Optimization of Nucleic Acid Sequences

Ingrid Lafontaine and Richard Lavery
Laboratoire de Biochimie Théorique UPR 9080 Centre National de la Recherche Scientifique, Institut de Biologie Physico-Chimique, Paris 75005, France

ABSTRACT Base sequence influences the structure, mechanics, dynamics, and interactions of nucleic acids. However, studying all possible sequences for a given fragment leads to a number of base combinations that increases exponentially with length. We present here a novel methodology based on a multi-copy approach enabling us to determine which base sequence favors a given structural change or interaction via a single energy minimization. This methodology, termed ADAPT, has been implemented starting from the JUMNA molecular mechanics program by adding special nucleotides, "lexides," containing all four bases, whose contribution to the energy of the system is weighted by continuously variable coefficients. We illustrate the application of this approach in the case of double-stranded DNA by determining the optimal sequences satisfying structural (B–Z transition), mechanical (intrinsic curvature), and interaction (ligand-binding) properties.

INTRODUCTION

Most properties of nucleic acids can only be understood in detail by taking into account base sequence effects. This is true for the fine structure of double-helical DNA or for the folded forms of RNA, and it is also true for the often highly specific interactions formed between nucleic acids and proteins or drug molecules. Although many biophysical techniques give access to sequence information, computational modeling is becoming a steadily more reliable tool for obtaining such data, and it has the advantage of being well-adapted to systematic studies of mechanical and dynamic properties (Auffinger and Westhof, 1998; Sprous et al., 1998).

Studying sequence effects by modeling, however, leads to a major problem involving the number of calculations to be performed. If we consider the example of DNA-protein binding, the nucleic acid site contacted typically involves 10 to 15 basepairs. For 10 basepairs, there are already 4^{10}, that is to say, more than a million possible sequences. For 15 basepairs, this number becomes 1.07×10^9. It should be remarked that specific gene control effectively requires sites of this length to ensure that they will be statistically unlikely to occur more than once within a complete genome. Although it is clearly impossible to carry out calculations on this number of sequences, such data are necessary to understand how a protein chooses its optimal binding site or, perhaps still more importantly, how a protein mutation would affect this choice. Experimentally, problems of this type have been treated very successfully with the SELEX approach (Tuerk and Gold, 1990), which iteratively selects molecules satisfying a chosen criterion from a combinatorial bank of starting sequences. We now propose a theoretical method to attack similar problems.

Our method involves extending the computer modeling of nucleic acids to incorporate the notion of an evolving base sequence. More explicitly, this implies using a sequence that can itself be energy-optimized in the same way that we normally optimize molecular conformations. Our idea is based on a mean-field approach and is closely related to a number of so-called multi-copy algorithms. These methods have been very successful in solving computationally complex problems ranging from the diffusion of a small ligand within a globular protein (Elber and Karplus, 1990) to the optimization of the structure of polypeptide side chains or loops (Koehl and Delarue, 1996).

In the case of nucleic acids, the small number of natural bases (G, A, C, and T or U) led us to the idea that a multi-copy approach could be used to generate "adaptive" sequences within nucleic acids. We have put this idea into practice by modifying the JUMNA algorithm (Lavery et al., 1995) and by creating a companion program, ADAPT, which can carry out sequence optimizations. The resulting algorithm has been applied to three practical problems of the type mentioned above with encouraging results and only modest computational effort.

METHODOLOGY

JUMNA

Our methodology is based on the JUMNA program, which uses a reduced coordinate approach to model nucleic acids (Lavery et al., 1995). JUMNA breaks each nucleic acid strand into its constituent nucleotides by introducing junctions at the O5′—C5′ bonds. These bonds are maintained during minimization by quadratic restraints. Nucleotides are positioned with respect to an axis system using helicoidal translation and rotation variables. The internal flexibility of each nucleotide is limited to torsions around single bonds and to certain valence angles, lying along the phosphodiester backbone and within the sugar rings (sugar flexibility again involves the introduction of restraints for the C4′—O4′ bonds). All bond lengths are assumed to be constant. These choices lead to roughly a 10-fold reduction in the total number of variables compared to a Cartesian coor-

Received for publication 31 January 2000 and in final form 11 April 2000.
Address reprint requests to Richard Lavery, Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, Paris 75005, France. Tel.: 33-1-58-41-50-16; Fax: 33-1-58-41-50-26; E-mail: rlavery@ibpc.fr.

0006-3495/00/08/680/06 $2.00

dinate representation. In addition, the direct use of helicoidal variables enables easy introduction of helical (or superhelical) symmetry, with further gains in the total number of variables. Conformational energies are calculated with the FLEX force field specifically developed for the nucleic acids (Lavery et al., 1995), but calculations can also be made using the versions parm94 (Cornell et al., 1995) and parm98 (Cheatham and Kollman, 1999) of the AMBER force field. Solvent and counterion electrostatic damping effects are introduced via a sigmoidal distance-dependent dielectric function (using a slope of 0.356 and a plateau value of 80) (Lavery et al., 1995; Hingerty et al., 1985) and reduced net charges on the phosphate groups (−0.5). A recent study has shown that, with an appropriate choice of damping parameters, this method can produce stable A and B forms of DNA (Flatters et al., 1997) and it has also enabled DNA deformations to be modeled in good agreement with experiment (Sanghani et al., 1996; Cluzel et al., 1996; Lebrun et al., 1997).

Lexides

Within JUMNA, each base is a flat, rigid body (with the exception of the rotating thymine methyl group). Nucleic acid starting conformations are built by plugging together pre-constructed nucleotides drawn from a library file. In order to introduce "adaptive" sequences into JUMNA, we started by adding a new type of nucleotide to the library. This entity is termed a "lexide" by analogy to the name lexitropsin, introduced for sequence-reading ligands (Kopka et al., 1985b). Lexides are distinguished by having the four standard bases (T, C, A, and G for DNA, or U, C, A, and G for RNA) bound to a single sugar C1′ atom. These bases are superposed on one another and share a common C1′-N1/N9 glycosidic bond vector (Fig. 1).

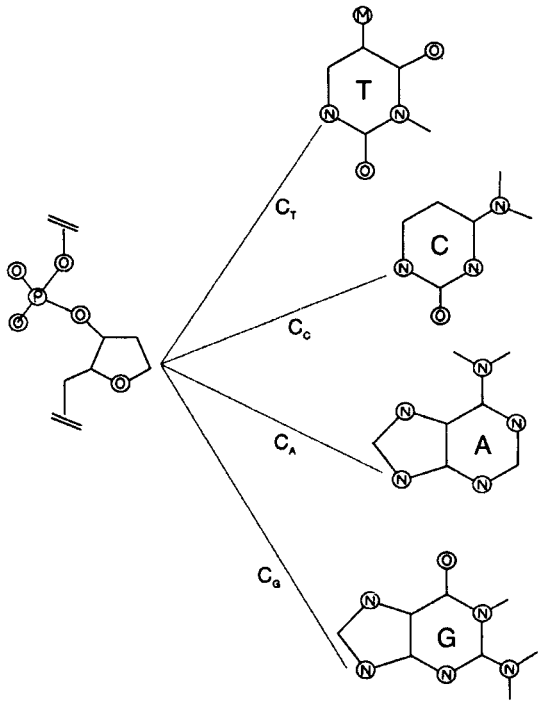

FIGURE 1 Schematic representation of a "lexide." The four standard bases (shown here in an exploded view) are bound to a common C1′ atom and superposed in space. The bases of a given lexide do not interact with one another. The contribution of each lexide base to the energy of the system is weighted by a coefficient (C_T, C_C, C_A, or C_G), whose sum is normalized (see Table 1).

TABLE 1 Examples of various choices of the lexide coefficients

Lexide	C_T	C_C	C_A	C_G	σ_L
T	1	0	0	0	1
R (purine)	0	0	0.5	0.5	0.577
Y (pyrimidine)	0.5	0.5	0	0	0.577
N	0.25	0.25	0.25	0.25	0
X	← variable →				

The normalized standard deviation of the coefficients, σ_L, is equal to 1 for a pure base and decreases to zero for a homogeneous mixture of all four bases (N).

When these lexides are built into a nucleic acid, several changes are made to the energy calculations. First, the four bases belonging to a given lexide do not interact with one another. Second, all energy contributions involving base b of the lexide l are multiplied by a coefficient C_{lb}, which indicates the degree to which this base is present within the structure. If the energy term involves base atoms from two lexides, then two coefficients will be involved. The meaning of the coefficients C_{lb} is given in Table 1. Note $\Sigma_{b=1,4} C_{lb} = 1$ for a given lexide l. By modifying these coefficients, a lexide can effectively become a standard nucleotide or can represent a partial mixture of bases, such as a purine or a pyrimidine. It can also become an averaged nucleotide "N" and be used to construct a "neutral" sequence. The state of a lexide can be usefully summarized by the value of the normalized standard deviation σ_l of its coefficients C_{lb}:

$$\sigma_l = \left[(4/3) \sum_{b=1,4} (C_{lb} - 0.25)^2 \right]^{1/2}$$

As shown in Table 1, the extreme values are $\sigma_l = 1$ for a "pure" base and $\sigma_l = 0$ for a homogeneous mixture of bases.

Sequence optimization

Taking one more step, it is possible to imagine minimizing the energy of a structure containing lexides with respect to their coefficients C_{lb}. In this case, the sequence will evolve. It is remarked that, during minimization, σ_l can be used to prevent, or to force, the appearance of a "pure" base sequence.

Minimizing the energy of a given nucleic acid structure with respect to its sequence in this way is, however, physically incorrect. Even though the internal energy of each base is ignored in JUMNA (this energy being constant for rigid bodies), minimizing with respect to sequence, via the lexide coefficients, would be equivalent to comparing the energies of molecules with different chemical compositions. This is invalid with molecular mechanics force fields that calculate conformational (and not formation) energies.

In fact, problems that interest us will always involve at least two conformational states of a given nucleic acid. For example, one can ask which sequence most favors the transition of a fragment from conformation A to a new conformation B. Our approach can be used to find this sequence by minimizing the energy difference $E_B - E_A$ with respect to the lexide base coefficients. Note that the coefficients of equivalent lexides in the two conformational states are identical at all times. The energy difference being minimized thus has a sense since both molecules maintain identical chemical compositions.

In practice, such a study involves generating initial conformations for the two fragments using "neutral" sequences composed of "N" lexides (where all base coefficients equal 0.25). We then minimize the energy difference between the fragments with respect to their sequence (that is, their lexide coefficients). If necessary, it is possible to re-minimize the conformation of the two fragments to take into account the conformational

impact of their new sequence and then to loop back to the sequence minimization step.

It is actually possible to make the sequence optimization step very rapid. We first store the energy calculated by JUMNA. This is done using a matrix where each element contains the energy terms multiplied by a given lexide coefficient, or pair of coefficients (plus a single element containing the energy terms not involving lexides). One such matrix is constructed for the conformation created at the end of each structural minimization procedure. This matrix contains all the information necessary to calculate the energy of the corresponding conformation for any chosen base sequence. A new program, ADAPT, can then be used to obtain the optimal energy difference between the conformations with respect to the lexide coefficients. At each sequence minimization step, this program can construct the necessary energies by simply multiplying the appropriate matrix terms by the appropriate values of the lexide coefficients.

It should be remarked that although each lexide is characterized by four coefficients, the normalization condition $\Sigma_{b=1,4} C_{lb} = 1$ must be satisfied. This can be done in a number of ways. We presently define the coefficients as $C_{lb} = V_{lb}^2/\Sigma_{b=1,4} V_{lb}^2$, where V_{lb} are the variables actually used by the minimization algorithm. We have compared this approach with the use of three curvilinear variables that automatically respect the normalization constraint for each lexide, and identical results were obtained.

Treating basepairs

A special feature has been added to JUMNA and ADAPT to simplify the treatment of canonical basepairs. In such cases, the coefficients of the two lexides forming the pair are coupled so that $C_{lA} = C_{l'T}$, $C_{lG} = C_{l'C}$, $C_{lT} = C_{l'A}$, and $C_{lC} = C_{l'G}$ (where l and l' are the complementary lexides of a basepair). Energy interactions within such a pair are also limited to complementary bases. This choice requires the addition of a normalization factor $[\Sigma_{b=1,4} C_{lb}^2]^{-1/2}$ because of missing cross-terms. This is easily illustrated for the case of a basepair made from an equal mixture of AT and GC. The left-hand lexide will have coefficients 0/0/0.5/0.5 and the right-hand lexide 0.5/0.5/0/0 (the base coefficients being given in the order TCAG). Since A only interacts with T and G with C, the total energy will consist of two terms, each multiplied by $0.5 \times 0.5 = 0.25$, and therefore its total weight would be 0.5 and not 1.0, as it would be with pure bases, for example, G (0/0/0/1) interacting with C (0/1/0/0).

Combinatorial sequence searches

Provided we are only interested in "pure" base sequences (where each lexide has one coefficient equal to unity) it is possible to replace minimization with a combinatorial search of all possible base sequences for the fragments under study. Although the number of these sequences increases exponentially with the number of lexides present, an efficient use of the energy matrices described above enables very rapid energy calculations.

RESULTS

To illustrate our new approach we present three types of problem that can be studied: a simple comparison of DNA allomorphic forms, a more subtle case of DNA sequence-dependent deformation, and an example of a DNA-ligand interaction.

DNA allomorphic forms

The B-DNA to Z-DNA transition is a good example for studying sequence effects on the allomorphic form of DNA because the Z conformation requires RY (purine-pyrimidine) alternating sequences and has a strong preference for GC alternation, with *syn* guanosine residues linked to C4′-exo sugars followed by *anti* cytidine residues linked to C2′-endo sugars. According to free energy measurements (Ho et al., 1986), the order of preference for forming Z-DNA is GC > AC > AT = GG > GA (where we adopt a notation corresponding to *syn*-p-*anti* for each nucleotide pair). To test whether our technique can reproduce this order, we first energy-minimized the B and Z forms of DNA for double-stranded 18-mers with "neutral" sequences ($dN_{18} \cdot dN_{18}$). Note that all calculations were carried out under dinucleotide symmetry constraints. End-effects were avoided by optimizing the energy of a single dinucleotide repeating unit in its polymeric environment (Lavery et al., 1995). Each strand of the Z form began with a *syn* nucleotide.

The energy matrices resulting from these calculations were used as a basis for optimizing the sequence to stabilize the Z form with respect to the B form. Because of dinucleotide symmetry and the imposition of Watson-Crick pairing (see Methodology, Treating Basepairs) this optimization involves only two independently variable lexides. After roughly 250 cycles of minimization, we obtained zero gradients (for the B → Z energy difference with respect to the lexide coefficients) with an energy of 2.7 kcal/mol and lexides representing a "pure" GC alternating sequence: "pure" meaning that each lexide has one coefficient equal to unity and thus has become a normal nucleotide. It was not immediately clear that such a minimization would lead to a pure sequence, but as we shall see below, this has invariably been the result of our calculations to date.

Assuming that a pure sequence is indeed the best result for this study, we can use our combinatorial approach to test the transition energy for the 10 unique dinucleotide sequences. The results are shown in Table 2, where it can be seen that GC is indeed the most favorable sequence and thus

TABLE 2 B→Z transition energies (kcal/mol) as a function of sequence

ADAPT		ADAPT + structure optimization	
Sequence	$\Delta E_{B\to Z}$	Sequence	$\Delta E_{B\to Z}$
GC	2.7	GC	2.0
AC/GT	3.1	AC/GT	2.8
GA/TC	4.8	AT	3.7
AT	4.9	GG/CC	3.7
GG/CC	5.2	CA/TG	3.7
TA	5.3	GA/TC	4.2
AA/TT	5.7	TA	4.4
CA/TG	5.9	CG	4.5
AG/CT	6.4	AG/CT	5.7
CG	7.1	AA/TT	8.1

The first set of results refers to the initial structures of B-DNA and Z-DNA optimized with neutral $dN_{18} \cdot dN_{18}$ sequences, while the second set has allowed for structural adjustment to the corresponding sequences.

that minimization with ADAPT effectively found the optimal result. It can also be seen below that the best five sequences found are those favored experimentally (although it should be recalled that we are only estimating enthalpies of transition and not free energies). If we make one more step and allow the B and Z conformations to adjust to the base sequence present, only minor reordering occurs (see Table 1 and below).

ADAPT: $GC > AC > GA \approx AT > GG$

ADAPT + structure optimization:

$$GC > AC > AT = GG = CA$$

Experiment (Ho et al., 1986):

$$GC > AC > AT = GG > GA$$

DNA curvature

Predicting the sequence-dependence of DNA curvature presents a more subtle problem. First, the conformational changes necessary to induce DNA curvature are small, and second, we need to consider much longer variable sequences in order to produce significant DNA curvature. JUMNA offers the possibility of building polymeric curved DNA conformations through the introduction of superhelical symmetry constraints (Sanghani et al., 1996). This implies using a number of basepairs within the symmetry repeat unit that is equal to an integer number of turns of the double helix. In the present case we use 10 basepairs per symmetry unit. Under superhelical symmetry, the DNA helical axis itself becomes a helix with a defined radius R and pitch P. All present calculations are performed with zero pitch, implying that the corresponding DNAs will be planar circles. For convenience, the superhelical radius R is expressed as a curvature index, $C = 45/R$. This implies that $C = 0$ for straight DNA ($R = \infty$) and $C = 1$ when the radius of DNA is that of a nucleosome particle ($R = 45$ Å). We have built DNAs with radii varying from 900 Å ($C = 0.05$) to 45 Å ($C = 1$).

Conformational optimizations were performed for each chosen curvature using ($dN_{18} \cdot dN_{18}$) "neutral" sequence oligomers. Sequence optimizations were then carried out for the passage from a straight DNA to each of the increasingly curved conformations.

The results, which are presented in Table 3, show that optimizing the lexides again leads to pure base sequences. As concerns the nature of the optimal sequences, even small degrees of curvature show a preference for A_nT_n tracts separated by CG basepairs. More precisely, $(A_4T_4CG)_n$ is favored for curvatures varying from 0.2 to 0.8, while $(A_3T_4CGC)_n$ and $(A_3T_5CG)_n$ are favored for the most strongly curved conformations. This is an encouraging result because such sequences are known experimentally to be associated with intrinsic DNA curvature (Hagerman, 1986). It is also worth noting that these sequences have been found despite the fact that the optimal energy difference between the straight and curved DNAs amounts to less than a kcal/mol per turn of the double helix (and, incidentally, that our approach excludes any effects due to explicit water molecules or counter ions). Combinatorial sequence searches confirm the results of the minimization for curvatures in the range 0.2–0.6. Outside this range there are some minor changes, but in each case the energy difference involved is very small, typically only hundredths of a kcal/mol.

The fact that the optimal sequences found by our approach indeed lead to curvature can be seen in Fig. 2, where we plot the energy as a function of curvature. While the "neutral" sequence resists bending, the optimal sequence found by our approach for intermediate curvatures, $(A_4T_4CG)_n$ effectively favors a bent conformation. It is also important to note that, in line with experimental data (Drew and Travers, 1985), the minor groove of the A_nT_n tract faces in the direction of the induced curvature.

TABLE 3 Optimized sequences obtained by comparing straight (S) with curved DNA (U) for different radii of curvature (R)

C	R (Å)	Minimization	$\Delta E_{S\rightarrow U}$	Combinatorial	$\Delta E_{S\rightarrow U}$
0.05	900	CGAAGCCTTT	−0.102	GAAAACTTTC	−0.103
0.1	450	GAAAACTTTC	−0.200	GAAAAATTTC	−0.202
0.2	225	AAAATTTTCG	−0.396	AAAATTTTCG	−0.396
0.3	150	AAAATTTTCG	−0.520	AAAATTTTCG	−0.520
0.4	112	AAAATTTTCG	−0.593	AAAATTTTCG	−0.593
0.5	90	AAAATTTTCG	−0.632	AAAATTTTCG	−0.632
0.6	75	AAAATTTTCG	−0.698	AAAATTTTCG	−0.698
0.7	64	AAAATTTTCG	−0.917	GAAATTTCGC	−0.978
0.8	56	AAAATTTTCG	−0.923	GAAATTTCGC	−0.981
0.9	50	AAATTTTCCG	−0.833	AATTTTCGCG	−0.845
1.0	45	AAATTTTTCG	−0.626	GAATTTTCGC	−0.637

The curvature index C is equal to $45/R$. Energies are given in kcal/mol and refer to the conformational energy of the 10-basepair repeating unit. The result of a single minimization of the lexide coefficients is compared with an exhaustive combinatorial search of all possible sequences respecting the symmetry constraints.

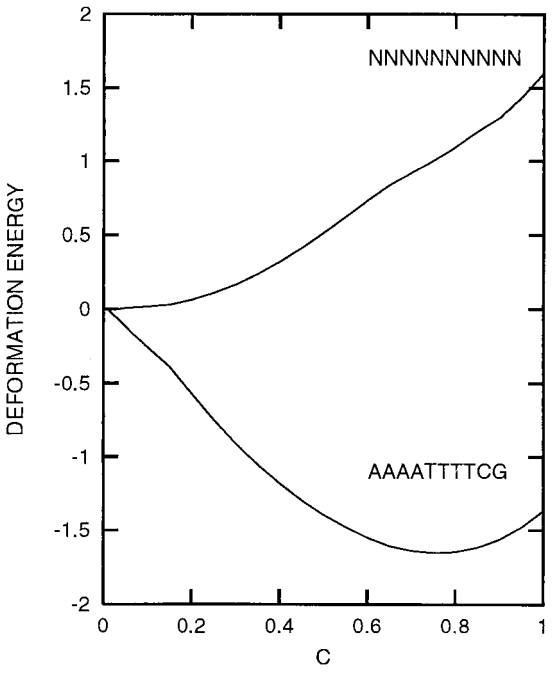

FIGURE 2 Deformation energy per turn of DNA (kcal/mol) as a function of its curvature ($C = 45/R$, where R is the radius of curvature) using superhelical symmetry constraints and a repeating unit of 10 basepairs. A "neutral" base sequence $(dNNNNNNNNNN)_n$ is compared with the sequence optimized by ADAPT at $C = 0.2–0.6$ $(dAAAATTTTCG)_n$. Note that, in contrast to the "neutral" sequence, the optimized sequence shows a stable curvature.

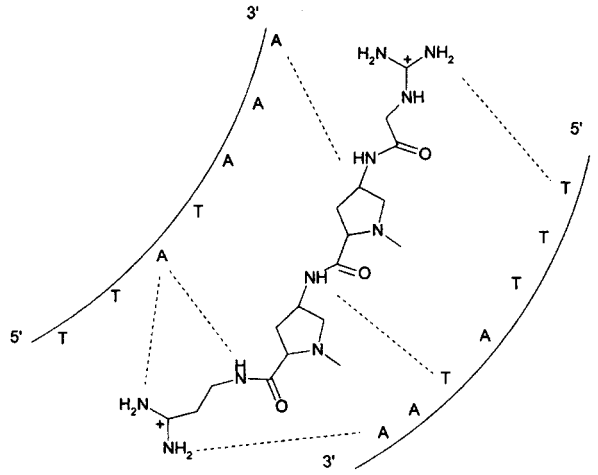

FIGURE 3 Schematic diagram of netropsin bound to the optimized binding sequence obtained with ADAPT. Netropsin is positioned in the minor groove and is hydrogen-bonded (*dotted lines*) to a six-basepair site, TTTATA. Hydrogen bonds involve both A(N3) and T(O2) atoms and, in the case of the proprioamidinium group at the bottom of the figure, two bonds to sugar O4′ atoms. The electrostatic influence of netropsin results in AT basepairs being created beyond the binding site for 9 of the 12 "adaptable" base positions (see Table 4).

DNA-ligand binding

Finally, we consider the case of a DNA-ligand interaction. We have chosen netropsin as a well-known example of a simple, sequence-specific ligand. Netropsin is a cationic (net charge +2), peptide-like antibiotic and antitumoral agent that binds in the minor groove of DNA (Kopka et al., 1985a) most strongly to AT basepairs and with a preference for alternating sequences (Marky and Breslauer, 1987). To generate the necessary matrices for sequence optimization, we first energy-minimized a canonical B-DNA 18-mer with a fixed $CGCN_{12}CGC$ sequence. We then energy-minimized a complex between this oligomer and netropsin (held in its crystallographic conformation) (Kopka et al., 1985a), placing the ligand at the entrance to the central part of the minor groove, several angströms away from the DNA bases. An optimal sequence was then generated in two steps: 1) preventing a full sequence from appearing by limiting all σ_L to 0.25, and 2) re-optimizing the structures of the isolated DNA and of the complex and allowing an explicit sequence to appear. The result obtained was CGCGGT**TTTATA**-AACCGC, where the bold characters indicate the ligand binding site. As observed experimentally, this site is found to consist of a partially alternating AT tract. The resulting complex was finally energy-optimized, allowing the ligand to adapt its internal conformation. Fig. 3 shows the ligand orientation and the hydrogen bonds formed. Netropsin is well situated within the minor groove, interacting with six basepairs in a manner similar to that seen in the crystallographic complex (Kopka et al., 1985a).

It is also interesting to note that the electrostatic properties of netropsin reach beyond its physical binding site and favor AT pairs (which are associated with more negative

TABLE 4 Netropsin binding to DNA

Site	Sequence	ΔE_{DNA}	ΔE_{LIG}	ΔE_{INT}	ΔE_{COMP}
Minor	cgcGGT**TTTATA**AACcgc	8	4	−91	−79
Major	cgcAGC**TTGTTCA**CCgc	6	1	−53	−46
Entry minor	cgcGCCCCATATGGGcgc	3	2	−28	−23

The optimized binding sequence and the complexation energy (ΔE_{COMP}) of netropsin are shown for three sites of the ligand: within the minor groove, within the major groove, and held at the entrance to the minor groove. Bold letters indicate the ligand binding site. Binding is made to a B-DNA oligomer with 18 basepairs, of which the 12 central pairs (upper case) are variable lexides. The complexation energy is decomposed into three components describing the deformation energies of DNA (ΔE_{DNA}) and of netropsin (ΔE_{LIG}) upon binding and the interaction energy between the two partners (ΔE_{INT}). All energies are in kcal/mol.

minor groove potentials) for 9 of the 12 basepairs in the sequence adaptable region. The results in Table 4 confirm this electrostatic effect by showing that when netropsin is held at the entrance to the minor groove (preventing any direct interaction with the basepairs), sequence optimization already generates an ATAT tract, although no direct DNA-ligand interactions exist. Table 4 also confirms that netropsin preferentially binds to the minor groove. When the ligand is placed in the major groove and an optimal sequence is generated, the complexation energy is only 60% of that in the minor groove. This decrease corresponds to a poorer steric fit and to fewer hydrogen bonds, which involve only the proprioamidinium and guanidinium endgroups.

CONCLUSIONS

We have presented an algorithm capable of dealing with nucleic acid sequence problems that would otherwise pose great computational difficulties. For two of the examples treated here, conventional modeling techniques would require several million energy minimizations to reliably obtain the optimal base sequences: for DNA curvature, it would be necessary to calculate the energies for straight and curved segments with all possible 10-basepair sequences, that is, 2.1 million (2×4^{10}) possibilities; for netropsin binding, it would be necessary to compare the energies of free and bound DNA with all possible sequences for the 12-basepair tract we studied, that is, 3.3 million (2×4^{12}) possibilities. By introducing the notion of "adaptive" sequences, only one or two minimizations become sufficient. Despite the simplicity of the current modeling approach, the results obtained are found to be in line with available experimental data.

As with earlier mean-field approaches, the use of multiple base copies appears to smooth out the conformational energy surface and facilitate minimization (Elber and Karplus, 1990; Koehl and Delarue, 1996). For the examples studied here, combinatorial searches have confirmed that our energy minimizations lead directly to the global sequence minimum or, in rare cases, to a sequence differing by a few bases (and having an energy within a few hundredths of a kcal/mol compared to the global minimum).

It should be stressed that this methodology could be used in other nucleic acid modeling contexts. The basic ideas are by no means restricted to the simple force field and solvent model that we have used here. It should nevertheless be remarked that our simple treatment associated with either the FLEX or AMBER force fields lead to encouragingly good agreement with experiment for the problems we have studied. This is particularly striking in the case of sequence-dependent curvature where very small deformation energies are involved and where we do not consider any effects due to explicit water or ions. Finally, it is also possible to use this approach with more complex energy criteria involving many conformational states. In this way it should become possible to define sequences that exhibit given dynamic properties.

The authors thank the Indo-French Centre for the Promotion of Advanced Research for their support of this work through Project 1804-1.

REFERENCES

Auffinger, P., and E. Westhof. 1998. Simulations of the molecular dynamics of nucleic acids. *Curr. Opin. Struct. Biol.* 8:227–236.

Cheatham, T. E. III, and P. A. Kollman. 1999. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* 4:845–862.

Cluzel, P., A. Lebrun, C. Heller, R. Lavery, J.-L. Viovy, D. Chatenay, and F. Caron. 1996. DNA: an extensible molecule. *Science.* 271:792–794.

Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. 1995. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* 117:5179–5197.

Drew, H. R., and A. A. Travers. 1985. DNA bending and its relation to nucleosome positioning. *J. Mol. Biol.* 186:773–790.

Elber, R., and M. Karplus. 1990. Enhanced sampling in molecular dynamics: use of the time-dependent Hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *J. Am. Chem. Soc.* 112:9161–9175.

Flatters, D., K. Zakrzewska, and R. Lavery. 1997. Internal coordinate modeling of DNA: force field comparisons. *J. Comp. Chem.* 18: 1043–1055.

Hagerman, P. J. 1986. Sequence-directed curvature of DNA. *Nature.* 321:449–450.

Hingerty, B., R. H. Ritchie, T. L. Ferrel, and J. E. Turner. 1985. Dielectric effects in biopolymers: the theory of ionic saturation revisited. *Biopolymers.* 24:427–439.

Ho, P. S., M. J. Ellison, G. J. Quigley, and A. Rich. 1986. A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J.* 5:2737–2744.

Koehl, P., and M. Delarue. 1996. Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.* 6:222–226.

Kopka, M. L., C. Yoon, D. Goodsell, P. Pjura, and R. E. Dickerson. 1985a. Binding of an antitumor drug to DNA, Netropsin and C-G-C-G-A-A-T-T-BrC-G-C-G. *J. Mol. Biol.* 183:553–563.

Kopka, M. L., C. Yoon, D. Goodsell, P. Pjura, and R. E. Dickerson. 1985b. The molecular origin of DNA-drug specificity in netropsin and distamycin. *Proc. Natl. Acad. Sci. USA.* 82:1376–1380.

Lavery, R., K. Zakrzewska, and H. Sklenar. 1995. JUMNA: junction minimisation of nucleic acids. *Comp. Phys. Commun.* 91:135–158.

Lebrun, A., Z. Shakked, and R. Lavery. 1997. Local DNA stretching mimics the distortion caused by the TATA box-binding protein. *Proc. Natl. Acad. Sci. USA.* 94:2993–2998.

Marky, L. A., and K. J. Breslauer. 1987. Origins of netropsin binding affinity and specificity: correlations of thermodynamic and structural data. *Proc. Natl. Acad. Sci. USA.* 84:4359–4363.

Sanghani, S. R., K. Zakrzewska, S. C. Harvey, and R. Lavery. 1996. Molecular modelling of $(A_4T_4NN)_n$ and $(T_4A_4NN)_n$: sequence elements responsible for curvature. *Nucleic Acids Res.* 24:1632–1637.

Sprous, D., M. A. Young, and D. L. Beveridge. 1998. Molecular dynamics studies of axis bending in $d(G_5\text{-}(GA_4T_4C)_2\text{-}C_5)$ and $d(G_5\text{-}(GT_4A_4C)_2\text{-}C_5)$: effects of sequence polarity on DNA curvature. *J. Phys. Chem. B.* 102:4658–4667.

Tuerk, C., and L. Gold. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science.* 249:505–510.