Sequence−Structure Relationships in DNA Oligomers: A Computational Approach

Martin J. Packer* and Christopher A. Hunter

Contribution from the Krebs Institute for Biomolecular Science, Department of Chemistry, University of Sheffield, Sheffield, S3 7HF England

Received September 13, 2000

Abstract: A collective-variable model for DNA structure is used to predict the conformation of a set of 30 octamer, decamer, and dodecamer oligomers for which high-resolution crystal structures are available. The model combines an all-atom base pair representation with an empirical backbone, emphasizing the role of base stacking in fixing sequence-dependent structure. We are able to reproduce trends in roll and twist to within 5° across a large database of both A- and B-DNA oligomers. A genetic algorithm approach is used to search for global minimum structures and this is augmented by a grid search to identify local minimums. We find that the number of local minimums is highly sequence dependent, with certain sequences having a set of minimums that span the entire range between canonical A- and B-DNA conformations. Although the global minimum does not always agree with the crystal structure, for 24 of the 30 oligomers, we find low-energy local minimums that match the experimental step parameters. Discrepancies throw some light on the role of crystal packing in determining the solid-state conformation of double-helical DNA.

Collective-variable models for DNA continue to be of use in predicting and rationalizing structure and function. While all-atom models have developed to the point that valid predictions can be made about sequence-dependent structure,[1,2] they are computationally expensive. This means that one cannot easily explore sequence space (even at the dodecamer level there are more than 8 million possible sequences), and it is not easy to rationalize sequence-dependent variations based on a limited computational data set. Collective-variable models, by contrast, represent various aspects of DNA structure in a concerted fashion and so remove many degrees of freedom that seem to contribute little to overall sequence dependence.[3] This enables study of the structure and dynamics of much longer sequences. In this paper, we demonstrate the utility of a simple collective-variable model which we have recently described.[4] We have already used this approach to explain sequence context effects in tetranucleotides,[4] and we now apply it to a set of 30 DNA oligomers for which high-resolution X-ray crystal structures are available.

A significant number of collective-variable models have been proposed and are in use (see ref 3 for a recent review). They emphasize different aspects of DNA structure and are targeted at the prediction of different properties. They range from models that include both base and backbone explicitly[5] to single-variable wedge models used to explain the bending of oligomers.[6] While single-variable models capture the essence of sequence-dependent structure, they are unable to account for sequence context effects, where steps have different conformations dependent on neighboring steps. Such context effects are clearly apparent in crystal structure surveys.[7,8] Extending to models with trinucleotide motifs[9,10] can help to account for context. However, the large number of possible tetranucleotides means that there is insufficient experimental data available to parametrize models that go beyond the trinucleotide building block. An important advantage of the model we describe here is that context effects are implicit and operate cooperatively along the entire sequence. In addition, the simplicity of the model means that it can be applied to very long oligomers.

The accurate prediction of DNA oligomer structure has many potential applications. Processes such as packaging and transcription involve looping of DNA around multiprotein complexes.[11,12] Bending models that use a single value of roll for each step are not sufficiently flexible to describe the structures of such complexes. They also fail to capture the intrinsic flexibility of some sequences that can adopt multiple low-energy conformations. The model we describe here uses all six step parameters as variables and is based on a potential energy surface over the full range of slide and shift, with roll, twist, tilt, and rise optimized at all values of these two primary degrees of freedom. It is therefore possible to make detailed predictions about the conformation and flexibility of extended sequences. The aim of this paper is to benchmark the performance of the

* Corresponding author. Tel: (+44) 0114 2229456. Fax: (+44) 0114 2738673. E-mail: m.j.packer@shef.ac.uk.

(1) Sprous, D.; Young, M. A.; Beveridge, D. L. *J. Phys. Chem. B* **1998**, *102*, 4658−4667.
(2) Tsui, V.; Case, D. A. *J. Am. Chem. Soc.* **2000**, *122*, 2489−2498.
(3) Lafontaine, I.; Lavery, R. *Curr. Opin. Struct. Biol.* **1999**, *9*, 170−176.
(4) Packer, M. J.; Dauncey, M.; Hunter, C. A. *J. Mol. Biol.* **2000**, *295*, 85−103.
(5) Lavery, R.; Zakrzewska, K.; Sklenar, H. *Comput. Phys. Commun.* **1995**, *91*, 135−158.
(6) Bolshoy, A.; McNamara, P.; Harrington, R. E.; Trifonov, E. N. *Biophys. J.* **1990**, *57*, A 454−A 454.
(7) El Hassan, M. A.; Calladine, C. R. *Philos. Trans. A* **1997**, *355*, 43−100.
(8) Suzuki, M.; Amano, N.; Kakinuma, J.; Tateno, M. *J. Mol. Biol.* **1997**, *274*, 421−435.
(9) Crothers, D. M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 15163−15165.
(10) Brukner, I.; Sanchez, R.; Suck, D.; Pongor, S. *Embo J.* **1995**, *14*, 1812−1818.
(11) Wentzell, L. M.; Halford, S. E. *J. Mol. Biol.* **1998**, *281*, 433−444.
(12) Travers, A. *DNA-protein interactions*, 1st ed.; Chapman & Hall: London, 1994.

10.1021/ja003385u CCC: $20.00
Published on Web 07/07/2001

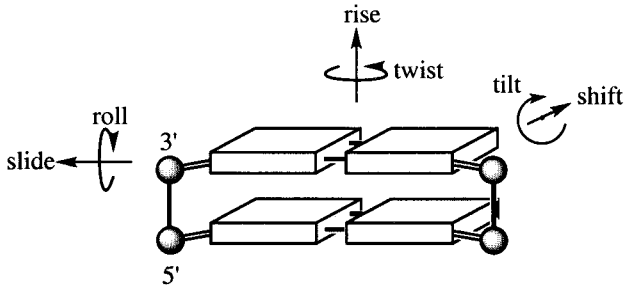

Figure 1. Definition of the six base step parameters defined with respect to the upper base pair. Arrows point in the direction of positive values. The minor groove is toward the viewer.

model against X-ray crystal structures of DNA oligomers for which correspondingly detailed experimental structural data are available. Reproducing the conformations of these oligomers is a crucial test of the model and provides confidence that longer sequences where high-resolution structural data are lacking can also be modeled accurately.

Materials and Methods

The model we have used to study oligomers is similar to that introduced in previous work,[4,13,14] and we refer to those papers for full details. A new feature of the work described here is an additional potential energy term that accounts for the base−backbone interaction (see below).

Base Step Geometry. We used a local description for the base step geometry, as described in the Cambridge convention on DNA conformation.[15] The definition of the six base step parameters is shown in Figure 1. We adopted the implementation of El Hassan and Calladine,[16] which defines a midstep triad for each step. This ensures that the step parameters are independent of both the step context and the direction in which the step is reckoned (with the exception of shift and tilt, which change sign when the step is reckoned in opposite directions).

Experimental Data. The experimental database with which we compare our results was generated by El Hassan and Calladine.[7] It consists of 400 individual dinucleotide steps, extracted from 60 crystal structures of naked DNA oligomers (i.e., not complexed to protein or any other molecule). Only 30 of the oligomers are represented in the database by all their constituent steps. In other oligomers, it was necessary to exclude various steps, because of non-Watson−Crick base pairing, inosine or uracil substitutions, or inconsistent backbone conformations. We refer to the paper of El Hassan and Calladine[7] for full details of these exclusions. We concentrated on predicting the structures of the 30 complete oligomers from this database: their NDB accession codes are listed in Table 2. The experimental step and base pair parameters were obtained using the SCHNAaP nucleic acid analysis program.[17]

Base Step Energy. The interaction energy of each of the 16 base steps was calculated using a standard exp-6/van der Waals potential with atom-centered σ-charges and off-atom π-charges.[18] Out-of-plane π-charges provide a better description of the electrostatic potential around an aromatic system than atom-centered charges alone.[19] The constraints on stacking geometry due to the backbone were modeled using a semiflexible rod, whose properties are dependent only on the values of slide and shift.[13] The base pair parameters were fixed at their average values from the dinucleotide database: for example, for AA/TT the lower propeller was −18.0° and the upper propeller −19.5°, while for CG both propellers were −10.1°.[18] A grid of 32 slide values between −3 and +3 Å and a grid of 32 shift values between −2 and +2 Å was constructed, and at each grid point the step parameters twist, roll, rise, and tilt were optimized with respect to the base step energy. This gave a total of 1024 grid points for each step. This model has been used previously to generate dinucleotide conformational maps which account for many aspects of sequence-dependent behavior.[14]

Optimization of Oligomer Energy. The energy of an oligomer of length N was expressed as the sum of the base step energies, E_{step} plus step junction contributions, E_{junction}, as previously applied to tetranucleotides:[4]

$$E^N_{\text{oligomer}} = \sum_{n=1}^{N-1} E^n_{\text{step}} + \sum_{n=1}^{N-2} E^n_{\text{junction}} \qquad (1)$$

E^n_{step} is the energy of the nth step, taken from the slide/shift grid described above. E^n_{junction} is a backbone penalty function which gives a positive (i.e., destabilizing) energy contribution if two neighboring steps have different slide or do not have shift values of opposite sign:

$$E_{\text{junction}} = 5.0(\Delta D_y)^2 + 3.0(\sum D_x)^2 + 0.5(\Delta D_x)^2 \qquad (2)$$

where ΔD_y is the difference in slide between two neighboring steps, ΔD_x is the difference in shift, and $\sum D_x$ the sum of shift. The structural features of the backbone that lead to the correlation of slide and anticorrelation of shift were discussed in detail in a previous paper.[4] The small ΔD_x term reflects the fact that shift tends to zero in most structures. Note that E^N_{oligomer} is a function of slide and shift only: the other step parameters have been optimized for each pair of slide/shift values with respect to E_{step}.[14]

Optimization Strategy. Two complementary approaches were used to optimize E^N_{oligomer}: a genetic algorithm search for the global minimum structure and a grid search for local minimums. We add the caveat that, for oligomers of N base pairs, there are potentially $1024^{(N-1)}$ possible conformations available from our grids, but a much smaller number of minimums. It is not possible to locate the global minimum with certainty, since this would involve a complete enumeration of all possible conformations. We use the term global minimum to refer to the lowest energy structure located from our calculations. The fact that the genetic algorithm and grid search usually locate the same global minimum using very different criteria indicates that we are searching the conformational space thoroughly.

Genetic Algorithm Optimization. Genetic algorithm (GA) optimization proceeds by encoding solutions to a problem as a string of binary digits called a "chromosome".[20,21] The coding dictates that each chromosome has a specific fitness which corresponds to the quantity that is to be optimized (in our case the energy of the oligomer). The chromosome can be further subdivided into discrete "genes", where each gene represents a variable. The optimum solution corresponds to the chromosome of highest possible fitness. Optimization is usually started with a randomly generated population of chromosomes, which are combined via "crossover" and "mutation" operations to produce a new generation of chromosomes of higher average fitness. Crossover involves two chromosomes that are split at a random position (one-point crossover) and then recombined with each other to produce two new chromosomes. It is also possible to use two-point or higher crossover, splitting each chromosome at several points; this has the effect of increasing coverage of the search space at the expense of a greater number of generations. Mutation involves a single chromosome, with one or more bits changed from 0 to 1 or vice versa. Both crossover and mutation have the effect of producing chromosomes of a fitness different from the original population. A "reproduction" cycle selects chromosomes for crossover and mutation on the basis of their fitness, with high-fitness chromosomes having an enhanced reproduction rate. The old population is then replaced by the new population (although certain GA protocols retain a fraction of the old population). Selection

(13) Packer, M. J.; Hunter, C. A. *J. Mol. Biol.* **1998**, *280*, 407−420.
(14) Packer, M. J.; Dauncey, M.; Hunter, C. A. *J. Mol. Biol.* **2000**, *295*, 71−83.
(15) Diekmann, S. *J. Mol. Biol.* **1989**, *205*, 787−791.
(16) El Hassan, M. A.; Calladine, C. R. *J. Mol. Biol.* **1995**, *251*, 648−664.
(17) Lu, X. J.; El Hassan, M. A.; Hunter, C. A. *J. Mol. Biol.* **1997**, *273*, 668−680.
(18) Hunter, C. A.; Lu, X. J. *J. Mol. Biol.* **1997**, *265*, 603−619.
(19) Hunter, C. A.; Sanders, J. K. M. *J. Am. Chem. Soc.* **1990**, *112*, 5525−5534.
(20) Goldberg, D. E. *Genetic algorithms in search, optimization, and machine learning*; Addison-Wesley: Reading, MA, 1989.
(21) Davies, L. *Handbook of Genetic Algorithms*, 1st ed.; Van Nostrand Weinhold: New York, 1991.

of chromosomes for reproduction based on their fitness ensures that the new population will have a higher average fitness than the old. As the number of generations increases the fitness will increase, producing better solutions to the problem. The search can be run for as many generations as desired or halted when the fitness has failed to change for a certain number of generations. The ability to reach a good solution depends on numerous factors, including population size, number of chromosomes that are replaced in each generation, and crossover and mutation rates. In general, a higher number of generations will produce a better solution, but the population will finally reach a point where no crossover or mutation operation can produce a chromosome of higher fitness than those that already exist (e.g., if the global minimum has been located).

An oligomer sequence consisting of N base pairs was represented by a chromosome of $(N - 1)$ genes, each gene denoting a slide/shift grid point for a base step. Since the grids consist of 1024 (2^{10}) points, genes were encoded by an 10-bit binary string. The fitness of a chromosome was given by $-E^N_{\text{oligomer}}$ and optimization consisted of maximizing this fitness. A steady state without duplicates genetic algorithm was used,[21] a protocol designed for optimization problems in which a global minimum can be located. The replacement rate was set at 10 per generation for a population of 50 chromosomes, but replacement occurred only if the new chromosome was better than all existing chromosomes. This maintained diversity within the population but ensured that the GA made steady progress toward a minimum energy structure, since the fittest chromosome was never discarded. One-point crossover occurred in every generation with a probability of 0.8, and two- and three-point crossover occurred with a probability of 0.2. The mutation rate was set at one mutation per chromosome per generation. A population of 50 randomly generated chromosomes was used as a starting point for each run, which continued until the oligomer energy had not changed for 30 000 generations. At this point, crossover was switched off, and the mutation rate was increased to 1 in every 10 genes with an average mutation rate of 10 grid points (i.e., if the grid point is 700, it can mutate in the range 700 ± 10 with a Gaussian distribution). This final stage provided a local search and was continued until a further 30 000 generations passed without variation in the energy. The last step was a grid search from the best chromosome to locate a minimum energy structure within the grid, since the GA is unable to do this (see below for details). Each search required in the region of 100 000 generations.

In the case of a tetranucleotide sequence, it was possible to locate the global minimum by examining all 1024^3 conformations. This provided a benchmark for the GA optimization, enabling us to test the number of generations required to reach the global minimum for different values of crossover and mutation rate and population size. The parameter set described above was able to locate the global minimum within 2000 generations, but reducing the population size or neglecting two- and three-point crossover led to searches that required more generations or that located a local minimum only. The nonzero probability for two- and three-point crossover is desirable to maintain diversity within the population, but the probability for one-point crossover should always dominate.[21] Higher rates of mutation are also possible but would slow convergence to the global minimum by exploring low-fitness regions. In principle, each sequence will have a different set of optimum parameters that yield the global minimum in the smallest number of generations.[21] We therefore made no attempt to define an optimum set based on the tetranucleotide.

Given the simplicity of the fitness function, the search for a dodecamer global minimum requires only 6 CPU min on a Pentium III 500-MHz workstation running under GNULinux. The GA routines were taken from the SUGAL package[22] and adapted to reproduce the steady-state protocol.

Grid Search for Local Minimums. While the GA provides an efficient method for exploring conformational space and exploiting regions of low energy, it is not suited to generating ensembles of low-energy structures. For this reason, we complemented the GA with a simple grid search. Given a set of grid points, the first base step was moved to each of its four neighboring points (i.e., two for slide and two for shift), and the resulting energy was evaluated at each new point. If one or more of these moves resulted in an oligomer of lower energy, the base step was moved to that grid point giving the biggest improvement; otherwise it was unchanged. This process was repeated for each base step along the oligomer, returning to the first base step at the end of the sequence. The process continued until every move was uphill, resulting in a local minimum structure for the oligomer. Since the step junction term in the potential energy function forces similar slide values in neighboring steps (eq 1), our starting points for the grid search were taken as the 1024 structures in which all grid points are the same.

The GA provides a highly efficient method for searching conformational space. It is especially suited to a problem of this type, since crossover and mutation operations have the effect of retaining base step motifs of low energy. It is also insensitive to its starting point, which is important in searching a large space. The grid search, by contrast, is completely dependent on its starting point but has the advantage that it can readily identify local minimums by using a range of initial conformations and can aid in identifying conformational basins of attraction in a polymorphic oligomer. The two methods are therefore complementary and provide more information when used in tandem than either could in isolation.

Initial application of eq 1 generated a large number of minimums which had very high positive slide: 5 of the 30 oligomers gave this structure as the global minimum, but there are no examples of such a conformation in the crystal structures examined. We previously noted low-energy high-slide structures in our analysis of tetranucleotide conformations, but these appear to be an artifact of the model. The stability of high positive slide conformations in CG/CG, GG/CC, and CA/TG steps can be clearly understood in terms of electrostatic interactions.[14] The fact that it is not generally observed experimentally therefore suggests that we have neglected a factor unrelated to base stacking. One important term that is missing from the model is the interaction between the bases and the backbone. At the simplest level, this can be represented as the torsional potential of the χ torsion angle[13] which shows a distinct minimum for $\chi \approx 180°$,[23] since this minimizes the steric clash between the base and the furanose ring. There is a strong correlation between the value of χ and slide, and so we can account for the base−backbone interaction by parametrizing a new potential energy term, E^n_{sugar}, which is a function of slide

$$E^N_{\text{oligomer}} = \sum_{n=1}^{N-1} E^n_{\text{step}} + \sum_{n=1}^{N-2} E^n_{\text{junction}} + \sum_{n=1}^{N} E^n_{\text{sugar}} \qquad (3)$$

where

$$E^n_{\text{sugar}} = F_y(D^n_y - D^{\text{opt}}_y)^2 \qquad (4)$$

This introduces an energy penalty if slide is not equal to D^{opt}_y at each base pair. For nonterminal base pairs, the value of slide (D^n_y) was averaged over the two steps in which the base pair was located. The parameters F_y and D^{opt}_y were obtained by a chi-squared (χ^2) minimization procedure applied to the 229 tetranucleotide sequences in the 30 complete oligomers. For each tetranucleotide, the two outer steps were fixed at their experimental geometries, and E^N_{oligomer} was minimized with respect to slide and shift of the central step. A χ^2 value was then calculated as a sum over twist, roll, slide, and shift. F_y was varied in increments of 0.1 from 0 to 2.0 and D^{opt}_y in increments of 0.1 from −3.0 to +3.0 Å. The χ^2 minimum occurred for $F_y = 0.5$ and $D^{\text{opt}}_y = -1.0$ Å. The value of D^{opt}_y is consistent with a minimum energy base−backbone interaction when the glycosidic torsion $\chi \approx 180°$,[23] and the function E^n_{sugar} will clearly increase the potential energy of the very high slide conformations which are not observed experimentally. The final model includes five adjustable parameters: the three parameters in eq 2, which were previously determined,[4] and F_y and D^{opt}_y in eq 3.

(22) Hunter, A. *SUGAL−Sunderland Genetic Algorithms package*; University of Sunderland, 1985.

(23) Foloppe, N.; Mackerell, A. D. *J. Phys. Chem. B* **1999**, *103*, 10955−10964.

Table 1. Root-Mean-Square Deviations between Calculated and Experimental Step Parameters Obtained from the Global Minimum Structures of the 30 Complete Oligomers (see Table 2)[a]

	$E_{step} + E_{junction}$	$E_{step} + E_{junction} + E_{sugar}$
twist (deg)	6.78	5.20
roll (deg)	6.05	5.59
tilt (deg)	2.20	2.14
rise (Å)	0.24	0.24
slide (Å)	1.22	0.93
shift (Å)	0.52	0.50

[a] Introducing E_{sugar} significantly improves the slide−roll−twist degree of freedom.

The performance of this refined model applied to the complete oligomers is summarized in Table 1. The root-mean-square deviations (RMSD) with respect to experiment for eq 3 are compared with the old model (eq 1). Introducing the base−backbone function leads to a clear improvement in the slide−roll−twist degree of freedom, which is strongly coupled to χ. The RMSDs for shift and tilt in the oligomer calculations are not improved by the introduction of the base−backbone interaction.

Results

We first discuss the structures of the 30 oligomers from the dinucleotide database that formed part of our test set for parametrizing the model and then apply the approach to two recently crystallized sequences that were not in our test set.

The results of the conformational searches are summarized in Table 2. The first thing to note is that the number of local minimums located in the grid search differs dramatically from one sequence to the next. For each sequence. the standard deviations of slide and energy (σ_{D_y} and σ_E) for these local minimums provide a crude measure of flexibility. Some sequences adopt a small number of well-defined structures; for example, adh038 has only five local minimums, spread over a small range of slide and energy ($\sigma_{D_y} = 0.1$, $\sigma_E = 0.3$). Others are conformationally mobile and have a large number of accessible structures, for example. adh007, which has 30 local minimums with a large range of slide and energy ($\sigma_{D_y} = 0.5$, $\sigma_E = 0.8$).

To compare the quality of the calculated structures with the crystal structures, we will focus on the value of slide, since this is the primary degree of freedom which defines major structural differences.[13] In Table 2, the calculated structures are ranked in order of the RMSD between experimental slide and the value at the global minimum. In addition, the difference in energy per base step between the calculated and experimental structures, ΔE^{exp}, is listed. Since we used a discrete energy function, the energy of the experimental oligomer structure was defined using the slide and shift values of the closest grid point. The predicted structures have been classified as either A- or B-form, based on the mean value of slide over the whole oligomer (mean slide less than −0.5 Å is an A-type structure). In the case of the global minimums, eight oligomers predicted the wrong polymorph (shown as lowercase a or b), but if we consider the best local minimums, there are only three discrepancies.

The A- and B-DNA oligomers with the lowest RMSD values at the global minimum are adh030 and bdl001, respectively. These structures are examined in detail in Figures 2 and 3. We note that both sequences are symmetric, as are the calculated structures, but that the crystal structures do not reflect this symmetry. This presumably reflects the impact of crystal packing forces which lead to minor distortions away from the optimum structure. Table 2 shows that the energy required to distort the global minimum structure into the experimental conformation is generally very small, and this provides an indication of the magnitude of the crystal packing forces involved. The bdl001 structure is particularly informative from this point of view. This sequence has been crystallized four times under different conditions and in different space groups,[24,25,26] yet the structure of the DNA is remarkably similar in all cases (Figure 3). Thus, the intrinsic conformational preferences of the DNA are more important than crystal packing forces in this system.

Figures 2 and 3 illustrate that the agreement between experimental and calculated slide is generally mirrored in the behavior of roll, twist, and shift (rise and tilt are essentially sequence-independent and will not be discussed here). We will consider the four major step parameters in turn.

Slide: There is generally only a small absolute error in slide across the entire sequence with most variation in the terminal steps, as exemplified by bdl001 (Figure 3) where the end steps have higher positive slide than we predict. It is possible that crystal contacts play a role in fixing the conformation of the terminal steps,[27] but we made no attempt to account for these in the optimization.

Roll: The alternation of roll observed experimentally is reproduced. In bdl001, for example, we find that the central AT step has negative roll and is flanked by AA/TT steps with small positive roll, while in adh030, we correctly predict high positive roll for the central TA step. Our success in predicting the variation in roll along the oligomer indicates that this model will be able to reproduce patterns of bending in longer sequences which rely most strongly on roll.

Twist: The prediction of twist is generally inferior to other step parameters. While we generally predict the correct trend for twist, we find much less variation between steps than is observed experimentally. This probably arises because we impose a strong correlation between twist and slide via the model backbone. Experimentally there is somewhat more variation in twist than we allow.[28] We fail to predict unusually low values of twist, such as that seen in the inner CG steps of bdl001 (Figure 3).

Shift: The anticorrelation of shift between steps apparent in the experimental structures is clear from our calculations. However, we noted previously that it is possible for shift to be out of phase with experiment, and this has indeed happened in bdl001 (Figure 3). There is a local minimum with step parameters identical to the one plotted, but with shift of opposite sign in each step. This conformation is therefore in phase with the experimental shift. The energy of this local minimum is only 0.05 kJ mol^{-1} higher than the global minimum, illustrating the degeneracy in shift for symmetric steps such as CG and GC. By contrast, correct phasing is obtained for adh030, probably because the AC/GT step has a strong preference for positive shift.[14]

We now turn to some of the structures at the bottom of Table 2, where our model does not perform so well. The results of the grid search provide a straightforward explanation for some of these structures. For example, although the global minimum structure for adj022 is a B-DNA structure, this sequence has 30 local minimums, one of which agrees extremely well with the A-DNA structure which was crystallized. Figure 4 shows

(24) Shui, X.; Sines, C. C.; McFail-Isom, L.; van der Veer, D.; Williams, L. D. *Biochemistry* **1998**, *37*, 16877−16887.
(25) Liu, J.; Subirana, J. *J. Biol. Chem.* **1999**, *274*, 24749−24752.
(26) Johansson, E.; Parkinson, G.; Neidle, S. *J. Mol. Biol.* **2000**, *300*, 551−561.
(27) Dickerson, R. E.; Goodsell, D. S.; Neidle, S. *Proc. Nat. Acad. Sci. U.S.A.* **1994**, *91*, 3579−3583.
(28) Gorin, A. A.; Zhurkin, V. B.; Olson, W. K. *J. Mol. Biol.* **1995**, *247*, 34−48.

Table 2. Results of the Conformational Searches for 30 Oligomers[a]

					global minimum from GA			best local minimum from grid search		
NDB code	sequence	no.	σ_{D_y}	σ_E	R	ΔE^{exp}	form	R	ΔE^{min}	form
adh030	GGGTACCC	15	0.6	2.6	0.2	0.7	A	0.2	0.0	A
adh012	CCCCGGGG	9	1.0	0.3	0.4	2.1	A	0.4	0.0	A
adh038	GTGTACAC	5	0.1	0.3	0.4	0.7	A	0.3	0.1	A
adh026	GGGCGCCC	12	0.7	4.3	0.5	2.7	A	0.4	0.0	A
adh008	GCCCGGGC	32	0.6	2.1	0.5	1.7	A	0.5	0.0	A
bdl001	CGCGAATTCGCG	26	0.1	0.6	0.5	2.3	B	0.5	0.2	B
ahj040	GGGTATACGC	16	0.3	1.0	0.5	1.5	A	0.4	0.1	A
bdj031	CGATTAATCG	13	0.1	0.5	0.6	3.3	B	0.4	0.1	B
bdl029	CGTGAATTCACG	12	0.1	0.6	0.6	2.0	B	0.4	0.2	B
adh006	GGGGCCCC	22	0.8	6.6	0.6	2.4	A	0.5	0.1	A
bdl006	CGCAAAAAAGCG	18	0.1	0.3	0.6	4.0	B	0.5	0.0	B
bdl047	CGCGAAAAAACG	17	0.1	0.1	0.6	3.1	B	0.5	0.0	B
adh014	GTGTACAC	5	0.1	0.3	0.7	3.4	A	0.5	0.1	A
adh007	GGGATCCC	30	0.5	0.8	0.7	1.8	A	0.4	0.1	A
bdl038	CGCAAATTTGCG	27	0.1	0.3	0.8	1.9	B	0.6	0.1	B
bdl015	CGCAAAAATGCG	24	0.1	0.4	0.8	5.0	B	0.7	0.1	B
bdj036	CGATATATCG	12	0.1	0.2	0.8	3.4	a	0.7	0.0	B
bdl047	CGCGAAAAAACG	17	0.1	0.1	0.9	3.6	B	0.7	0.0	B
bdl042	CGTAGATCTACG	14	0.1	0.2	0.9	6.9	B	0.8	0.0	B
bdl047	CGCGAAAAAACG	17	0.1	0.1	0.9	3.2	B	0.7	0.0	B
bdl015	CGCAAAAATGCG	24	0.1	0.4	0.9	8.8	B	0.8	0.1	B
bdj039	CCGGCGCCGG	39	0.6	2.4	0.9	2.4	B	0.8	0.1	B
adh024	GTACGTAC	7	0.1	0.1	0.9	3.8	A	0.9	0.1	A
bdl007	CGCATATATGCG	20	0.2	0.6	1.0	3.7	a	0.9	0.1	B
adl046	GCGTACGTACGC	14	0.1	0.4	1.0	3.5	A	0.9	0.0	A
bdj017	CCAGGCCTGG	30	0.8	3.9	1.2	11.0	B	1.0	0.1	B
adl045	CCGTACGTACGG	14	0.1	0.3	1.2	2.4	b	1.2	0.0	b
bdj019	CCAACGTTGG	8	0.1	0.2	1.3	11.3	B	1.2	0.0	B
adj022	ACCGGCCGGT	30	0.6	1.5	1.3	4.1	b	0.4	0.4	A
adh041	GTCTAGAC	4	0.1	0.4	1.3	4.1	b	1.1	0.1	b
adh020	CTCTAGAG	6	0.1	0.6	1.6	4.2	b	1.4	0.2	b
addb01	CCGG	11	0.9	2.4	1.8	4.5	b	0.3	1.6	A
bdj051	CATGGCCATG	19	0.3	0.3	1.9	13.5	a	1.5	0.1	B

[a] The base sequences are reckoned in the 5′-3′ direction. The number of local minimums located for each oligomer is listed (no.), along with the standard deviation of slide for this set of minimums (σ_{D_y}) and the standard deviation of energy (σ_E). *R* is the root-mean-square deviation for calculated slide (Å). ΔE^{exp} is the difference in energy per base step, in kJ mol^{-1}, between the experimental and global minimum conformation; ΔE^{min} is the difference in energy per base step between the global minimum and the best local minimum. The first letter of the NDB code indicates the polymorph found experimentally. The calculated structures are designated A- or B-form based on the mean slide. Where the calculated and experimental polymorph differ, a lowercase a or b is used in the table entry.

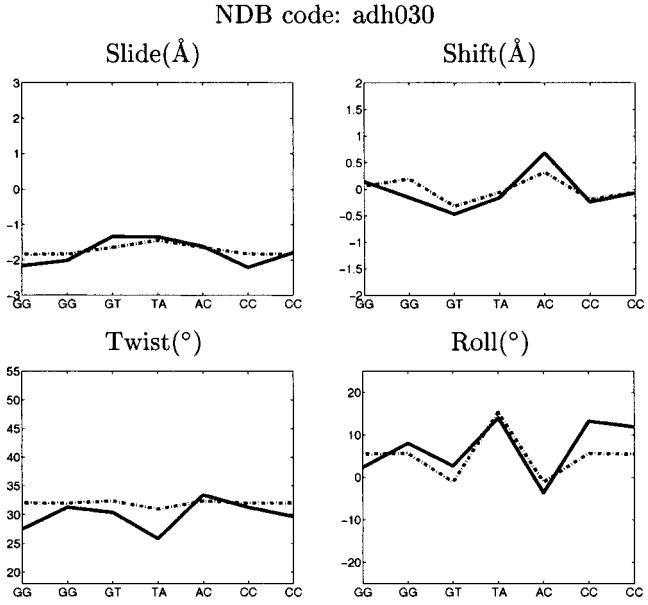

Figure 2. Experimental (solid line) and calculated values of slide, shift, twist, and roll for oligomer adh030.

the step parameters for the experimental structure, the global minimum structure, and all of the local minimums for this sequence. It is clear that adj022 is a very flexible structure and can adopt a range of low-energy conformations. Thus, the slide RMSD can be improved from 1.27 to 0.35 Å at a cost of only 0.38 kJ mol^{-1} per step. Table 2 lists the best local minimum′ for each sequence, i.e., the structure that is most similar to experiment, along with the energy required to reach this conformation starting from the global minimum, ΔE^{min}. The global minimum does not usually correspond to the lowest RMSD, but in most cases, it occupies the same conformational basin of attraction as the best local minimum and differs only slightly in slide, shift and energy.

If we consider the best local minimum, the sequences fall clearly into two groups. Twenty-four of the 30 sequences have an RMSD less than 0.9 Å; i.e., the structure we predict is basically the correct one. For two of these sequences, the global minimum is a B-DNA structure, but the experiment and best local minimum are A-form. These require relatively large energies to reach the experimental structures, but for the other 22 sequences, the global minimum is very close to the best local minimum. The structures of the remaining six sequences cannot be explained in the context of our model. For the three B-structures in this category (bdj017, bdj019, bdj051), the source of the error becomes obvious on examining the slide profiles (Figure 5). In all three cases, there are two CA/TG steps that adopt unusually high positive slide. The large change in slide between neighboring steps is not compatible with our backbone model, which constrains neighboring slides to similar values.

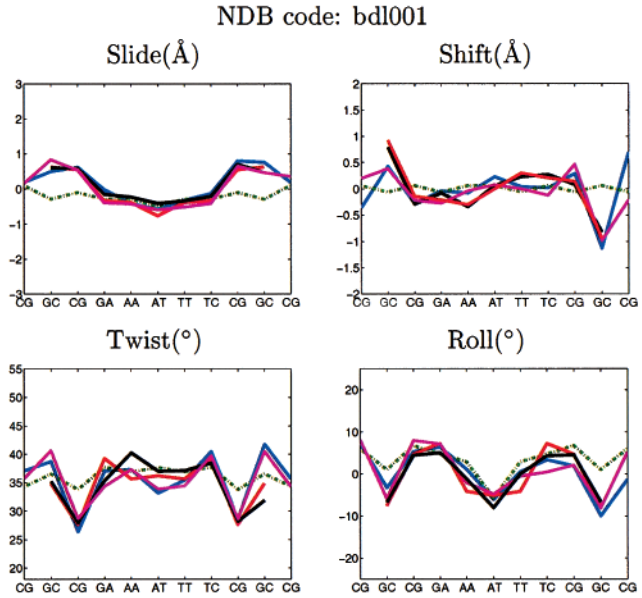

Figure 3. Experimental (solid lines) and calculated (dashed line) values of slide, shift, twist, and roll for bdl001 (blue). Three other sets of X-ray data for this sequence, which were not included in our database, are also plotted: bd0005 (magenta),[24] bd0014 (black),[25] and bd0032 (red).[26] Only bd0032 has crystallized in a symmetric conformation consistent with the sequence, although the end bases are unpaired meaning that there are no terminal step parameters.

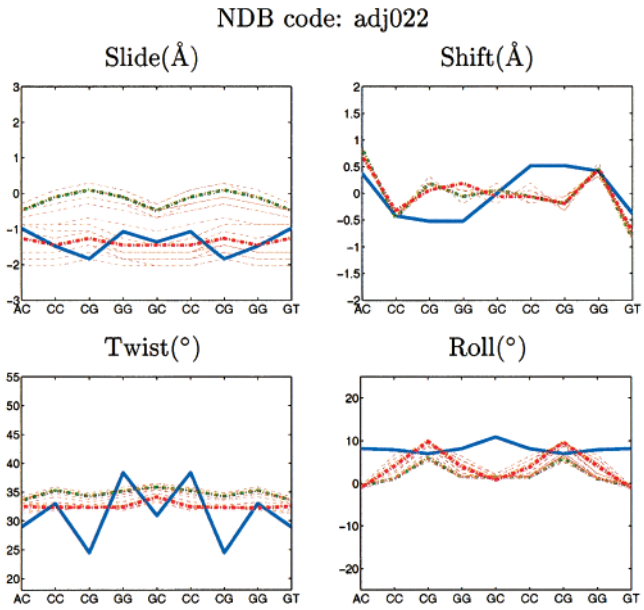

Figure 4. Experimental (solid line) and calculated values of slide, shift, twist, and roll for adj022. This oligomer shows the largest improvement in RMSD between the global minimum structure (green) and the best local minimum (red). All 30 local minimums located by the grid search are plotted.

On examining the X-ray crystal structures in detail, we found that these steps all have a nonstandard B_{II} conformation in both backbones and that this is associated with close intermolecular phosphate contacts. It seems likely that these crystal structures therefore represent distorted conformations stabilized by strong intermolecular interactions in the crystal. The three A-structures that we are unable to predict do not appear to share any common features that might account for the failure of the model.

In Figures 6 and 7, we show step parameter profiles for one A- and one B-DNA oligomer which were not represented in our test set: adj0113 [29] and bdl078.[30] For bdl078, there are very few local minimums, and the slide RMSD for the global minimum structure is 0.75 Å. There is generally good agreement for the other step parameters, particularly roll. For adj0113, there are more local minimums, but the global minimum coincides with the experimental structure with a slide RMSD of 0.65 Å. Again the trends for the other step parameters are good.

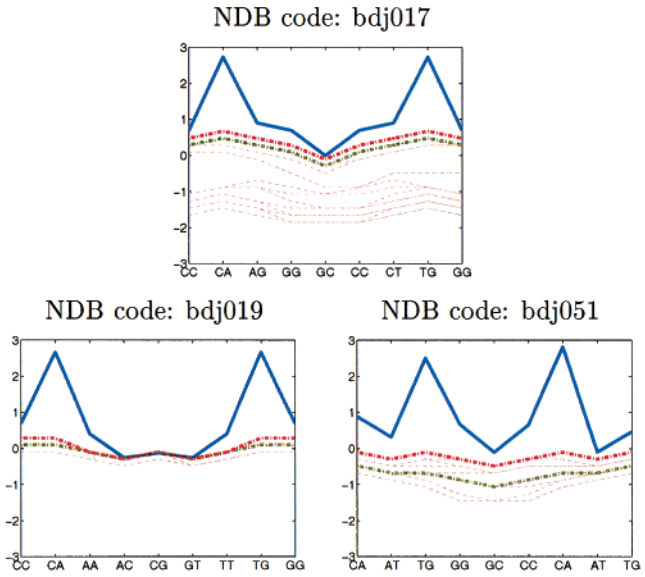

Figure 5. Experimental (solid line) and calculated values of slide for three oligomers which show large RMSD values at both global and local minimums and have large values of ΔE (Table 2). All three have high positive slide at CA/TG, which we predict to be unfavorable conformations due to the large change in slide between neighboring steps.

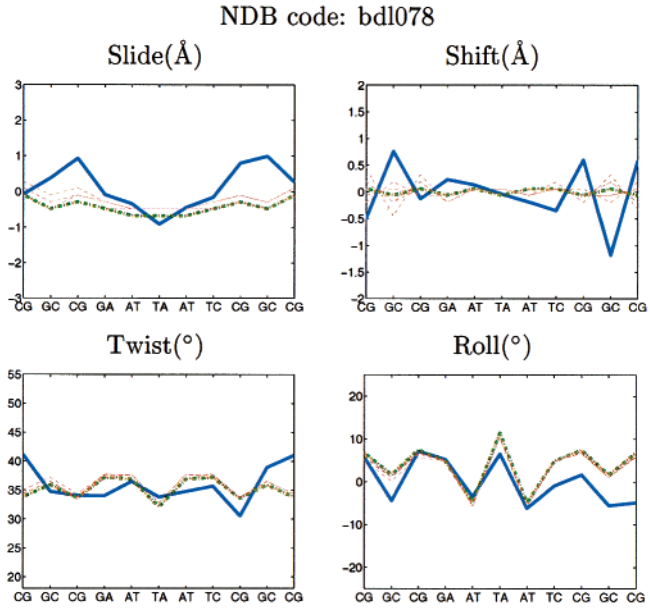

Figure 6. Experimental (blue) and calculated values of slide, shift, twist, and roll for bdl078. The global minimum structure is plotted in green. This oligomer was not in the test set used to parametrize the backbone model.

Scope and Limitations. The quality of our structural predictions for octamer, decamer, and dodecamer structures suggests

(29) Ban, C.; Sundaralingam, M. *Biophys. J.* **1996**, *71*, 1222−1227.
(30) Shatzky-Schwartz, M.; Arbuckle, N. D.; Eisenstein, M.; Rabinovich, D.; Bareketsamish, A.; Haran, T. E.; Luisi, B. F.; Shakked, Z. *J. Mol. Biol.* **1997**, *267*, 595−623.

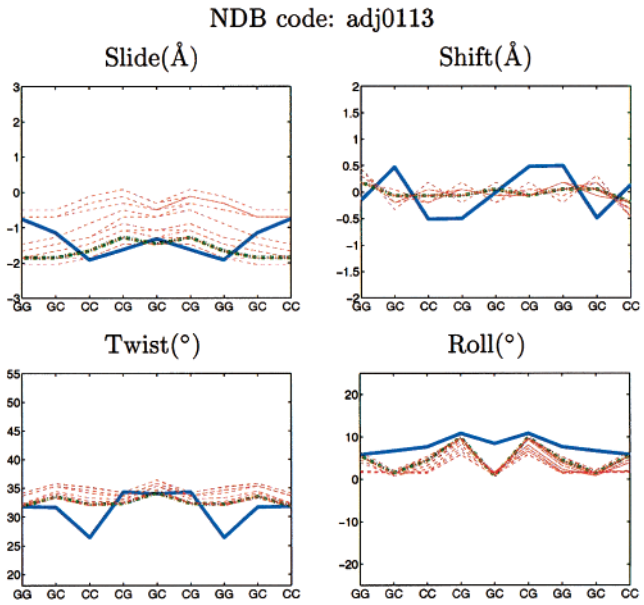

Figure 7. Experimental (blue) and calculated values of slide, shift, twist, and roll for adj0113. The global minimum structure is plotted in green. This oligomer was not in the test set used to parametrize the backbone model.

that we should be able to treat much longer oligomers with similar accuracy. An important aim in modeling extended sequences is to understand how long-range curvature varies with the presence of different sequences.[31] Such curvature is related principally to variations in roll, with superhelical or plane curve regions being induced by phasing of certain motifs. The slide/shift adiabatic maps on which our oligomer structures are based show a strong coupling between roll and slide.[14] Optimizing extended oligomers using slide and shift as the primary degrees of freedom will therefore provide a viable model for sequence-dependent long-range curvature, since roll will vary in a manner consistent with experimental observation.

While we have parametrized against X-ray crystal data at the level of di- and tetranucleotides,[14] optimization of the oligomer structures did not include any factors to account for crystal contacts or any effects specific to the crystal environment. The agreement between our global minimum conformations and experiment is therefore persuasive evidence that our predictions will be valid for solution-phase structures. Base stacking effects have been calculated using a force field, without reference to the crystal structure database, and so modeling modified bases should give results of similar accuracy. In contrast, the backbone is modeled via an empirical function which encapsulates conformational constraints and solvent and environmental effects in the crystal. This means that the model cannot readily be extended to modified backbones or to significantly distorted backbone conformations (e.g., in intercalated base steps). It will also be unable to describe structures that depend on unusual solvent environments (e.g., Z-DNA).

Conclusion

We have developed a relatively simple model for predicting the structural properties of DNA oligomers. The model has three components. The conformation of individual steps is determined by the base stacking interactions in conjunction with the constraints imposed by a backbone that behaves as a semiflexible rod. Within this model, the properties of a base step can be completely described by two parameters, slide and shift.[13] Second, to model context effects in oligomers, we impose correlation of slide and anticorrelation of shift, based on our observation of these trends in experimental data. Third, we include a penalty term, E_{sugar} (eq 4), to account for the base−backbone interactions associated with the χ torsion. Our results indicate that A- and B-DNA are representative of a single conformational family, spanning a wide range of slide.[32] Benchmarking the methodology against a database of 30 different DNA oligomers for which high-resolution X-ray crystal structures are available has provided good evidence for the validity of the approach. The model predicts the occurrence of two structural families, A- or B-DNA, and is able to place different sequences into the correct family with reasonable reliability. Twenty-four of the structures were predicted accurately, and for three of the remaining six structures, the conformation appears to be significantly perturbed by phosphate contacts in the crystal, and so these examples are not representative of the structures of the individual molecules. When applied to two sequences that were not used in the development of the model, excellent results were obtained with RMSDs in the values of slide of 0.65 and 0.75 Å.

Both molecular dynamics modeling[33] and high-resolution X-ray studies[34] have previously found that there are multiple minimums in DNA oligomers. We have identified families of local minimum structures in our calculations, and sequence-dependent variations in flexibility are clearly apparent. While only GG/CC steps are bistable at the dinucleotide level, other steps become bistable when placed in different sequence contexts.[4] This bistability opens up the possibility of multiple minimums in longer oligomers. The X-ray crystal structure observed in an experiment reflects a low-energy conformation which is also compatible with the environment and organization of the crystal itself. This causes subtle distortions of the DNA, but in most cases, the experimental conformation is very close in structure and energy to our calculated global minimum structure.

There is no restriction on the length of oligomer that can be studied with this type of model, and the simplicity of the energy function opens the way to studying oligomers that are hundreds of base pairs in length, giving information on nonlocal interactions and long-range curvature.[11] The fact that we use a grid of energy values for each step, rather than a single preferred conformation[6,35] not only means that we can explore sequence context variations but also means that we can study the effect of constraints, such as protein-induced bending or wrapping onto nucleosome core particles. A potential drawback is the lack of resolution, with the accuracy of step parameters in the region of 5° for angles and 0.5−1.0 Å for translations, as indicated by Table 1. However, we clearly obtain correct trends for the step parameters and could use the predicted structures as a first approximation for more detailed modeling. There is no doubt that the model captures sequence dependence at the dinucleotide level,[13] sequence context effects in tetranucleotides,[4] and the influence of longer sequences on step conformation and context. Combined with efficient minimization and a higher resolution model for the backbone, it promises to provide further insight

(31) Calladine, C. R.; Drew, H. R. *Understanding DNA−The molecule and how it works*, 2nd ed.; Academic Press: London, 1997.

(32) Ng, H. L.; Kopka, M. L.; Dickerson, R. E. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2035−2039.

(33) Mackerell, A. D.; Banavali, N. K. *J. Comput. Chem.* **2000**, *21*, 105−120.

(34) Kielkopf, C. L.; Ding, S.; Kuhn, P.; Rees, D. C. *J. Mol. Biol.* **2000**, *296*, 787−801.

(35) Goodsell, D. S.; Dickerson, R. E. *Nucleic Acids Res.* **1994**, *22*, 5497−5503.

into the role of sequence-dependent variations in the structure and function of extended DNA sequences.

Acknowledgment. We thank Peter Willett for his original suggestion that a genetic algorithm approach would be useful for this problem and Mark Dauncey for contributing to its development. We also thank Stephen Neidle for providing us with coordinates for BD0032 prior to their publication.

JA003385U