Ingrid Lafontaine
Richard Lavery
Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, Paris 75005, France

Received 4 August 2001; accepted 9 August 2001

ADAPT: A Molecular Mechanics Approach for Studying the Structural Properties of Long DNA Sequences

Abstract: *We describe an original approach to determining sequence–structure relationships for DNA. This approach, termed ADAPT, combines all-atom molecular mechanics with a multicopy algorithm to build nucleotides that contain all four standard bases in variable proportions. These nucleotides enable us to search very rapidly for base sequences that energetically favor chosen types of DNA deformation or chosen DNA–protein or DNA–ligand interactions. Sequences satisfying the chosen criteria can be found by energy minimization, combinatorial sequence searching, or genome scanning, in a manner similar to the threading approaches developed for protein structure prediction. In the latter case, we are able to analyze roughly 2000 base pairs per second. Applications of the method to DNA allomorphic transitions, DNA deformation, and specific DNA interactions are presented.* © 2001 John Wiley & Sons, Inc. Biopoly (Nucleic Acid Sci) 56: 292–310, 2001

Keywords: *molecular modeling; multicopy algorithm; DNA; DNA–protein binding; sequence–structure relationships; genome analysis*

INTRODUCTION

The primary function of DNA is to provide the genetic code for protein synthesis within the cell. However, DNA contains many other signals that regulate gene expression, recombination, replication, and even its own spatial organization. These signals are also contained in the base sequence, but they are expressed through a more subtle code that involves local DNA structure and flexibility. Understanding sequence effects on DNA structure is therefore a necessary step to deciphering this code.

Over the last few decades, many examples of sequence–structure–function relationships have been identified for DNA. The best known is certainly intrinsic curvature that can be induced by a variety of base sequences, most notably A-tracts, and leads to both reduced electrophoretic mobility on polyacrylamide gels and enhanced ring closure probabilities.[1–3] A-tracts appear to induce curvature via special structural features including narrow minor grooves and high propeller twists,[4–6] although the exact mechanism remains unclear. It has been shown that intrinsic curvature plays a determinant role in many essential biological process including transcription,[7,8] recombination,[9] and DNA compaction within the cell.[10,11]

A number of proteins specifically recognize curved DNA, or favor sequences that can easily be

Correspondence to: Richard Lavery; email: rlavery@ibpc.fr
Biopolymers (Nucleic Acid Sciences), Vol. 56, 292–310 (2001)

curved.[12–14] Such properties thus play an important role in the indirect recognition of base sequence.[15,16] A good example is provided by the TATA-box binding protein (TBP) that specifically recognizes a region within eukaryotic RNA polymerase II (RNA pol II) promoters termed the TATA-box.[17,18] As part of the transcription factor complex TFIID,[19,20] TBP binds to DNA enabling RNA pol II to be correctly positioned with respect to the transcription start site (TSS).[21] It has to be noted that TBP also exists in archae.[22–24] Crystallographic studies have shown that TBP binding induces strong DNA curvature, local unwinding and minor groove opening.[25–33] The intrinsic bending and flexibility of the base sequences that constitute the TATA box have been shown to be important for TBP recognition both experimentally[30,34–39] and theoretically.[40–44] Although such structural properties are probably common to all TBP binding sites, many of these sites do not conform to the experimentally established consensus sequence (TATAWAWN, where W implies A or T and N any base[45]),[34,46–48] while other sites that fit the consensus are not actually bound by TBP. This makes it difficult for lexical approaches to use TATA boxes as signals for promoter regions and thus as pointers to adjacent coding regions.[49,50] It also suggests that adding structural properties to such sequences analyses could be a useful step forward.[51]

The structural properties of DNA are also known to be important for its organization within the cell. In eukaryotic cells, DNA compaction involves histone binding to form chromatin.[52–55] The basal unit of chromatin is the nucleosome, which results from the interaction of 146 base pairs of DNA with a histone octamer.[56] The crystal structure of the nucleosome shows that the main protein–DNA contacts involve the phosphodiester backbones of DNA[57–59] and it is consequently not surprising that nucleosome reconstitution in vitro is possible for nearly all type of sequences.[60] Nucleosome positioning is nevertheless influenced by the structural properties of certain sequences,[10,11,61–65] and regions showing preferential nucleosome binding are known to play an important role in the regulation of genes expression.[66–71] Although methods have been proposed to predict nucleosome positions from sequence data both experimentally[63,72] and theoretically,[64] it is clear that the underlying mechanisms are not yet well understood.[11]

In view of these multiple roles for sequence-dependent structural and mechanical properties, it would certainly be useful to predict such properties for given sequences. Unfortunately, this goal is hindered by simple combinatorial problems. Since the number of sequences to be studied grows exponentially with fragment length (already exceeding a million for ten base pairs), we rapidly reach the limits of standard experimental or theoretical approaches. One experimental solution to this problem is the SELEX method,[73] which involves an iterative selection of nucleic acid oligomers satisfying a chosen criteria from a combinatorial bank of starting sequences. However, while SELEX can create DNA sequences with given properties, it cannot be used to analyze existing sequences or to formulate a predictive theory linking sequence to structure.

Because of the limited amount of experimental data available, most theoretical approaches for studying the structural and mechanical properties of DNA sequences are still based on di- or trinucleotide models. As an example, Pedersen and co-workers have used both these approaches to produce structural profiles for a group of human promoter sequences that had previously been aligned using hidden Markov models.[74] The resulting "bendability" profiles showed that these sequences tend to be highly bent downstream of the TSS and weakly bent upstream.[75] The generation of so-called structural atlases for eighteen complete bacterial genomes indicated that coding sequences, intergenic regions, and promoter regions each present different structural properties, apparently confirming the importance of sequence–structural–function relationships for DNA.[76] In another approach, Kono and co-workers scanned long DNA sequences with an empirical potential derived from the analysis of 52 protein–DNA complexes whose structure has been resolved by crystallography and showed that they could detect the binding sites for a set of regulatory proteins.[77] Other authors have replaced structural data with related information on the sequence-dependent stability of the double helix.[78–80] The results show that coding regions can be discriminated by their high thermal stability and that regions of destabilization are correlated with promoters.

Despite the useful information that can be obtained by such approaches, it is necessary to recall that structural analysis of crystallographic and NMR data shows that local heterogeneity cannot be fully predicted by di- or even trinucleotide models and many examples of cooperative behavior over relatively long base sequences are known.

From a theoretical standpoint, the most flexible approach would therefore be one that could analyze any chosen structural or mechanical property of existing DNA sequences, or could, alternatively, generate sequences that would best express a chosen property, without having to resort to a deconvolution of existing experimental results. We have recently attempted to formulate such an approach. Our technique is termed ADAPT. It is based on mean-field theory

and uses a multicopy algorithm in combination with an internal coordinate representation of DNA to permit systematic studies of the relationships between sequence and structural properties.[81,82] By effectively making the base sequence itself a variable of the problem, this approach overcomes the combinatorial barrier to studying large numbers of sequences, without obliging us to give up an all-atom representation of DNA. After presenting the theoretical approach we have adopted, we will show how ADAPT can be used to explain the sequence dependence of both the intrinsic and interactive properties of DNA. We also show that the method can be applied to analyzing properties on the scale of whole genomes.

METHODS

DNA Representation

We have chosen to carry out our simulations using internal coordinate molecular mechanics. This choice enables us to greatly improve the performance of energy minimization by reducing the number of variables needed to model a DNA fragment by a factor ranging from roughly ten to one hundred.[83] We worked with the molecular mechanics algorithm JUMNA (*JU*nction *M*inimization of *N*ucleic *A*cids),[84] which offers convenient tools for studying DNA helical structures and their deformations. In this program, a DNA fragment is built by associating 3′-monophosphate nucleotides. Each nucleotide is positioned using a set of six helicoidal parameters, three translations, and three rotations, with respect to a referential axis system. Junctions are introduced at the O5′—C5′ bonds and maintained by quadratic distance restraints during minimization. Similar junctions are introduced into the sugar rings at the C4′—O4′ bonds. These choices allow easy construction of both canonical and irregular structures. It is also possible to introduce helical or superhelical symmetry. In the latter case, the possibility to use a helical axis that itself follows a helical pathway in space[85] makes it possible to build and energy minimize polymeric DNA conformations with a defined pitch P and radius of curvature R. This option, which implies using a number of base pairs within the symmetry repeat unit equal to an integer number of turns of the double helix, makes it simple to study the properties of curved DNA in a systematic way.

The internal flexibility of each nucleotide within JUMNA is limited to two torsion angles (ε, C3′—O3′ and ζ, O3′—P) and two valence angles (C3′—O3′—P and O3′—P—O5′) along the backbone, two torsions (O4′—C1′—C2′—C3′ and C1′—C2′—C3′—C4′) and three valence angles (O4′—C1′—C2′, C1′—C2′—C3′ and C2′—C3′—C4′) within the sugar, and to the glycosidic angle (see Figure 1). JUMNA calculates the conformational energies using either the FLEX force field[86,87] or the AMBER force field.[88,89] The present studies were all performed using the FLEX parameters. To mimic the damping effects occurring between two charges in a polar solvent, a sigmoidal distance dependent dielectric function is used,[84,90] while the net charges on phosphate groups are reduced from −1 e to −0.5 e to mimic counterion effects. A conjugate gradient method is used to minimize the energy. A recent study has shown that, with an appropriate choice of electrostatic damping parameters, this method can produce stable A and B forms of DNA,[40] and it has also enabled a number of DNA deformations to be modeled in good agreement with experiment.[85,91–93]

Variable Base Sequences

We have used a new application of mean-field theory in order to make the base sequence a variable of our method. Algorithms based on mean-field theory are very powerful in solving computationally complex problems and they are of special interest for studying biological processes.[94] In the case of conformational searches, a particularly useful approach involves using many copies of part of the system in order to enhance conformational sampling. The multiple copies do not interact with one another and the rest of the system senses the mean field generated by the copies. This approach has already been used to successfully study the specificity of ligand–protein interactions, using many copies of the ligand to search for binding pathways to a macromolecular target site[95,96] and extensions to multiple copies of the protein are possible.[97] It can also be used for positioning amino acid side chains, or complete amino acid loops, within a globular proteins.[98–103]

We have applied mean-field theory to model the influence of the four bases of DNA (adenine, thymine, guanine, and cytosine) within a single multicopy nucleotide (see Figure 1). These nucleotides, which we have termed "Lexides" (in reference to "Lexitropsins," which were conceived as "sequence-reading" ligands),[104] are built from the standard library used within JUMNA. The pyrimidine N1 (C and T) and the purine N9 (A and G) atoms are linked to the sugar C1′ atom. The bases are coplanar and, as for standard nucleotides, have no flexibility, excepting the rotation of the thymine C5 methyl group. The base orientation with respect to the sugar moiety varies with the glycosidic angle (O4′—C1′—N9—C4 for purines and O4′—C1′—N1—C2 for pyrimidines). Within a given lexide i, each base k is associated with a coefficient of presence $C(i, k)$. The sum of all $C(i, k)$ being set equal to one. Generally, within an oligomer, each base k of lexide i feels the mean field formed by the four bases of all other lexides j, but does not feel the presence of the other three bases of i.

By setting all $C(i, k)$ equal to 0.25, each of the four bases composing the lexide contribute equally to the conformational energy calculations and it becomes possible to study the conformational properties of a DNA with an averaged base sequence. We term such lexides N, implying "sequence neutral" nucleotides. An oligomer containing only "neutral" lexides is denoted ($dN_n \cdot dN_n$). Conventional "pure" nucleotides may be obtained by setting any single

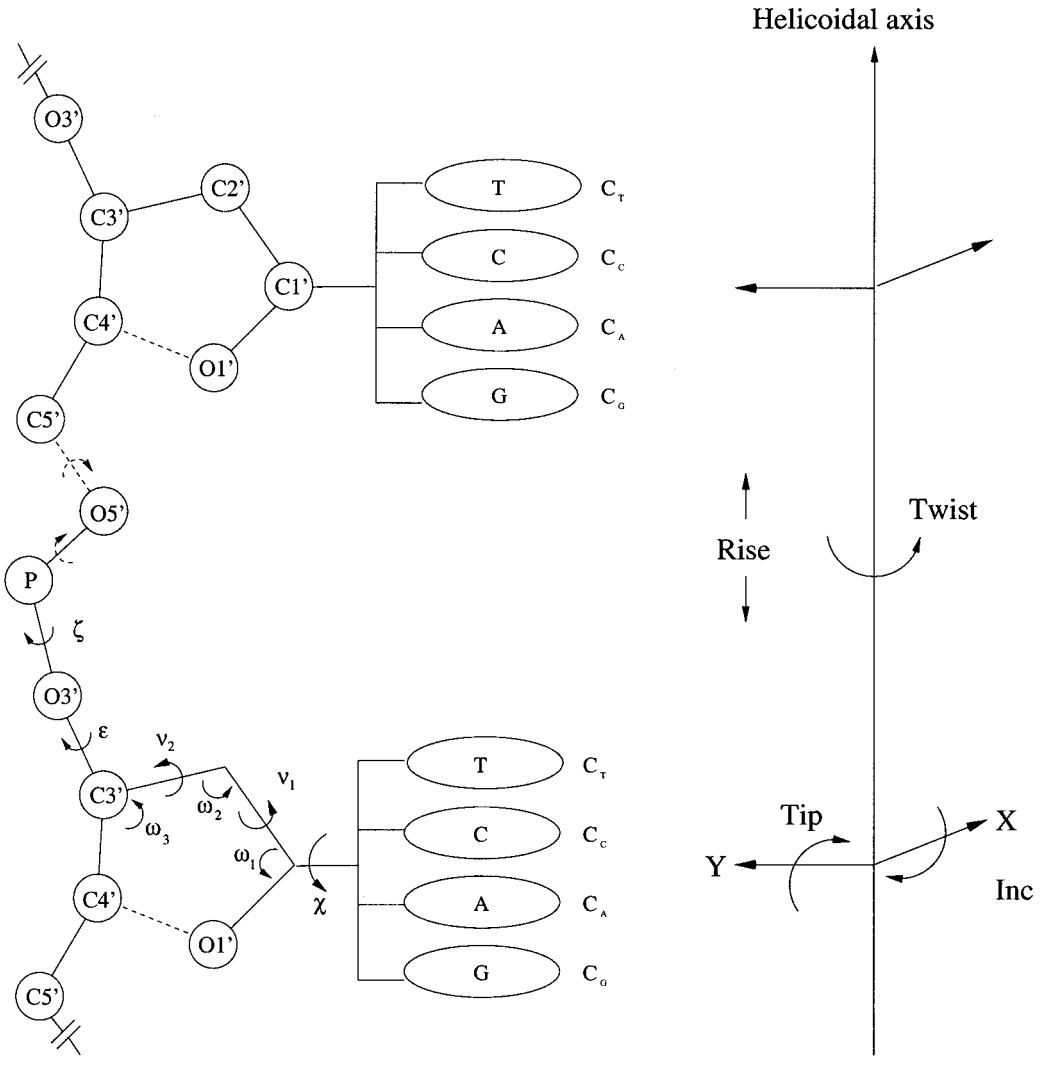

FIGURE 1 Schematic diagram of a lexide within the internal and helicoidal representation of DNA used within JUMNA. Each lexide, shown here in an exploded view, is constituted of a coplanar arrangement of the four standard bases bound to a common C1′ atom. The contribution of each lexide base to the conformational energy of DNA is regulated by the coefficients of presence: C_G, C_A, C_C and C_T.

coefficient to unity, e.g., dG implies using the base coefficients: $C_T = C_C = C_A = 0$ and $C_G = 1$, while a purine can be created by setting $C_T = C_C = 0$, and $C_A = C_G = 0.5$. It is also remarked that it is possible to save time in treating canonical Watson–Crick base pairs by using only a single set of lexide coefficients for both bases. Each coefficient then controls the presence of a given base in one strand and of its paired partner in the other strand.

By simply varying the lexide coefficients it becomes possible to study all possible base sequences within a given DNA fragment. This can represent enormous time savings and can also simplify conformational searches, since, as other studies have shown,[105,106] the use of multicopy algorithms tends to lower the energy barriers in the effective energy landscape.

Energy Calculations in the Presence of Lexides

The presence of lexides within a DNA fragment implies that any pairwise energy contribution involving either one or two atoms from the bases of the lexides will be multiplied by the corresponding coefficients $C(i, k)$ of these bases. In order to rapidly calculate the energy of a given molecular conformation for any given set of lexide coefficients (or, more specifically, any given base sequence), it is convenient to store the pairwise energy contributions in a matrix of dimension $n \times n$ for n lexides (or lexide pairs), grouping together all those terms multiplied by a given coefficient (along the diagonal, ii, of the matrix) or by a given pair of coefficients (in the upper triangle, ij, of the matrix). This

matrix has a depth of 16 elements containing the grouped energy terms corresponding to all possible combinations of the base indices, k, belonging to lexide i, and l, belonging to lexide j (although only four of these elements will be used if $i = j$). All the terms involving atoms from the phosphodiester backbones or normal (nonvariable) bases are independent of the coefficients and are stored in an additional matrix element, OO.

Once this matrix is constituted for a given molecular conformation the energy of any base sequence involving the lexides within this conformation can be calculated very rapidly by simply summing the terms of the matrix multiplied by the appropriate choice of lexide coefficients. If we begin by optimizing the molecular conformation using a neutral ($dN_n \cdot dN_n$) sequence, then to a first approximation we can assume that this conformation can be considered valid for any given base sequence. However, once we have located the sequence that satisfies our energy criteria (see below), nothing prevents us from reoptimizing the conformation for this sequence to improve the quality of the model. In fact, in many cases that we have studied so far, staying with the conformation obtained with the neutral sequence turns out to be a very reasonable approximation.

Optimization and Sequence Scanning

Although it is possible to minimize the energy of a structure containing lexides with respect to their coefficients, that is to say, changing the base sequence, it is impossible to compare two molecules of different chemical nature. This is due to the fact that molecular mechanics force fields calculate conformational, and not formation, energies. This is not a problem since our goal, the study of the influence of the sequence upon the structural properties of DNA, can be reached by comparing two distinct conformations of the same molecule, which we will term the target and reference conformations. If we keep the lexide coefficients identical within these two conformations, the energy difference between them is a valid quantity since we are comparing two chemically identical molecules (Figure 2). If we then allow the lexide coefficients to evolve, we can search for the set of coefficients that best stabilizes the target conformation with respect to the reference conformation. In energy matrix terms, this means that we will need to calculate an energy matrix for both the target and reference conformations. These matrices will then be subtracted from one another, before being multiplied by the chosen lexide coefficients. We can thus imagine searching for a base sequence that will stabilize, for example, a curved DNA with respect to a straight one. We can also imagine asking which sequence will most stabilize a protein–DNA complex with respect to a free DNA. In this case, the bound molecule is not associated with any lexide coefficients, and as far as the energy matrix is concerned, can be simply treated in the same way as the phosphodiester backbones of DNA.

How can we go about finding the optimal coefficients to solve such problems? ADAPT in fact offers three ways of doing this. Historically, the first approach we developed was

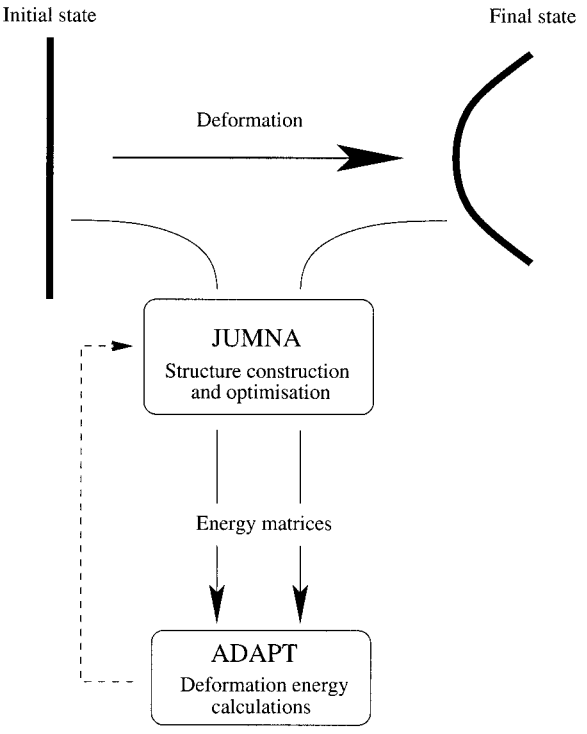

FIGURE 2 Calculation steps involved in sequence optimization. Reference and target conformations are optimized with respect to their conformational energy using JUMNA. They generally have an averaged base pair sequence, $(dN)_n \cdot (dN)_n$. The energy matrices resulting from these calculations are used within ADAPT to find the sequence that optimally reduces the deformation energy separating the reference and target conformations. If necessary, one can loop back to JUMNA for further conformational optimization.

energy minimization. The variables of the minimization are simply the lexide coefficients. However, it is necessary to respect the normalization constraint imposed on each set of four coefficients. This constraint can be integrated into the problem analytically by redefining a set of four independent variables, V_k, such that,

$$V_K = [C(i, k)/C(i, \max)]^{1/2}$$

where $C(i, \max)$ is the maximum of the four coefficients. Similarly, coefficients respecting the normalization criteria can be recovered as

$$C(i, k) = V_k^2 / \sum_{l=1}^{4} V_l^2$$

This problem can also be solved by generating three curvilinear variables that respect the normalization constraint. These variables can be thought of as defining the position on the surface of a four-dimensional hypersphere with unit

radius. We have found that both these methods lead to identical results.

Carrying out energy minimization implies calculating the energy derivatives with respect to the lexide coefficients. This can be done very easily since all the energy terms are either linearly dependent on the coefficients or are completely independent. The only complication comes from the fact that the minimizer requires the $C(i, k)$ derivatives to be converted into V_k derivatives. It should be noted that whichever the technique used to define the minimizer variables, the energy derivatives become identically zero for pure sequences (where one coefficient $C(i, k)$ is unity and the others are zero). In order to determine whether the stationary points that we locate by energy minimization are true minima and not saddle points, we have also calculated the Hessian matrix of the second derivatives of the energy with respect to the coefficients. Diagonalizing this matrix leads to a set of eigenvalues that will all be positive in the case of a true energy minimum.

It should be noted that there is no theoretical requirement for energy minimization to lead to pure base sequences. For this reason, we included the possibility of forcing a pure sequence to appear. This involves using the normalized standard deviation σ_L of the lexide coefficients $C(i, k)$, which equals unity for a pure base and zero for a neutral base:

$$\sigma_L = [(4/3) \sum_{k=1}^{4} (C(i, k)^2 - 0.25)]^{1/2}$$

To obtain a pure sequence, we force σ_L to become unity using a simple quadratic restraint. In fact, for reasons that are still not clear, almost all the problems we have studied using ADAPT lead to perfectly, or almost perfectly, pure sequences (within 1–2%). Consequently, σ_L has rather been used to prevent pure sequences from appearing too quickly (e.g., in the case of ligand binding) than for forcing pure sequences to appear.

If we are only interested in "pure" base sequences as solutions to our problem, then energy minimization can be replaced by a simple search of all possible base combinations. For sequences of length n, this implies calculating energies from the energy matrices for 4^n combinations of lexide coefficients. This can be done in a matter of minutes for $n < 12$. For longer sequences, an analysis of the energy matrices shows that, for both standard and deformed double helices, a given lexide only interacts significantly with only two neighboring base pairs on each side. This implies that combinatorial solutions for long sequences can be accelerated by building up the overall result from a set of connected pentanucleotide solutions.[82]

Finally, we can study the energy difference between our target and reference sequences for all possible sites along a given genomic sequence.[82] This "sequence scanning" involves calculating the deformation energy inside a window (having the same length as the reference and target oligomers), which is moved base by base along the sequence to be analyzed. This approach, which is analogous to the "threading" methods used for knowledge-based protein structural predictions,[107] can be used to characterize specific structural properties of an entire genome directly from its sequence. The present version of the program can scan roughly 2000 base pairs per second. The most exciting application of genome scanning is clearly with target conformations that involve protein binding. In such cases, we can hope that the regions presenting the lowest deformation energy will be likely protein binding sites.

Building Reference and Target Conformations

As mentioned above, the reference conformation, which in all the cases considered here will be a standard B-DNA state, is built using a neutral sequence, ($dN_n \cdot dN_n$), and then energy minimized. With the FLEX force field this leads to a conformation that corresponds well with the canonical parameters of B-DNA deduced from crystallographic and NMR studies of DNA oligomers. In most of the cases studied, the DNA fragments used contain only N lexides; however, a lexide segment can also be combined with standard nucleotides—for example, by placing a lexide segment between fixed sequence segments. Target conformations are also constructed using neutral sequences, but energy minimization is carried out in the presence of constraints that impose the desired change in structure. The restraints available within JUMNA are of many types, acting upon axial curvature, interatomic distances, helicoidal parameters, sugar puckers, etc. In the case of protein binding, we use the CONTACT utility program[108] to create a set of interatomic distances between the DNA atoms belonging to the protein–DNA interface that are able to reproduce the conformational impact of the protein. These restraints may be supplemented by including part or all of the protein in the energy matrix calculations carried out for the target conformation. We created the PCHEM utility program for this purpose. To allow for imprecision in the experimental data and also for the limited flexibility of the JUMNA internal coordinate model, protein–DNA interface atoms can be given the freedom to move within spheres of a chosen radius. In the calculations presented here a radius of 0.4 Å was used.

RESULTS AND DISCUSSION

In order to illustrate the possibilities of our new approach, we will present six different applications that treat successively an allomorphic transition of the double helix, a uniform structural deformation, ligand binding, single protein binding, and then genome scanning applied to both the TATA-box binding protein and a model of the nucleosome.

B–Z Allomorphic Transition

Experimental studies have revealed that the purine–pyrimidine (RY) alternating sequences strongly favor the B to Z transition.[109] The most stable Z-DNA structure is obtained with an alternating GC sequence, *syn*-guanosine-p-*anti*-cytosine, with C4′-*exo* and C2′-*endo* sugars for G and C respectively. The relative ease of forming Z-DNA in terms of dinucleotide steps is as follows: GC > AC > AT = GG > GA (where we adopt a notation corresponding to *syn*-p-*anti* for each nucleotide pair).[110–112] We used the combinatorial sequence searching procedure of ADAPT to check whether we could reproduce this order. Calculations were carried out on ($dN_{18} \cdot dN_{18}$) oligomers, using a B-DNA reference conformation and a Z-DNA target conformation and imposing dinucleotide symmetry constraints in both cases. Using the energy matrices obtained from these calculations, ADAPT determined the deformation energy required for the B–Z transition as a function of base sequence. The combinatorial procedure led to results for the ten possible dinucleotide sequences. The sequences found were effectively those favored experimentally, GC > AC > GA ≈ AT > GG, and this order is surprisingly close to the experimental result. Given that we are only estimating transition enthalpies and not free energies, this result also suggests that the sequence dependence of the B–Z transition is predominantly enthalpic. It is also remarked that if we use the energy minimization procedure of ADAPT, we also find a "pure" GC sequence to be the most favorable for creating Z-DNA.

Uniform Deformation—Curved DNA

To investigate the sequence dependence of DNA curvature, we constructed and energy optimized the conformations of ($dN_{18} \cdot dN_{18}$) neutral sequence oligomers with radii of curvature ranging from 900 down to 45 Å (the radius of DNA within a nucleosome core particle). This was done using the superhelical symmetry constraints of JUMNA.[85] A 10 base pair repeat symmetry was imposed on both the curved target conformation and the straight B-DNA reference conformation. Energy minimization was carried out within ADAPT to find the optimal 10 base-pair sequences favoring curvature.

The optimized sequences are once again found to be made of "pure" bases. For target conformations with very small curvature, the optimal sequences are made up of only GC base pairs. However, as the curvature increases, the experimentally observed preference for A-tracts separated by GC pairs appears. A single sequence, $(A_4T_4CG)_n$, is found to be optimal for radii of curvature ranging from 225 to 56 Å. For more extreme curvature, either $(A_3T_4CGC)_n$ or $(A_3T_5CG)_n$ are found to be energetically optimal. In each case, the sequence is placed within the curved oligomer so that the minor groove of the A-tract lies on the inside face. Combinatorial sequence searches confirm the results of the energy minimization for radii of curvature in the range 225–75 Å. Outside this range there are some minor changes, but in each case the energy difference involved is typically of the order of 0.01 kcal/mol.

All the optimal sequences identified resemble those used by Hagerman in his studies of the sequence dependence of curvature and are known experimentally to induce gel retardation.[113] It should be remarked that even if the energy differences between straight and curved DNA are less than a kcal/mol per turn of the double helix, ADAPT is still able to identify correct target sequences. It is also worth noting that, in agreement with a previous study made with the JUMNA algorithm, sequence-dependent intrinsic curvature can apparently be successfully modeled without considering the effects due to explicit water molecules or counter ions.[85]

Ligand Binding—Netropsin

Netropsin is a well-known sequence-specific DNA-binding ligand. It is a cationic (net charge +2), peptide-like antibiotic and antitumoral agent, which binds in the minor groove of DNA[114] most strongly to AT base pairs and with a preference for alternating sequences.[115] Since ligand positioning could be expected to depend on the sequence of the target DNA, we used two rounds of the structure-sequence generation procedure presented in methods section (see Figure 2). We first energy minimized a canonical B-DNA reference conformation with a $CGCN_{12}CGC$ sequence. The target conformation was a complex between an oligomer of the same sequence and netropsin (held in its crystallographic conformation).[114] The ligand was placed at the entrance to the central part of the minor groove of the target oligomer, several ångstroms away from the DNA bases. In the first round of energy minimization with respect to the base coefficients, we restrained the appearance of a "pure" sequence by limiting σ_L, the normalized standard deviation rms (see Methods), to 0.25. In the second round of calculations with JUMNA, the structures of the isolated DNA and of the complex were reoptimized with the sequence obtained by the first ADAPT calculations. The final ADAPT cycle was then carried out allowing a pure sequence to appear. The result

obtained was CGCGGTTTTATAAACCGC, where the underlined characters indicate the ligand binding site.

As observed experimentally, this site consist of a partially alternating AT-tract. The resulting complex was finally energy optimized, allowing the ligand to adapt its internal conformation to the target site. Netropsin is then well situated within the minor groove, interacting with six base pairs in a manner similar to that seen in the crystallographic complex.[114] It is 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 minor groove potentials) for 9 out of the 12 base pairs in the sequence-adaptable region.

Single Protein Binding: TBP

Gene expression is controlled in a very sophisticated way by a large variety of transcription factors that act both inside and outside the promoter region. Identifying coding sequences and control regions is thus directly linked to identifying promoters. Unfortunately, as mentioned in the introduction to this article, transcription factors often have poorly defined consensus sequences. This implies an important role for indirect protein–DNA recognition and the need to understand the structural and mechanical properties of the DNA target sequences. This is notably the case for TBP, whose role in gene transcription has already been discussed. As a first step to solving such problems, we used the energy minimization and combinatorial search options of ADAPT to determine the optimal sequences for TBP binding sites. The calculations were performed with three pairs of target and reference conformations that varied only in terms of length: ($dN_{10} \cdot dN_{10}$), ($dN_{16} \cdot dN_{16}$), ($dN_{24} \cdot dN_{24}$). In each case, the reference conformation is an energy-minimized B-DNA, while the central ten base pairs of the target conformation reproduce the DNA conformation within the human TBP/DNA complex.[29] Target oligomers beyond ten base pairs in length were used to see if the additional energy necessary to form junctions with B-DNA on either side of the protein-binding site influenced the optimal target sequences.

Energy minimization and combinatorial search approaches gave very close results. Each of the three pairs of target and reference conformations led to pure base sequences in roughly 200–300 cycles of conjugate gradient minimization. Identical binding site sequences, TATTTAAA, were obtained with all three fragments, although the deformation energy increased from 69 kcal/mol for 10 base pairs (bp) to 73 kcal/mol for 16 bp and finally 85 kcal/mol for 24 bp. Combinatorial searching found the sequence TATTTTTA as the global optimum binding site, once again for all the three fragments. The global optimum differs from the energy minimum sequence by two TA inversions in positions 6 and 7. The energy gain for the three fragments is of the order of 1.7 kcal/mol. The Hessian matrices calculated for the deformation process reveal that there are in fact only a few low energy minima and that the energy landscape is strongly funneling. The energy-optimized sequences indeed turn out to be true minima, even if the global energy minimum is not reached. Given the nature of the energy surface, the performance of the minimization procedure could certainly be improved by using a minimum-hopping algorithm.

By interchanging the target and reference conformations (i.e., changing the order of subtraction of the energy matrices), it is possible to search for the sequences that least favor TBP binding. This trial led to deformation energies between 122 and 138 kcal/mol with pure binding sequences made principally of GC base pairs: GGGCCCTC for the 10 bp fragments and GGGCCCTT for the 16 bp and 24 bp fragments. This result implies that the range of deformation energies for TBP binding as a function of sequence lies between 40 and 50 kcal/mol for the three different fragments used. It is also possible to obtain this result from the energy matrices themselves, by summing all the energy variations with respect to sequence for the 8 base pairs of the binding site.

By using the combinatorial search on the 10 bp fragment, we were able to define a consensus sequence by grouping all the binding sequences that fell within 5 kcal/mol of the global optimum energy of 67.4 kcal/mol. We obtained the consensus kTATWWWWRn, which is surprisingly close to the human TBP binding consensus of sTATAAAWRn [TRANSFAC data base accession number M00252][116] given that the current model of TBP binding is limited to the deformation induced in DNA (note K = G/T, S = G/C, R = A/G, W = A/T, and N = A/C/G/T). We can analyze the origins of this consensus in more detail by looking at the energy variations for each lexide pair forming the binding site as a function of the DNA sequence. The results shown in Figure 3 have been obtained by taking into account the fact that a given lexide pair only interacts significantly with two nearest neighbor pairs on either side (see Methods). The colored bars in the figure indicate the variations in energy for each possible base at each lexide position and the TBP binding site occupies positions 2–9. The bars at positions 2, 3, and 4 enable us to see the domination of T, A and T respectively for these sites, and it is also clear that adenine dominates at

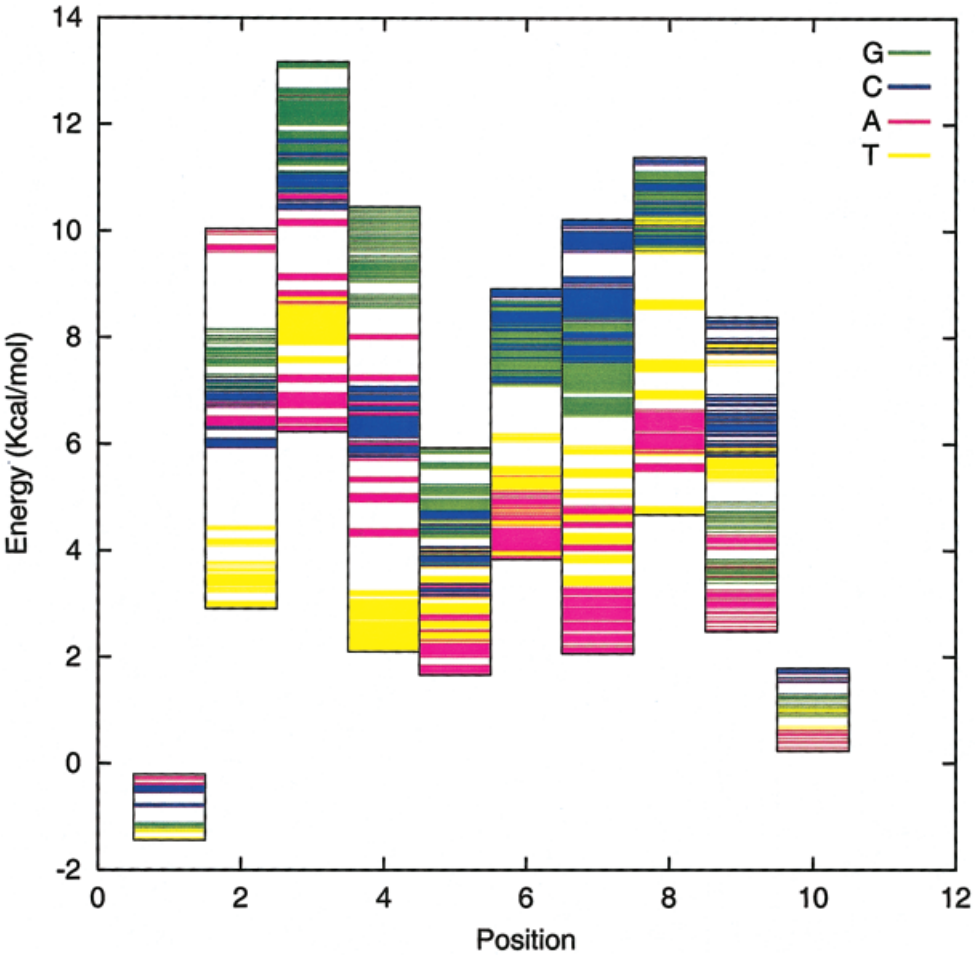

FIGURE 3 Variation of the deformation energy for TBP binding as a function of sequence. Each vertical bar shows the energy range for a given base pair along the binding site (positions 2–9). The colored bars show the energy range for each of the four bases at each site as a function of the sequence of the two nearest neighboring base pairs on either side of the site. The bases with the lowest deformation energies at each site constitute the computationally derived consensus binding sequence. The bars with the largest energy range represent the sites with the highest sequence discrimination for TBP binding.

position 5. If we look at the overall energy range at each site, we can see sequence dependence is only strong for the positions 2–9 belonging to the binding site. It is also worth noting that, within the binding site, positions 5 and 6 have relatively small ranges of deformation energy. We can assume that the energy range (i.e., sequence discrimination) becomes stronger when the specific base–amino acid side chains hydrogen bonds that involve these two positions are present.[26]

Genome Scanning: TBP

Maintaining the same target and reference conformations discussed above, we can use the sequence scanning option of ADAPT to make a first attempt at "reading" the structural properties of genomic sequences. Given the results obtained with the combinatorially defined consensus discussed in the previous section, it is reasonable to suppose that sequences with low deformation energies will be sites favoring TBP binding. We first tested this approach on five human promoter sequences, roughly 2000 bp in length and containing a TATA box.[82] The results showed that the experimentally identified TATA boxes fell within the 20% of the lowest deformation energies for each sequence and one of them was actually the global energy minimum for the sequence studied. These results were encouraging, especially since most of the sites we targeted were rather far from the standard TBP consensus.

Table I Distribution of Global Energy Minima for the Promoter Sequences from (a) Human Genome Sets hs1000 and hs2000, and (b) Vertebrate Genome Sets vs1000 and vs2000[a]

(a) Human Sequences

	hs1000				hs2000			
	Minima	Mean	Mode	σ	Minima	Mean	Mode	σ
DNA_{Alone}	192	−33	−29	267	192	−73	−30	579
DNA_{AA}	192	−35	−29	266	192	−79	−30	551
	Percentage of Global Energy Minima Within the Specified Windows							
	[Mode ± 2]	[−50:1]	[−200:100]		[Mode ± 2]	[−50:1]	[−200:100]	
DNA_{Alone}	11	27	42		6	16	25	
DNA_{AA}	13	30	41		8	20	28	

(b) Vertebrate Sequences

	vs1000				vs2000			
	Minima	Mean	Mode	σ	Minima	Mean	Mode	σ
DNA_{Alone}	620	−76	−30	246	619	−119	−30	541
DNA_{AA}	609	−71	−30	250	626	−96	−30	534
	Percentage of Global Energy Minima Within the Specified Windows							
	[Mode ± 2]	[−50:+1]	[−200:100]		[Mode ± 2]	[−50:1]	[−200:100]	
DNA_{Alone}	12	30	50		7	18	28	
DNA_{AA}	15	32	51		10	22	32	

[a] σ is the standard deviation of the distributions.

In an attempt to systematize this study, we extended our analysis to a much larger number of well-defined promoter sequences. We considered a total of 192 human sequences and the 605 vertebrate sequences that were available within the eukaryotic promoter database (EPD).[117,118] The EPD is a collection of nonredundant RNA pol II promoters for which the TSS has been determined experimentally. In each case, sequences with lengths of 1000 or 2000 bp were considered, with the TSS lying exactly in the midposition (hereafter numbered as position +1). The groups of human sequences, termed hs1000 and hs2000 according to their length, and numbered [−499:+500] and [−999:+1000], respectively. The vertebrate sequences are similarly termed vs1000 and vs2000. If the sequence fragment around the TSS is smaller than the required length, it is completed by a segment of neutral bases. The percentage of such filling segments was 14 and 22% for hs1000 and hs2000 and 5 and 8% for vs1000 and vs2000. It is also assumed that each promoter sequence effectively contains a TBP binding site.

Based upon the study of 502 eukaryotic promoters, Bucher estimated the position of the TATA box as lying within the [−39:−9] region.[45] If the experimentally defined positions of the TSS are correct, if there is really a TBP binding site, and if our structural approach indeed detects TATA boxes, we should find a high concentration of low energy deformation sites within this region. In order to judge the results, we have used two acceptance windows, one roughly 50 bp in width positioned around Bucher's estimate at [−50:+1] and a second 300 bp window centered around the same region, [−200:+100]. This larger window is chosen to take into account more completely experimental uncertainties in the location of the TSS, following earlier comparative study of promoter detection algorithms.[49]

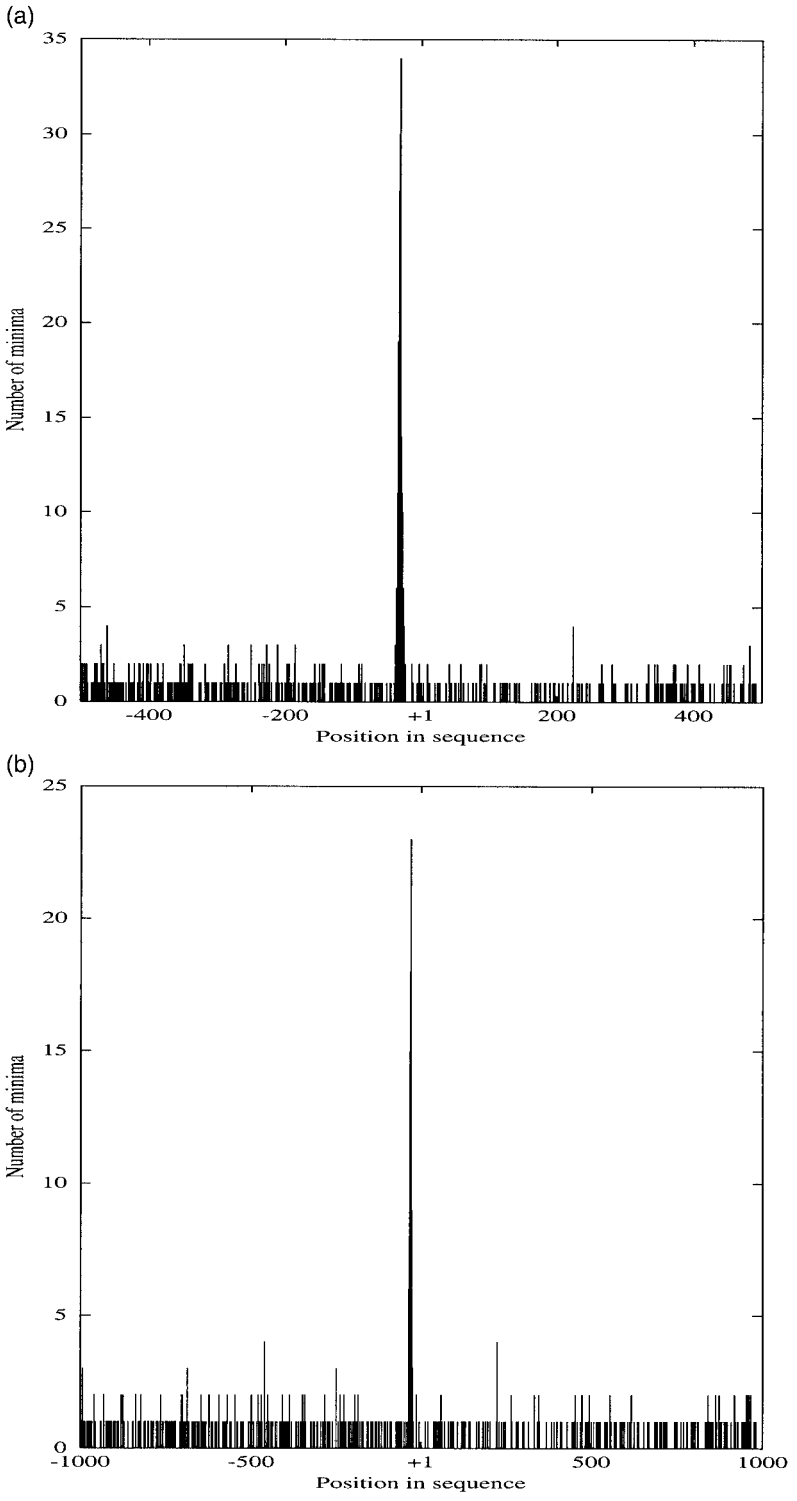

FIGURE 4 Histogram of the global minima distribution for (a) the vertebrate vs1000 and (b) vs2000 sets of promoter sequences. The TSS is situated at position +1 and the presumed TBP binding site should lie in the window [−39:−9].[45] Similar results (not shown) are obtained for the human genome sequences.

Table II Distribution of Local Energy Minima Within Two Windows Containing the Supposed TBP Binding Sites: (a) [−50:+1] and (b) [−200:+100][a]

(a) [−50:+1]

	hs1000				vs1000			
	Minima	Mean	Mode	σ	Minima	Mean	Mode	σ
DNA_{Alone}	192	−27	10	605	−28	−30	10	
DNA_{AA}	192	−26	−28	10	605	−27	−30	9
	Percentage of Local Energy Minima Within the Specified Windows							
	[Mode ± 2]		[−39:−9]		[Mode ± 2]		[−39:−9]	
DNA_{Alone}	41		100		49		90	
DNA_{AA}	44		100		50		90	

(b) [−200:+100]

	hs1000				vs1000			
	Minima	Mean	Mode	σ	Minima	Mean	Mode	σ
DNA_{Alone}	192	−53	−28	71	605	−52	−29	72
DNA_{AA}	192	−46	−28	68	605	−50	−30	71
	Percentage of Local Energy Minima Within the Specified Windows							
	[Mode ± 2]		[−39:−9]		[Mode ± 2]		[−39:−9]	
DNA_{Alone}	29		60		30		55	
DNA_{AA}	30		55		32		55	

[a] Results are again shown for the promoter sequences from the human genome, hs1000 and hs2000, and from vertebrate genomes, vs1000 and vs2000. σ is the standard deviation of the distributions.

Since TBP is a well-conserved protein among eukaryotes,[25–30] we have used the structural information from the TBP/DNA human complex[29] to analyze both the human and vertebrate promoter sequences chosen. Sequence scanning has been carried out in two ways. First, we used a target conformation (10 bp in length) deformed to match the crystallographic TBP binding site, as described in the section on single protein binding, and termed here DNA_{Alone}. Second, in order to estimate the role of the protein partner in the binding, we constructed a new target, DNA_{AA}, where the deformed DNA fragment is supplemented by the presence of the amino acids belonging to the protein–DNA interface (see Methods). In this case, the energy difference between the reference B-DNA conformation and the target conformation includes both the DNA deformation and a significant part of the protein–DNA interaction energy.

The statistics summarizing the location of the energy minima are presented in Table I and shown graphically in Figure 4. For both human and vertebrate promoter sequences, similar mean positions and rms deviations are found whether we use the DNA_{Alone} or DNA_{AA} targets. The modal point of the energy minima distribution is expected to fall around −30 position if TATA-box binding sequences are indeed being detected. Table I shows that this is the case for both human and vertebrate sequences and also that the concentration of minima around this point is significant: the [−50:+1] window containing approximately a third of the total observations and the [−200:+100] region containing almost half of them for the 1000 bp sequences. Concentrations are somewhat lower for the longer 2000 bp sequences. These results are encouraging; however, they indicate that the global minimum alone cannot be used directly to

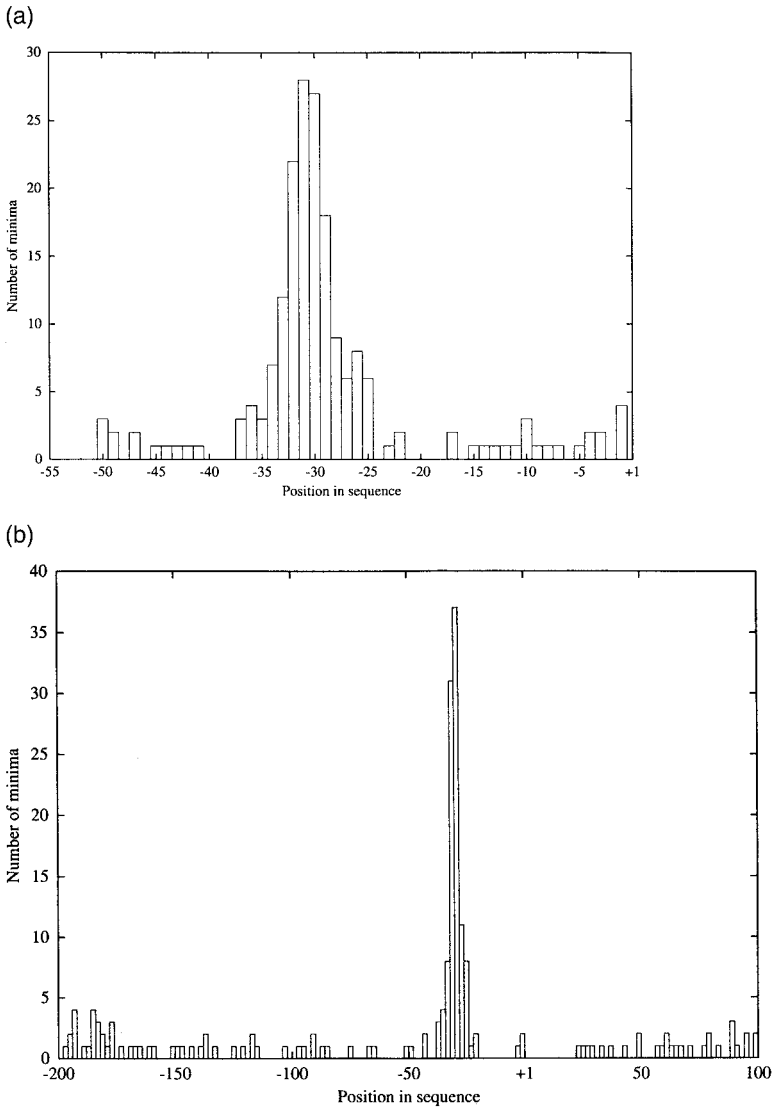

FIGURE 5 Histogram of the local minima distribution for the vertebrate vs1000 set of promoter sequences. The TSS is situated at the position +1. (a) [−50:+1] window, (b) [−200:+100] window.

detect the TBP binding sites. It is also found that correctly detected sites do not obey any fixed energy criteria. If we now limit ourselves to the energy distribution within the [−50:+1] and [−200:+100] windows, it is again found that there is a very significant proportion of the local energy minima within these windows that lie in the vicinity of the supposed TATA-box sites. Using Bucher's definition of such sites as [−39:−9], more than 50% of the minima lie within this region for the 300 bp window and at least 90% for the 50 bp window (see Table II and Figure 5). These results confirm that ADAPT can discriminate the region directly upstream the TSS as corresponding to TBP binding sites for a non-negligible part of the sequences studied.

On the basis of this analysis, it seems unreasonable to limit detection to the single global energy minimum. We have therefore extended our detection criteria to include energy minima that fall above the global minimum, but within a given percentage, p_{lim}, of the total energy variation seen for the fragments analyzed. Table III contains the results of this relaxed criteria, using a p_{lim} of 10 and 15% above the global minimum. Any sequence that has an energy minimum satisfying this criteria within the chosen 50 or 300 bp window is considered to be a positive detection.

Table III Number of TATA Boxes Correctly Described by ADAPT[a]

(a) Human Sequences

	hs1000						hs2000					
	[−50:+1]			[−200:+100]			[−50:+1]			[−200:+100]		
p_{lim}	0	10	15	0	10	15	0	10	15	0	10	15
	Percentage of Sites Detected											
DNA_{Alone}	27	57	72	42	79	88	16	44	57	25	58	74
DNA_{AA}	30	55	69	41	75	89	20	46	60	28	61	79
Total	32	61	74	47	82	92	21	51	62	31	66	82

(b) Vertebrate Sequences

	vs1000						vs2000					
	[−50:+1]			[−200:+100]			[−50:+1]			[−200:+100]		
p_{lim}	0	10	15	0	10	15	0	10	15	0	10	15
	Percentage of Sites Detected											
DNA_{Alone}	30	60	73	50	82	91	18	51	62	28	67	82
DNA_{AA}	32	60	72	51	81	93	22	51	65	32	69	85
Total	37	65	76	59	86	95	25	57	68	36	74	88

[a] Results are presented for the 50 and 300 bp windows around the TSS, with the TBP binding site alone (DNA_{Alone}) and the TBP binding site with the amino acids defining the surface contact (DNA_{AA}). The total line corresponds to all the unique TATA boxes detected by the two targets. (a) Human genome sets and (b) vertebrate genome sets.

When p_{lim} is set to either 10 or 15%, roughly 50% of sites are detected for the 50 bp window and roughly 70–80% for the 300 bp window. It can also be noted, in the last line of the table, that if we combine the sites detected using the DNA_{Alone} and DNA_{AA} target conformations, then we get a slightly higher detection rate than with either target alone.

The consensus sequences obtained with the sites detected using DNA_{Alone}, DNA_{AA}, or using both targets are presented in Table IV. They are very close to each other and to the experimental consensus. When p_{lim} is nonzero, sites detected by each of the systems show a larger sequence tolerance, but with variations for DNA_{Alone} and DNA_{AA}. These results confirm the previous observations (Table IV) that the two TBP targets localize somewhat different sequence sets and also that the presence of the interface amino acids makes for finer sequence discrimination.

Several refinements to this work can be envisaged, including using TBP conformations from other specific organisms, extending the target conformation to take into account the binding of multiple transcription factors, and naturally, improving the calculations of the DNA deformation and the DNA–protein interaction energies. Concerning the first of these possibilities, it is, however, interesting to note that the human TBP target conformation used presently in fact performed very well on the vertebrate promoter sequences we studied.

It should be noted that any sequence analyzed using ADAPT will lead to an energy minimum, whether or not the sequence in question contains the targeted binding site. Since we have already remarked that correct binding sites cannot be distinguished using an energy criteria alone, it will be necessary to supplement ADAPT results to obtain rigorous site detection. It is possible to imagine using aligned homologous sequences to improve detection in a manner analogous to phylogenetic footprinting.[119,120] However, the results already obtained suggest that ADAPT can provide information on protein binding sites that can represent a

Table IV Consensus Sequences of the Described TATA Boxes Using the vs1000 Set of Sequences with DNA_{Alone}, DNA_{AA}, or for the Combination of the Two Targets[a]

	[−50:+1]								[−200:+100]							
								p_{lim} 0%								
DNA_{Alone}	Y	A	T	A	W	A	A	A	T	A	T	A	W	A	W	R
DNA_{AA}	T	A	T	A	W	A	D	R	T	A	T	A	A	A	W	G
Total	T	A	T	A	A	A	A	R	T	A	T	A	W	A	W	R
								p_{lim} 10%								
DNA_{Alone}	Y	W	T	W	W	A	A	A	T	A	T	A	W	A	W	R
DNA_{AA}	T	A	T	A	W	W	D	R	T	A	T	W	W	A	D	R
Total	T	A	T	A	W	A	W	R	T	A	T	A	A	A	W	R
								p_{lim} 15%								
DNA_{Alone}	Y	Y	T	W	T	A	A	A	Y	H	T	H	W	W	W	R
DNA_{AA}	T	A	T	A	H	D	K	G	T	A	Y	Y	M	Y	D	R
Total	T	A	T	A	A	A	A	G	T	A	T	W	W	W	W	R

[a] At each position of the motif, the base (or combination of bases) whose occurrence is higher than 60% is represented.
Note: K = G/T; M = A/C; R = A/G; W = A/T; Y = C/T; H = A/C/T; V = A/C/G; D = A/G/T.

very useful complement to more standard lexical analyses of genome sequences and it is already feasible to couple ADAPT with consensus-search algorithms.

Genome Scanning—The Nucleosome

To illustrate the application of ADAPT to a multiple protein binding situation, we have looked at the possibility of detecting preferential nucleosome binding sites. This is a difficult problem since the histone core shows only slight base sequence preferences that, despite their biological importance, often involve free energy changes of the order of fractions of a kcal/mol. In this preliminary study, we have used a simplified model of the DNA wound around the nucleosome core. This model was built with an averaged sequence, $(dN_{146} \cdot dN_{146})$, using the superhelical symmetry option of JUMNA with a 10 bp repeat, a radius of curvature of 42 Å and a pitch of 24 Å taken from crystallographic data.[58] We have thus ignored the fine details of histone-induced DNA deformation, which may themselves be dependent on the sequence used crystallographically, and are also probably insufficiently resolved for direct use in molecular modeling. Lastly, DNA–protein interaction energies have not been taken into account since our simplified model is not compatible with the precise placement of amino acid side chains from the crystallographic data.

The target sequence of 146 bp was tested against a standard B-DNA reference conformation on a 860 bp sequence of the 5SRNA gene of *Xenopus borealis* for which preferential nucleosome binding positions have been determined experimentally using nuclease digestion.[121] The results obtained [Figure 6(a)] show two striking features. First, there is a clear 10 bp frequency oscillation in the deformation energy all along the sequence. This oscillation reflects the 10 bp helical symmetry repeat imposed on our target conformation. It implies that whatever the sequence being scanned, a 146 bp fragment of the sequence will always have a preferential direction of curvature. As the ADAPT sampling window moves along the sequence, this direction will be satisfied at some point i. Ten base pairs further along the sequence at $i + 10$, this direction will again be satisfied since a 10 bp move corresponds to a full rotation of the DNA double helix around its superhelical axis. However, at $i + 5$ the direction of curvature will be exactly opposed to the preferential direction, leading to a higher deformation energy.

The second notable feature is an amplitude modulation of the basic 10 bp oscillation of the deformation energy. This amplitude modulation leads to roughly sinusoidal envelopes, which cover about 150–200 bp. In several cases these envelopes seem to have some relationship to nucleosome binding locations within the *X. borealis* sequence [shown by the vertical dashed lines in Figure 5(a)]. If we contrast these results with those obtained by scanning a random sequence, with the same AT/GC content as that of *X. borealis,* it is seen that the 10 bp oscillation is always

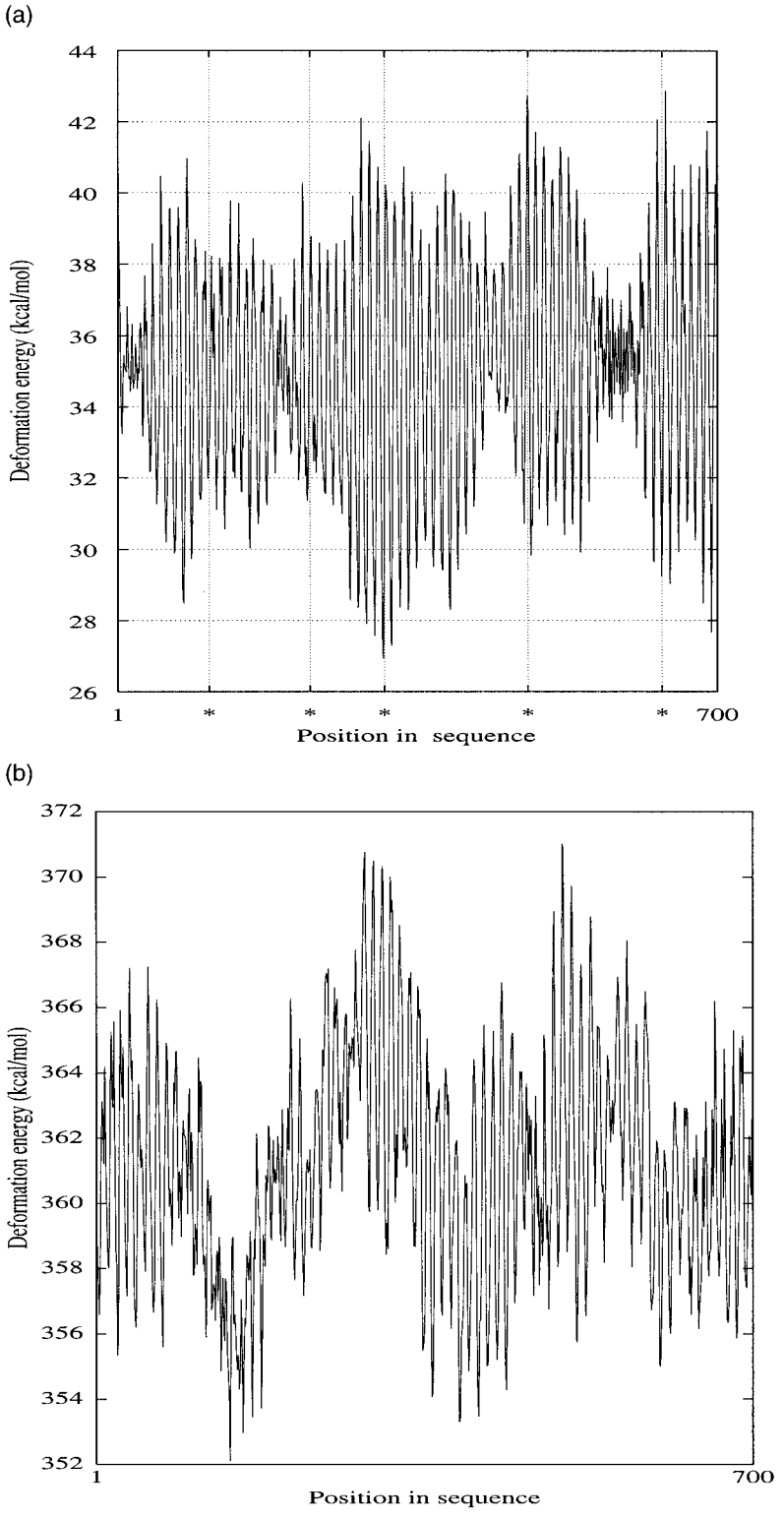

FIGURE 6 Probing genomic sequences with the deformation induced by the nucleosome formation. Deformation energy is plotted in function of the position in the analyzed sequence. (a) *Xenopus borealis* 5SRNA sequence (experimentally determined positions are indicated by vertical dashed lines); (b) random sequence.

present, but the amplitude modulation is much less structured, see Figure 6(b). It is too early to give a definitive explanation of the better defined envelopes seen for the biologically relevant sequence. Several improvements can be made to our nucleosome model by taking into account deviations from perfect superhelical symmetry and by including partial or complete DNA–protein interaction energies in the scanning. However, the results obtained suggest that there indeed seems to be a link between deformation energy and preferential nucleosome positioning.

CONCLUSIONS

The base sequence of DNA leads to variations in structure and flexibility that are now known to play a significant role in many biological functions via their contribution to protein–DNA recognition processes. It is, however, difficult to predict such sequence-dependent variations directly, and in general, there is insufficient experimental data to be able to formulate models that apply to more than one type of deformation or go beyond di- or trinucleotide parameter sets. Although molecular simulations can now be used to obtain both structural and dynamic information on short DNA fragments, they are also limited by the computational effort involved.

In this article, we have described an original approach, termed ADAPT, which is aimed at determining the physical properties of given base pair sequences within DNA and is sufficiently rapid to be applied to the analysis of complete genomes. ADAPT is based on making the base sequence of DNA a variable, via a multicopy approach. This involves introducing special nucleotides ("lexides") into the JUMNA internal/helicoidal molecular modeling program. These lexides contain all four standard bases, whose contribution to the conformational energy is controlled by variable coefficients. Using energy matrices taken from such calculations, it is then possible to compare the energies of a reference and a target conformation as a function of their common base sequence. ADAPT enables sequences to be tested using energy minimization, complete combinatorial searching or sequence scanning, in a manner analogous to the protein threading approach used in protein structure prediction.

We have presented applications of ADAPT to a number of examples involving DNA allomorphic transitions, DNA deformations and protein–DNA binding. In each case it has been shown that ADAPT is able to translate base sequences into physical properties that show a good correlation with available experimental data. Importantly, because of the multicopy approach, ADAPT calculations are very much faster than the equivalent molecular mechanics calculations, by a factor that in some cases can exceed 10^9. This also means that specific physical properties of genomic sequences can be scanned at a rate of roughly 2000 bp per second.

ADAPT allows the detection of conserved DNA features that are not directly visible from the sequence. It can be used alone, but it could be profitably combined with other DNA sequence analysis methods. Moreover, as nucleosome positioning can play a role in gene regulation,[67–69] it could be interesting to take into account the data coming from our studies of histone core deformation energy when considering the TBP target analysis of promoter regions.[51,75,122,123] Such combined approaches could be one step in the direction of a more integrated description of genomic information.

ADAPT can equally be extended to problems involving structural flexibility, which is also an important factor for DNA function.[124] It can be coupled with more sophisticated force fields, molecular representations, and solvent models than those used in the present tests. As such, it will hopefully prove to be a useful and complementary tool to standard lexical analyses of genome sequences.

REFERENCES

1. Marini, J. C.; Levene, S. D.; Crothers, D. M.; Englund, P. T. Proc Natl Acad Sci USA 1983, 79, 7664–7683.
2. Hagerman, P. J. Proc Natl Acad Sci USA 1984, 81, 4632–4636.
3. Hagerman, P. J. Nature 1986, 321, 449–450.
4. Olson, W. K.; Zhurkin, V. B. In Biological Structure and Dynamics; Proceedings of the Ninth Conversation, State University of New York 1995, Albany, NY; Ramaswamy, H.; Sarma, R. H.; Sarma, M. H., Eds.; Adenine Press: New York, 1996; p 341.
5. Shatzky-Schwartz, M.; Arbuckle, N. D.; Eisenstein, M.; Rabinovitch, D.; Bareket-Samish, A.; Haran, T. E.; Luisi, B. F.; Shakked, Z. J Mol Biol 1997, 267, 595–623.
6. Crothers, D. M.; Shakked, Z. In Oxford Handbook of Nucleic Acid Structures; Neidle, S., Ed.; Oxford University Press: New York, 1999; pp 455–470.
7. Bracco, L.; Kotlarz, D.; Kolb, A.; Diekmann, S.; Buc, H. EMBO J 1989, 8, 4289–4296.
8. Carmona, M.; Magasanik, B. J Mol Biol 1996, 261, 348–356.
9. Goodman, S. D.; Nash, H. A. Nature 1989, 341, 251–254.
10. Drew, H. R.; Travers, A. A. J Mol Biol 1985, 186, 773–790.

11. Travers, A. A.; Drew, H. R. Biopolymers 1997, 44, 423–433.
12. Wu, H. M.; Crothers, D. M. Nature 1984, 308, 509–513.
13. Travers, A. A. Curr Opin Struct Biol 1991, 1, 114–122.
14. Harrington, R. E. Mol Microbiol 1992, 6, 2549–2555.
15. Travers, A. A. DNA–Protein Interactions: Chapman & Hall: London, 1993.
16. Drew, H. R.; Travers, A. A. Nucleic Acid Res 1985, 13, 4445–4467.
17. Buratowski, S. Cell 1994, 77, 1–3.
18. Burley, S. K. Curr Opin Struct Biol 1996, 6, 69–75.
19. Pabo, C. O.; Sauer, R. T. Annu Rev Biochem 1992, 61, 1053–1095.
20. Burley, S. K.; Roeder, R. G. Annu Rev Biochem 1996, 65, 769–799.
21. Benoist, C.; Chambon, P. Nature 1981, 290, 304–310.
22. Marsh, T. L.; Reich, C. I.; Whitelock, R. B.; Olsen, G. J. Proc Natl Acad Sci USA 1994, 91, 4180–4184.
23. Rowlands, T.; Baumann, P.; Jackson, S. P. Science 1994, 264, 1326–1329.
24. Qureshi, S. A.; Bauman, P.; Rowlands, T.; Khoo, B.; Jackson, S. P. Nucleic Acids Res 1995, 23, 1775–1781.
25. Kim, J. L.; Nikolov, D. B.; Burley, S. K. Nature 1993, 365, 520–527.
26. Kim, Y.; Geiger, J. H.; Sigler, P. B. Nature 1993b, 365, 512–520.
27. Nikolov, D.; Burley, S. K. Nature Struct Biol 1994, 1, 621–637.
28. Kim, J. L.; Burley, S. K. Nature Struct Biol 1994, 1, 638–653.
29. Nikolov, D.; Chen, H.; Halay, A. D.; Hoffmann, A.; Roeder, R. G.; Burley, S. K. Proc Natl Acad Sci USA 1996, 93, 4862–4867.
30. Juo, Z. S.; Chiu, T. K.; Leiberman, P. M.; Baikalov, I.; Berk, A. J.; Dickerson, R. E. J Mol Biol 1996, 261, 239–254.
31. Guzikevich-Guerstein, G.; Shakked, Z. Nature Struct Biol 1996, 3, 32–37.
32. DeDecker, B. S.; O'Brien, R.; Fleming, P. J.; Geiger, J. H.; Jackson, S. P.; Sigler, P. J Mol Biol 1996, 264, 1072–1084.
33. Kosa, P. F.; Ghosh, G.; DeDecker, B. S.; Sigler, P. B. Proc Natl Acad Sci USA 1997, 94, 6042–6047.
34. Wobbe, C. R.; Struhl, K. Mol Cell Biol 1990, 267, 807–817.
35. Kim, J.; Klooster, S.; Shapiro, D. J. J Biol Chem 1995, 270, 1282–1288.
36. Parvin, J. D.; McCormick, R. J.; Sharp, P. A.; Fisher, D. E. Nature 1995, 23, 724–727.
37. Starr, D. B.; Hoopes, B. C.; Hawley, D. K. J Mol Biol 1995, 250, 434–446.
38. Grove, A.; Galeone, A.; Yu, E.; Mayol, L.; Geiduscheck, E. P. J Mol Biol 1998, 282, 731–739.
39. Bareket-Samish, A.; Cohen, I.; Haran, T. E. J Mol Biol 2000, 299, 955–977.
40. Flatters, D.; Zakrzewska, K.; Lavery, R. J Comp Chem 1997, 18, 1043–1055.
41. Pastor, N.; Pardo, L.; Weinstein, H. Biophys J 1997, 73, 640–652.
42. Pastor, N.; Pardo, L.; Weinstein, H. In Molecular Modeling of Nucleic Acids; ACS Symposium Series 682; Leontis, N. B.; SantaLucia, J., Jr., Eds.; American Chemical Society, Washington, DC, 1998; p 329.
43. de Souza, O. N.; Ornstein, R. L. Biopolymers 1998, 46, 403–415.
44. Strahs, D.; Schlick, T. J Mol Biol 2000, 301, 643–663.
45. Bucher, P. J Mol Biol 1990, 212, 563–578.
46. Hahn, S.; Buratowski, S.; Sharp, P. A.; Guarente, L. Proc Natl Acad Sci USA 1989, 86, 5718–5722.
47. Singer, V. L.; Wobbe, C. R.; Struhl, K. Genes Dev 1990, 4, 636–645.
48. Coleman, R. A.; Pugh, B. F. J Biol Chem 1995, 23, 13850–13859.
49. Fickett, J. W.; Hatzigeorgiou, A. G. Genome Res 1997, 7, 861–878.
50. Prestridge, D. S.; Burks, C. Human Mol Genet 1993, 2, 1449–1453.
51. Pedersen, A. G.; Baldi, P.; Chauvin, Y.; Brunak, S. Comput Chem 1999, 23, 191–207.
52. Kornberg, R. D. Annu Rev Biochem 1977, 46, 931–954.
53. McGhee, J. D.; Felsenfeld, G. Annu Rev Biochem 1980, 49, 1115–1156.
54. Widom, J. Annu Rev Biophys Biophys Chem 1989, 18, 365–395.
55. Widom, J. Annu Rev Biophys Biomol Struct 1998, 27, 285–327.
56. Kornberg, R. D.; Lorch, Y. Cell 1999, 98, 285–294.
57. Richmond, T. J.; Finch, J. T.; Rushton, B.; Rhodes, D.; Klug, A. Nature 1984, 311, 532–537.
58. Luger, K.; Mader, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. Nature 1997, 389, 251–260.
59. Luger, K.; Richmond, T. J. Curr Opin Struct Biol 1998, 8, 33–40.
60. Lowary, P. T.; Widom, J. Proc Natl Acad Sci USA 1997, 94, 1183–1188.
61. Trifonov, E. N.; Sussman, J. L. Proc Natl Acad Sci USA 1980, 77, 3816–3820.
62. Satchwell, S. C.; Drew, H. R.; Travers, A. A. J Mol Biol 1986, 191, 659–675.
63. Widlund, H. R.; Cao, H.; Simonsson, S.; Magnusson, E.; Simonsson, T.; Nielsen, P. E.; Kahn, J. D.; Crothers, D. M.; Kubista, M. J Mol Biol 1997, 267, 807–817.
64. Anselmi, C.; Bocchinfuso, G.; De Santis, P.; Savino, M.; Scipioni, A. Biophys J 2000, 79, 601–613.
65. Brukner, I.; Sanchez, R.; Suck, D.; Pongor, S. J Biomol Struct Dynam 1995, 13, 309–317.
66. Richard-Foy, H.; Hager, G. L. EMBO J 1987, 6, 2321–2328.
67. Svaren, J.; Hörz, W. Curr Opin Genet Dev 1996, 6, 164–170.
68. Zhu, Z.; Thiele, D. J. Cell 1996, 87, 459–470.

69. Panetta, G.; Buttinelli, M.; Flaus, A.; Richmond, T. J.; Rhodes, D. J Mol Biol 1998, 282, 683–697.
70. Spangenberg, C.; Eisfeld, K.; Stunkel, W.; Luger, K.; Flaus, A.; Richmond, T. J.; Truss, M.; Beato, M. J Mol Biol 1998, 278, 725–739.
71. Flaus, A.; Richmond, T. J. J Mol Biol 1998, 275, 427–441.
72. Flaus, A.; Luger, K.; Tan, S.; Richmond, T. J. Proc Natl Acad Sci USA 1996, 93, 1370–1375.
73. Tuerk, C.; Gold, L. Science 1990, 249, 505–510.
74. Pedersen, A. G.; Baldi, P.; Brunak, S.; Chauvin, Y. Intell Syst Mol Biol 1996, 4, 182–191.
75. Pedersen, A. G.; Baldi, P.; Chauvin, Y.; Brunak, S. J Mol Biol 1998, 281, 663–673.
76. Pedersen, A. G.; Jensen, L. J.; Brunak, S.; Staerfeldt, H. H.; Ussery, D. W. J Mol Biol 2000, 299, 907–930.
77. Kono, H.; Sarai, A. Proteins 1999, 35, 114–131.
78. Benham, C. J. CABIOS 1996, 12, 375–381.
79. Yeramian, E. Gene 2000, 255, 139–150.
80. Yeramian, E. Gene 2000, 255, 151–168.
81. Lafontaine, I.; Lavery, R. Biophys J 2000, 79, 680–685.
82. Lafontaine, I.; Lavery, R. Comb Chem High Throughput Screening 2001, in press.
83. Lafontaine, I.; Lavery, R. Curr Opin Struct Biol 1999, 9, 170–176.
84. Lavery, R.; Zakrzewska, K.; Sklenar, H. Comp Phys Comm 1995, 91, 135–158.
85. Sanghani, S. R.; Zakrzewska, K.; Harvey, S. C.; Lavery, R. Nucleic Acid Res 1996, 24, 1632–1637.
86. Lavery, R.; Parker, I.; Kendrick, J. J Biomol Struct Dynam 1986, 4, 443.
87. Lavery, R. In Structure and Expression Vol 3. DNA Bending and Curvature; Olson, W. K.; Sarma, R. H.; Sarma, M. H.; Sundaralingam, M., Eds.; Adenine Press: New York, 1988; p 191.
88. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. J.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1995, 117, 5179–5197.
89. Cheatham, T. E., III; Kollman, P. A. J Biomol Struct Dynam 1998, 4, 845–862.
90. Hingerty, B.; Ritchie, R. H.; Ferrel, T. L.; Turner, J. E. Biopolymers 1985, 24, 427–439.
91. Cluzel, P.; Lebrun, A.; Heller, C.; Lavery, R.; Viovy, J. L.; Chatenay, D.; Caron, F. Science 1996, 271, 792–794.
92. Lebrun, A.; Lavery, R. Curr Opin Struct Biol 1997, 7, 348–354.
93. Bernet, J.; Zakrzewska, K.; Lavery, R. J Mol Struct (Theochem) 1997, 399, 473–482.
94. Khoel, P.; Delarue, M. Curr Opin Struct Biol 1996, 6, 222–226.
95. Elber, R.; Karplus, M. J Am Chem Soc 1990, 112, 9161–9175.
96. Stultz, C. M.; Karplus, M. Proteins 1999, 37, 512–529.
97. Verkhivker, G.; Rejto, P. A. Proc Natl Acad Sci USA 1996, 93, 60–64.
98. Vasquez, M. Biopolymers 1995, 36, 53–70.
99. Zheng, Q.; Rosenfeld, R.; DeLisi, C.; Kyle, D. J. Protein Sci 1994, 3, 493–506.
100. Khoel, P.; Delarue, M. J Mol Biol 1994, 239, 249–275.
101. Kono, H.; Nishiyama, M.; Tanokura, M.; Doi, J. Protein Eng 1998, 11, 47–52.
102. Voigt, C. A.; Mayo, S. L.; Arnold, F. H.; Wang, Z. G. Proc Natl Acad Sci USA 2001, 98, 3778–3783.
103. Kono, H.; Doi, J. Proteins 1994, 19, 244–255.
104. Kopka, M. L.; Yoon, C.; Goodsell, D.; Pjura, P.; Dickerson, R. E. Proc Natl Acad Sci USA 82, 1376–1380.
105. Roitberg, A.; Elber, R. J Chem Phys 1991, 95, 9277–9287.
106. Zheng, Q.; Rosenfeld, R.; Kyle, J. D. J Chem Phys 1993, 99, 8892–8896.
107. Westhead, D. R.; Collura, V. P.; Eldridge, M. D.; Firth, M. A.; Li, J.; Murray, C. W. Protein Eng 1995, 18, 1197–1204.
108. Lebrun, A.; Lavery, R.; Weinstein, H. Protein Eng 2001, 14, 233–243.
109. Leng, M. Biochim Biphys Acta 1985, 825, 339–344.
110. Wang, A. H.-J.; Hakokshima, T.; Marel, G. A.; van Boom, J. H.; Rich, A. Cell 1984, 37, 321–331.
111. Ho, P. S.; Ellison, M. J.; Quigley, G. J.; Rich, A. EMBO J 1986, 5, 2737–2744.
112. Klump, H. H.; Schmid, E.; Wosgien, M. Nucleic Acids Res 1993, 21, 2343–2348.
113. Hagerman, P. J. Biochemistry 1985, 24, 7033–7037.
114. Kopka, M. L.; Yoon, C.; Goodsell, D.; Pjura, P.; Dickerson, R. E. J Mol Biol 1985, 183, 553–563.
115. Marky, L. A.; Breslauer, K. J. Proc Natl Acad Sci USA 1987, 84, 4359–4363.
116. Wingender, E.; Dietze, P.; Karas, H.; Knuppel, R. Nucleic Acids Res 1996, 24, 238–241.
117. Bucher, P.; Trivonov, E. N. Nucleic Acids Res 1986, 14, 10009–10026.
118. Périer, R. C.; Praz, V.; Junier, T.; Bonnard, C.; Bucher, P. Nucleic Acids Res 2000, 28, 302–303.
119. Tagle, D. A.; Koop, B. F.; Goodman, M.; Slightom, J. L.; Hess, D. L.; Jones, R. T. J Mol Biol 1998, 203, 439–455.
120. Gelfand, M. S.; Koonin, E. V.; Mironov, A. A. Nucleic Acids Res 2000, 28, 695–705.
121. Drew, H. R.; Calladine, C. R. J Mol Biol 1987, 195, 143–173.
122. Polach, K. J.; Widom, J. J Mol Biol 1996, 258, 800–812.
123. Ioshikhes, I.; Trifonov, E. N.; Zhang, M. Q. Proc Natl Acad Sci USA 1999, 96, 2891–2895.
124. Koenig, P.; Richmond, T. J. J Mol Biol 1993, 233, 139–154.