LA-UR- 95-4181

Title: **Predicting B-DNA Structure from Sequence**

DOC
MAY 16 1995
OSTI

Author(s):

**Chang-Shun Tung, T-10
Gerhard Hummer, T-10
Dikeos M. Soumpasis, Max Planck Institute**

Submitted to:

**DOE Office of Scientific and Technical
Information**

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

MASTER
DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED 35

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

Predicting B-DNA Structure from Sequence

Chang-Shun Tung*, Gerhard Hummer, and Dikeos M. Soumpasis (Max Planck Institute)

Abstract

This is the final report of a three-year Laboratory-Directed Research and Development (LDRD) project at the Los Alamos National Laboratory (LANL). This project developed a reliable method that is capable of predicting B-DNA duplex structure from sequence. From any given sequence, the method predicts a complete double helical structure at the atomic level. Tetramers (four base pairs) are used as a basic unit for the study to include the sequence effects from the neighboring base pairs. The equilibrium structures of the 136 distinct tetramers are deduced from Monte Carlo simulations on a set of reduced coordinates developed at LANL. The prediction methods by this project can be used for searching and defining structural motifs in the functional regions of the genes.

1. Background and Research Objectives

Since the late 1970's and the early 1980's, observation based on both solution study and single crystal data have revealed the concept that the structure of the DNA duplex is sequence dependent. Although DNA duplex overall maintains a B-conformation under physiological conditions, the local structures within the duplex can vary significantly (e.g., twist angle can vary from less than 20 to greater than 40 degrees, roll angle can vary from less than -10 to greater than 10 degrees). These local structural variations can accumulate and display some gross structural features (e.g., bent DNA, changes in groove width and depth, etc.) that are significantly deviated from those of the canonical B-DNA structure. These different structural features can serve as recognition motifs for molecules such as proteins, drugs, and ligands that bind to DNA.

The ability to systematically and accurately search and define structural motifs involved in the functional regions (e.g., promoter, operator, replication origin, etc.) of DNA is crucial to the understanding of the mechanisms involved in gene expression and gene control. With the continuing advances in sequencing technology, there is a vast amount of sequence information

*Principal investigator, e-mail: ct@lanl.gov

available today (approximately 150 million base pairs in Release 75 of GenBank). We are anticipating an even more rapid growth in the amount of sequence information due to efforts from the Human Genome Project. If there exists a reliable mechanism for predicting structure from sequence, this vast amount of sequence information can be used to study the structure-function relationship of DNA molecules.

There exist several prediction models that predict DNA structure from sequence. For example, a phenomenological prediction model was first developed by Calladine and later refined by Dickerson. Two prediction models based on data-fitting were developed by Trifonov's and De Santis' groups independently. Prediction models based on conformational energy calculations were developed by Tung & Harvey and Olson's group. Not only do these different prediction models predict different structures, but the next nearest base-pairs interactions (traditionally thought to be very important to the sequence-dependent structure) are also ignored in most of these prediction models.

Based on a set of reduced coordinates, a rigorous sampling approach, and an implicit way of treating solvent effects using the potential of mean force (PMF) approach, we developed a reliable method of predicting DNA duplex structure that uses tetramers as basic units to include the next nearest base-pairs interactions.

2. Importance to LANL's Science and Technology Base and National R&D Needs

The development of a reliable prediction model provides a tremendous tool in deciphering the vast amount of information encoded in genomic sequences produced by the Human Genome Project. The predicted atomic structures of the DNA molecules can be used directly for studying recognition mechanisms between DNA and DNA binding molecules and eventually leads to the complete understanding of gene expression and gene regulation. This type of problem is of great interest to the Structural Biology Initiative which is being emphasized by both LANL and DOE.

3. Scientific Approach and Results to Date

The fundamental concepts behind the development of the prediction method include:

(a) Using tetramer as the fundamental unit for the study to include the next nearest neighbor interactions.

(b) Developing an efficient sampling approach to run Monte Carlo simulations on a set of reduced coordinates to deduce the equilibrium structures of all 136 distinct tetramers.

(c) Constructing user-friendly software to predict DNA duplex structures from sequence.

In the first two years of the project, we concentrated our efforts on improving both the efficiency and accuracy of the sampling approach to deduce the equilibrium structures of different tetramers with boundary conditions set to emulate the condition such that the tetramers are embedded in a continuous B-DNA duplex. To improve the efficiency of the sampling method, we developed a revised set of reduced coordinates for constructing base/base-pair structures. In this revised set, it takes 12 parameters (minimum number required for two rigid bodies) instead of 18 (corresponds to the old set) to describe the structure of a base-pair. By reducing the number of degrees of freedom, efficiency of sampling is increased. A manuscript describing the revised set of reduced coordinates has been published (1). The accuracy of the method was improved by including solvent effects in the calculation and implementing this information in our simulation package. The solvent effects are treated using the PMF approach, with pair correlation calculated from the hypernetted-chain approximation. By implicitly including the solvent effects, the over-represented charge-charge interactions between the phosphate groups will be properly taken care of.

Our sampling method consists of two parts. The first part includes the generating of structures of the tetramer that cover evenly in the conformational space and have energies below a preselected cut-off value. A Latin hypercube method was used to generate structures that cover the conformational space. Five thousand structures were initially generated for each tetramer. The cut-off energy was chosen such that a desired number of structures (50 in this study) were selected. The second part includes the equilibration of these selected structures with a run (3,000 accepted structures) of Monte Carlo simulations. The equilibrium structure of the tetramer is calculated from the energy-weighted structures deduced from these 50 runs of Monte Carlo simulations.

A set of conceptually straightforward boundary conditions were used to replace the existing set to simulate the condition that the tetramer is embedded in a long polymer. These include: B-type sugar conformations for the end deoxyriboses, averaged propeller twist angles (-15 degrees), and zero buckle angles for the end base pairs. A small energetic penalty is applied to structures with an overall bending greater than 10 degrees and an overall twist either smaller than 102 degrees or greater than 108 degrees. The inversion rules (structural parameters deduced from the two strands of the base-pair step are identical except for the sign reversal for tilt and x-displacement) are carefully implemented and maintained in our simulation algorithm.

Figures 1 and 2 show the predicted twist angles and roll angles for the dodecamer (CGCGATTCGCG) together with structures predicted by three other prediction models. All

these predictions were compared to the averaged crystal value (plotted as diamond with error bar). The averaged crystal structure of a particular base-pair step embedded in its two neighboring base pairs (a tetramer) is calculated from all the tetramers of the same sequence in a set of 56 oligonucleotide structures in PDB. All predictions correlated well with the crystal data for this dodecamer. Besides being based on physical ground, our new prediction method is the only method that has taken into account the next nearest neighbor interactions directly. A manuscript for this work is in preparation. A program that generates predicted structure from sequence based on our new prediction method will be available to the public in the near future.

One of the necessary conditions for the sequence-dependent structure of DNA to involve directly in biological recognition is a structural correlation between the bases (internal of the duplex) and the sugar-phosphate backbone (external of the duplex) of the DNA. The narrowed minor grooves associated with the A, T rich regions and the zig-zag backbone pattern of the left-handed DNA duplexes are two experimental evidences that support the notion that the base structure and the backbone structure are highly correlated. Using tetramer as the basic unit and a least-square-fitting algorithm, we have developed a method to predict the complete structure of a DNA based on the structure of the phosphorus atoms. We tested this prediction method on ten oligonucleotides with their crystal structures solved and deposited in Protein Data Bank. Good agreement was obtained between the two sets of structures. The RMS differences for base atoms between the predicted and the crystal structures are less than 0.5 Angstrom. Due to the additional constraints imposed by base-pairing, the RMS differences for base atoms are smaller than those for sugar atoms in all molecules we had tested. Our results support a strong correlation between the structures of bases and phosphorous atoms. A manuscript about this work has been submitted to Biophysical Journal (2). This work opens the possibility of an alternative DNA sequencing method to determine base sequence from the structure of the phosphorus atoms.

Using the method as described in the previous paragraph, we have constructed an atomic modeled structure of a 17 base-pair DNA operator (cro, from phage lambda) with the phosphorus structures solved by x-ray crystallography. With this predicted DNA structure and modeled structures of the alpha-3 helix based on the C-alpha atoms solved by x-ray crystallography, we were able to predict (3) two specific interactions between the cro protein and the DNA (Ser-28 to Gua-14, Lys-32 and Gua-12) as shown in Figure 3. These interactions were partially verified by NMR using N-15 labeled DNA operator (3).

References

1. Tung, C-S., Soumpasis, D. M. and Hummer, G. "An extension of the rigorous base-unit oriented description of nucleic acid structures," J. Biomol. Struct. and Dyn. 11: 1327-1344 (1994).
2. Tung, C.-S., "Structural prediction of B-DNA duplexes based on coordinates of the phosphate atoms," Biophys. J., submitted (1995).
3. Edwards, C.A., Gatewood, G. M., Tung, C.-S., Fee, J. A. and Santhana Mariappan, S. V., "Binding of Cro protein to Nitrogen-15 labeled operator DNA indicates specific interaction of Cro with N7 of guanosine-14," J. Biol. Chem., submitted (1995).

Fig. 1. Twist angles of the dodecamer CGCGAATTGCGG from four different prediction methods and those from crystal structures.

Fig. 2. Roll angles of the dodecamer CGCGAATTGGCG from four different prediction methods and those from crystal structures.

Fig. 3. Modeled structure of the operator DNA and two side chains of the Cro protein that form specific H-bonds with the DNA. The all-atom structure of the DNA (in grey) together with the phosphorous and C-alpha atoms (in black) of the crystal structure (4CRO) are plotted on the left side of the image. A section of the operator DNA and the alpha-3 helix of the protein are plotted on the right side of the image. Ser-28 hydroxyl and Lys-32 amino protons form H-bonds with G-14 and G-12 respectively.