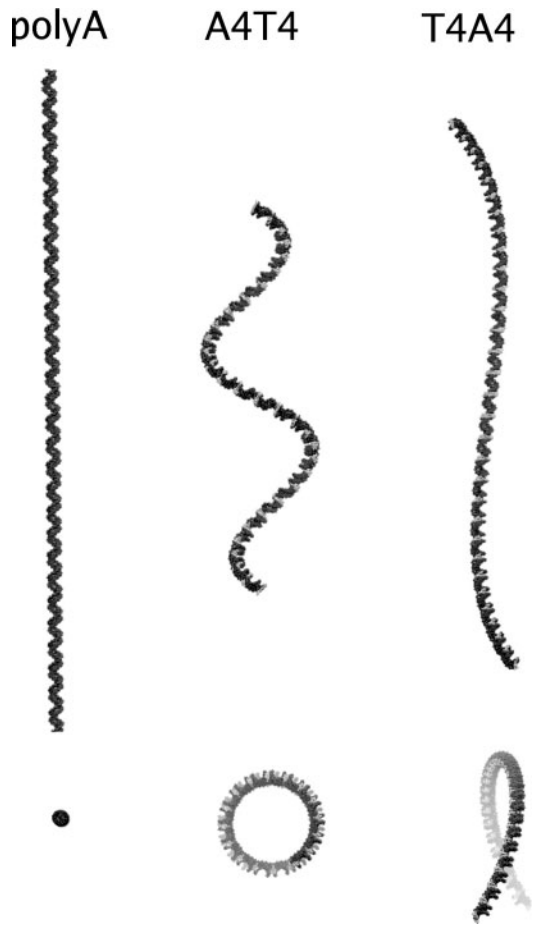

Fig. 1. Models of (left) polyA, (middle) A_4T_4: d(CAAAATTTTG)$_{15}$ and (right) T_4A_4: d(GTTTTAAAAC)$_{15}$ constructed using the structural parameters derived from the MD data. The A_4T_4 sequence forms a superhelix with a diameter of ~120 Å and a pitch of ~100 bp while the T_4A_4 sequence is almost straight with a mild super-helical twist. Both the sequences present a left-handed super-helical twist. The nucleotides are colored red (adenine), blue (thymine), green (guanine) and yellow (cytosine) (see Supplementary data for colour version of figure). The figure on the top presents the side view of the DNA while the lower array of figures presents an orthonormal view down the super-helical axis. Graphics rendered using VMD (Humphrey *et al*., 1996).

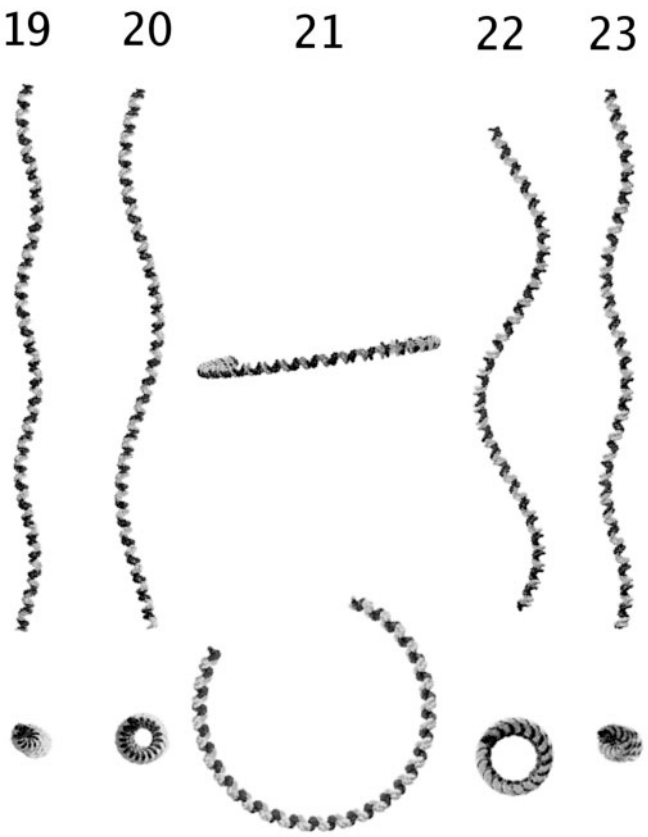

Fig. 2. All atom models of 19mer (AAAAAGGCCAAAAAAGGCCC), 20mer (AAAAAGGGCCAAAAAGGGCC), 21mer (AAAAAGGCCCAAA-AAGGGCCC), 22mer (AAAAAGGGCCCAAAAAAGGCCC) and the 23mer (AAAAAGGGCCCAAAAAAGGGCCC) DNA sequences, each 300 bp long. The 19 and 20mer DNA sequences present a left-handed super-helical twist while the 22 and 23mer sequences exhibit right-handed super-helical twist. The 21mer is closest to forming an almost planar super-helical twist. The observed super-helical twist translates into anamolous gel migration properties in experiments (Brukner *et al*., 1997). The figure on the top presents the side view of the super-helical DNA while the lower array of figures presents an orthonormal view, highlighting the differences in the diameter of the superhelix. The nucleotides are colored red (adenine), blue (thymine), green (guanine) and yellow (cytosine) (see Supplementary data for colour version of figure). Graphics rendered using VMD (Humphrey *et al*., 1996).

of the nucleic acid segments in each of the frames in these trajectories. Assuming the individual bases to be planar and rigid, the helicoidal parameters can be employed to reconstruct detailed DNA structure at the global level. A relational database system has been developed for querying this repository of simulation trajectories. Results from the structural analysis of all the trajectories using the CURVES (Lavery and Sklenar, 1988) and 3DNA (Lu and Olson, 2003) programs are stored in the database and have been indexed on the basis of the nucleotide position, the time step in the simulation, the sequence composition such as dinucleotide step, tetranucleotide step, etc. The use of structured query language at the backend permits execution of complex queries while the web interface provides easy access to a limited number of queries of interest. A component of this project is an initiative aimed at the structured storage and handling of results from large and numerous molecular simulations.

The web interface can perform a number of queries ranging from the analysis of single trajectories to results that rely on the mining of multiple trajectory data. The interface will permit queries to view numerically and graphically the mean and standard deviation profile of all the helicoidal parameters as a function of sequence in each of the trajectories or the time series of the properties in the course of the simulation. The correlations and coupled nature of various structural parameters can be studied. Searches comparing the statistical properties of the DNA structural parameters of a central base pair in all its relevant tetranucleotide contexts can be executed. While the current simulations results are based on the AMBER parm94 (Cornell *et al*., 1995) force field model in explicit solvent, the database setup can be readily expanded to include alternative force fields, simulations methods and solvent representations.

From an application perspective, the helicoidal properties of the unique dinucleotide and tetranucleotide building blocks derived

Downloaded from bioinformatics.oxfordjournals.org at Periodical Department on July 7, 2011

from the MD data can be employed to construct detailed atomistic models of DNA. We present a web interface that permits the user to build such models for any user-defined sequence and conclude here with some examples of structure prediction using this model. As shown in Figure 1, the pure A-tract sequences present the straight B′ DNA structure, the phased A4T4 sequences present a highly curved structure, while the curvature in the corresponding T4A4 sequence is comparatively smaller. The MD data clearly reproduce the experimentally known curved DNA structures obtained in phased A-tract containing sequences after accounting for the undertwisting observed (Beveridge *et al*., 2004a; Dixit *et al*., 2005) in the parm94 force field. In Figure 2 we show example structures of DNA that presents the change in super-helical twist of DNA as a function of the composition and phasing nature of the DNA sequence. The DNA models with different AAAAA and GGCC sequence composition and phasing, changes from the left-handed to almost planar and right-handed forms, in qualitative agreement with the experimental gel mobility result of these sequences (Brukner *et al*., 1997). In conclusion, this WWW resource provides a means for a general user to analyze the results of the ABC trajectories and construct sequence dependent right-handed double helical DNA models that can be useful for testing hypotheses relating sequence to structure.

ACKNOWLEDGEMENTS

The authors thank participants of the ABC for kindly sharing with us trajectories of the 39 DNA simulations and Prof. Michael Rice for useful discussions. We gratefully acknowledge support from NRAC award MCA94P011, NIH grant GM37909, The Keck Center for Integrative Genomics at Wesleyan University and the HHMI grant 52005211.

Conflict of Interest: none declared.

REFERENCES

Berman,H.M. *et al*. (1996) Nucleic acid crystallography: a view from the nucleic acid database. *Prog. Biophys. Mol. Biol*., **66**, 255–288.

Beveridge,D.L. *et al*. (2004a) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. *Biophys. J*., **87**, 3799–3813.

Beveridge,D.L. *et al*. (2004b) Molecular dynamics simulations of DNA curvature and flexibility: helix phasing and premelting. *Biopolymers*, **73**, 380–403.

Brukner,I. *et al*. (1997) Differential behavior of curved DNA upon untwisting. *Proc. Natl Acad. Sci. USA*, **94**, 403–406.

Cheatham,T.E.,III (2004) Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol*., **14**, 360–367.

Cornell,W.D. *et al*. (1995) A second generation force field for the simulation of proteins, nucleid acids and organic molecules. *J. Am. Chem. Soc*., **117**, 5179–5197.

Crothers,D.M. *et al*. (1990) Intrinsically bent DNA. *J. Biol. Chem*., **265**, 7093–7096.

Dickerson,R.E. (1983) The DNA helix and how it is read. *Sci. Am*., **249**, 94–111.

Dickerson *et al*. (1989), Definitions and nomenclature of nucleic acid structural parameters. *EMBO J*., **8**, 1–4.

Dixit,S.B. *et al*. (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J*., **89**, 3721–3740.

Farwer,J. *et al*. (2006) Prediction of atomic structure from sequence for double helical DNA oligomers. *Biopolymers*, **81**, 51–61.

Giudice,E. and Lavery,R. (2002) Simulations of nucleic acids and their complexes. *Acc. Chem. Res*., **35**, 350–357.

Hagerman,P.J. (1990) Sequence-directed curvature of DNA. *Annu. Rev. Biochem*., **59**, 755–781.

Hays,F.A. *et al*. (2005) How sequence defines structure: a crystallographic map of DNA structure and conformation. *Proc. Natl Acad. Sci. USA*, **102**, 7157–7162.

Humphrey,W. *et al*. (1996) VMD: visual molecular dynamics. *J. Mol. Graph*., **14**, 33–38, 27–28.

Lavery,R. and Sklenar,H. (1988) The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn*., **6**, 63–91.

Lu,X.J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*., **31**, 5108–5121.

MacDonald,D. and Lu,P. (2002) Residual dipolar couplings in nucleic acid structure determination. *Curr. Opin. Struct. Biol*., **12**, 337–343.

MacKerell,A.D.,Jr (2004) Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem*., **25**, 1584–1604.

Neidle,S. (ed.), *Oxford Handbook of Nucleic Acid Structure*. Oxford University Press, Oxford, New York.

Norberg,J. and Nilsson,L. (2002) Molecular dynamics applied to nucleic acids. *Acc. Chem. Res*., **35**, 465–472.

Olson,W.K. *et al*. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol*., **313**, 229–237.

Orozco,M. *et al*. (2003) Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev*., **32**, 350–364.

Travers,A. (1993) *DNA–Protein Interactions*. Chapman and Hall, London.

Yanagi,K. *et al*. (1991) Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J. Mol. Biol*., **217**, 201–214.

Downloaded from bioinformatics.oxfordjournals.org at Periodical Department on July 7, 2011