Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 11163–11168, September 1998
Biophysics

DNA sequence-dependent deformability deduced from protein–DNA crystal complexes

WILMA K. OLSON*†, ANDREY A. GORIN‡, XIANG-JUN LU*, LYNETTE M. HOCK*, AND VICTOR B. ZHURKIN†§

*Department of Chemistry, Rutgers University, New Brunswick, NJ 08903; ‡Sloan-Kettering Cancer Center, New York, NY 10021; and §National Cancer Institute, National Institutes of Health, Bethesda, MD 20892

Communicated by Donald M. Crothers, Yale University, New Haven, CT, June 30, 1998 (received for review April 13, 1998)

ABSTRACT **The deformability of double helical DNA is critical for its packaging in the cell, recognition by other molecules, and transient opening during biochemically important processes. Here, a complete set of sequence-dependent empirical energy functions suitable for describing such behavior is extracted from the fluctuations and correlations of structural parameters in DNA–protein crystal complexes. These elastic functions provide useful stereochemical measures of the local base step movements operative in sequence-specific recognition and protein-induced deformations. In particular, the pyrimidine-purine dimers stand out as the most variable steps in the DNA–protein complexes, apparently acting as flexible "hinges" fitting the duplex to the protein surface. In addition to the angular parameters widely used to describe DNA deformations (i.e., the bend and twist angles), the translational parameters describing the displacements of base pairs along and across the helical axis are analyzed. The observed correlations of base pair bending and shearing motions are important for nonplanar folding of DNA in nucleosomes and other nucleoprotein complexes. The knowledge-based energies also offer realistic three-dimensional models for the study of long DNA polymers at the global level, incorporating structural features beyond the scope of conventional elastic rod treatments and adding a new dimension to literal analyses of genomic sequences.**

In addition to the genetic message, DNA base sequence carries a multitude of other signals related to the manipulation of the long, thread-like molecule. Primary sequences of nucleic acid bases determine three-dimensional structures whose physical properties reflect the constituent residues. The existing library of solved DNA crystal structures (1), for example, reveals subtle sequence-dependent irregularities in the orientation and displacement of adjacent residues (2). Duplex stability under a given set of environmental conditions also depends to good approximation on the identity of the 10 nearest neighbor base pairs (3). The linear sequence of genetic information thus expands into a base sequence-dependent spatial and energetic code that governs the global organization of the double helix and its susceptibility to interactions with other molecules.

Interest in understanding the physical properties of genomic DNA has prompted the development of new approaches to analyze and depict the sequence-dependent bending and twisting of neighboring base pairs. Studies of gel migration (4), chain cyclization kinetics (4, 5), and nucleosome phasing (6) have yielded a variety of sequence-dependent models to account for the observed data. Furthermore, collected oligonucleotide crystal structures show similar conformational trends (2), although there are discrepancies between the x-ray and solution assessments of the direction of bending at a few dimer steps (7, 8). The solid state data additionally reveal sequence-dependent differences in the displacement of base pairs [see, e.g., Slide (2)], a feature not usually considered in simple models fitted to solution data (4–6).

The thermal fluctuations deduced from solution studies of long DNA (9) constitute yet another important feature of the double helix. These macroscopic properties, which are not easily interpreted at the base pair level, typically are described by changes in the elastic constants of idealized sequence-independent models that match the observed data.

The sequence-specific recognition of DNA by proteins and other ligands, however, calls for a more sophisticated description of chain flexibility than can be deduced from uniform elastic models. In particular, DNA responds to protein binding through sequence-dependent kinking and intercalation (10–12). This deformability is essential at both the global and local levels, serving as a potential long-range signal for molecular recognition as well as accommodating the local distortions of the double helix induced by tight binding, i.e., the so-called conformational recognition or indirect readout of DNA residues (13, 14). Analysis of such systems requires models that incorporate detailed stereochemical features of DNA sequences. One of the best sources of this information is the database of protein–DNA crystal complexes (1), which has grown to the point where there are now enough data to extract duplex properties (means and fluctuations) at the level of nearest neighbors, i.e., the 10 unique base pair steps. Although trimers and tetramers may be preferable in principle, there are not yet sufficient crystallographic data to address DNA deformability at this level. A further limitation of the current data is that anisotropic structure factors are infrequently reported, e.g., only one B-DNA structure (15) and no protein–DNA complexes. Therefore, the "real-time" flexibility of DNA in any given complex cannot be determined directly from the co-crystal data. Still, one can use the ensemble-averaged parameters characterizing the DNA variability in these complexes to learn about the intrinsic motions of the double helix.

This approach can be justified as follows. Different proteins impose different sorts of forces on DNA, so the distortions of DNA brought about by many kinds of proteins effectively cancel one another (major vs. minor groove bending, etc.). Therefore, we presume that, after averaging over a large ensemble of complexes, the natural conformational response of DNA will surface. (These general principles will not necessarily predict the interactions of DNA with amino acids in specific complexes.)

Here, we report the behavior of dimer steps collected from 92 protein–DNA crystal complexes and offer a set of empirical energy functions that describe the observed sequence-dependent deformability of the composite structures. To describe the DNA variability in this way, we apply a formal "culling" procedure to the x-ray data set to obtain Gaussian distributions of conformational parameters. Table 1 summarizes this information in terms of the spread of the six

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

†To whom reprint requests should be addressed. e-mail: olson@rutchem.rutgers.edu or zhurkin@structure.nci.nih.gov.

Table 1. Average values and dispersion of base pair step parameters* in DNA crystal complexes

Step	N	Twist, deg		Tilt, deg		Roll, deg		Shift, Å		Slide, Å		Rise, Å	
CG	88	36.1	(5.5)	0	(4.2)	5.4	(5.2)	0	(0.87)	0.41	(0.56)	3.39	(0.27)
CA	110	37.3	(6.5)	0.5	(3.7)	4.7	(5.1)	0.09	(0.55)	0.53	(0.89)	3.33	(0.26)
TA	134	37.8	(5.5)	0	(2.7)	3.3	(6.6)	0	(0.52)	0.05	(0.71)	3.42	(0.24)
AG	106	31.9	(4.5)	−1.7	(3.3)	4.5	(3.4)	0.09	(0.69)	−0.25	(0.41)	3.34	(0.23)
GG	97	32.9	(5.2)	−0.1	(3.7)	3.6	(4.5)	0.05	(0.76)	−0.22	(0.64)	3.42	(0.24)
AA	129	35.1	(3.9)	−1.4	(3.3)	0.7	(5.4)	−0.03	(0.57)	−0.08	(0.45)	3.27	(0.22)
GA	117	36.3	(4.4)	−1.5	(3.8)	1.9	(5.3)	−0.28	(0.46)	0.09	(0.70)	3.37	(0.26)
AT	140	29.3	(4.5)	0	(2.5)	1.1	(4.9)	0	(0.57)	−0.59	(0.31)	3.31	(0.21)
AC	137	31.5	(4.2)	−0.1	(3.1)	0.7	(3.9)	0.13	(0.59)	−0.58	(0.41)	3.36	(0.23)
GC	86	33.6	(4.7)	0	(3.9)	0.3	(4.6)	0	(0.61)	−0.38	(0.56)	3.40	(0.24)
MN[†]													
P·DNA[‡]	1,840	34.2	(5.5)	0	(3.6)	2.7	(5.2)	0	(0.64)	−0.09	(0.69)	3.36	(0.25)
P′·DNA[§]	2,114	33.2	(6.6)	0	(4.0)	4.0	(9.3)	0	(0.65)	−0.05	(0.77)	3.41	(0.47)
B–DNA[§]	724	35.4	(6.3)	0	(3.4)	1.4	(5.1)	0	(0.51)	0.35	(0.78)	3.32	(0.19)

Dimer steps taken from B–DNA and protein–DNA files in the Nucleic Acid Database (1). See http://rutchem.rutgers.edu/~olson/pdna.html for a complete listing with citations.

*Parameters computed with COMPDNA (2). Dispersion noted in parentheses.

[†] Average MN parameters for a generic MpN dimer step are based on equal weighting of average parameters of the 16 common dimers, i.e., AA and TT, AG and CT, etc., have identical averages except for different signs of Tilt and Shift (16). The number of MN entries thus exceeds the sum of the 10 unique dimers, and self-complementary steps, e.g., CG, are counted twice in the sample. The dispersion of MN steps is computed from weighted mean–square and mean values, $\Delta\theta_{\mathrm{MN}} = (\langle\theta^2_{\mathrm{MN}}\rangle - \langle\theta_{\mathrm{MN}}\rangle^2)^{1/2}$.

[‡] Averages $\langle\theta\rangle$ and rms deviations $\Delta\theta$ exclude terminal dimer units, which may adopt alternate conformations or be affected by crystal packing, and steps with single-stranded nicks and mismatches. Protein-bound DNA samples also omit "melted" residues, where displacements of complementary base pairs (16) deviate from their averages by $>3\Delta\theta$ before culling. To separate intrinsic from protein-induced conformational deformations and to obtain quasi-Gaussian distributions, we excluded outlying states of extreme bending, twisting, and/or stretching in a stepwise fashion until there were no base step parameters outside the $3\Delta\theta$ limit.

[§] Unrestricted samples that include dimer steps outside the $3\Delta\theta$ limit with secondary clusters of data points at some steps, i.e., CA in B–DNA (2). The unculled B–DNA data set presented here differs insignificantly from the B–DNA sample with no states outside the $3\Delta\theta$ limit.

parameters that relate successive base pair planes: three angles—Twist, Roll, Tilt—and three distances—Shift, Slide, Rise (16). To quantify the observed deformations of the DNA duplex in terms of harmonic energy functions, we adapted a methodology conventionally used to characterize the fluctuations of individual atoms in molecular structures (17).

Our analysis differs from standard energy treatments in two respects: (*i*) we express the DNA deformation energy in terms of the six "base pair step parameters" rather than the atomic coordinates or internal chemical parameters, i.e., bond lengths, valence angles, torsions; and (*ii*) we extract energies from the distributions of these six parameters in crystal complexes by using an "inverse harmonic analysis." In addition, our adoption of a complete set of interdependent dimer parameters, i.e., three rotations and three translations, refines current interpretations of DNA structure based on only three autonomous rotations (4–6, 18).

METHODS

Dimer Energy Functions. Conformational fluctuations of DNA are described by the increase in energy brought about by instantaneous fluctuations of the six-dimer step parameters, $\Delta\theta_{\mathrm{i}} = (\theta_{\mathrm{i}} - \theta_{\mathrm{i}}{}^{\mathrm{o}})$, $i = 1{,}6$, from their equilibrium (i.e., minimum energy) values, $\theta_{\mathrm{i}}{}^{\mathrm{o}}$. The energy of each dimer is approximated by a harmonic function,

$$E = E_{\mathrm{o}} + \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6} f_{ij}\,\Delta\theta_{\hat{i}}\,\Delta\theta_{\hat{j}}, \qquad [1]$$

where E_{o} is the minimum energy and the f_{ij} are elastic constants impeding deformations of the given step. If the f_{ij} are collected in the dimer stiffness matrix $\mathbf{F}$, the double summation in Eq. **1** reduces to $\mathbf{\Theta}^{\mathrm{T}}\mathbf{F}\mathbf{\Theta}$, where the elements of $\mathbf{\Theta}$ and its transpose $\mathbf{\Theta}^{\mathrm{T}}$ are the $\Delta\theta_i$.

According to conventional procedures (17), the force constants are evaluated from a known (typically all-atom) energy function, i.e., $f_{ij} = (\partial^2 E/\partial\theta_i\partial\theta_j)$ and the pairwise covariances of conformational variables are deduced from the inverse of matrix $\mathbf{F}$, i.e., $\langle\Delta\theta_i\Delta\theta_j\rangle k_BT = [\mathbf{F}^{-1}]_{ij}$, where k_B is the Boltzmann constant and T is the temperature in Kelvin. Here we carry out an "inverse harmonic analysis." The observed covariance of conformational parameters from x-ray structures is used as input, and the f_{ij} are obtained as output. That is, we substitute the observed pairwise parameter averages, $\langle\Delta\theta_i\Delta\theta_j\rangle = \langle\theta_i\ \theta_j\rangle - \langle\theta_i\rangle\langle\theta_j\rangle$, for the elements of $\mathbf{F}^{-1}$, i.e., the covariance matrix, and then find $\mathbf{F}$ by matrix inversion. The effective temperature is unknown but can be estimated by scaling the data against standard solution measurements, e.g., persistence length (9). The mean dimer conformation, measured by the $\langle\theta_i\rangle$ in Table 1, is taken as the minimum energy reference state. Although similar approaches are used in the analysis of all-atom dynamics trajectories (19), this formalism never has been applied to generalized coordinates such as Twist and Roll. Using these parameters (instead of Cartesian coordinates) reduces the number of independent degrees of freedom per base pair from ≈200 to 6 without loss of either the sequence specificity or the mobility of duplex structure.

Data Selection. In applying this approach, we restrict attention to dimer parameters that cluster tightly in quasi-normal distributions consistent with harmonic behavior. To obtain these distributions, we cull steps with obvious structural irregularities from the data and omit states of extreme bending, twisting, and/or stretching; see Table 1 for details and Fig. 1 for representative images. Overall, the chosen set of protein-bound DNA steps (Table 1) follows structural tendencies well known for B-DNA; the dimer steps preferentially bend via Roll (20) and are displaced locally through Shift or Slide (2). Moreover, the spread of parameters in our "unperturbed" P·DNA data set resembles the variability of dimer steps in ligand-free B-DNA crystals (compare areas covered by ellipses in Fig. 1). On the other hand, there are important differences between the protein-bound and unbound DNA dimer steps; the association of protein decreases Twist, increases Roll, and reduces Slide compared with the mean values in pure B-DNA, suggesting a partial shift of conformational parameters toward the A-form (21).

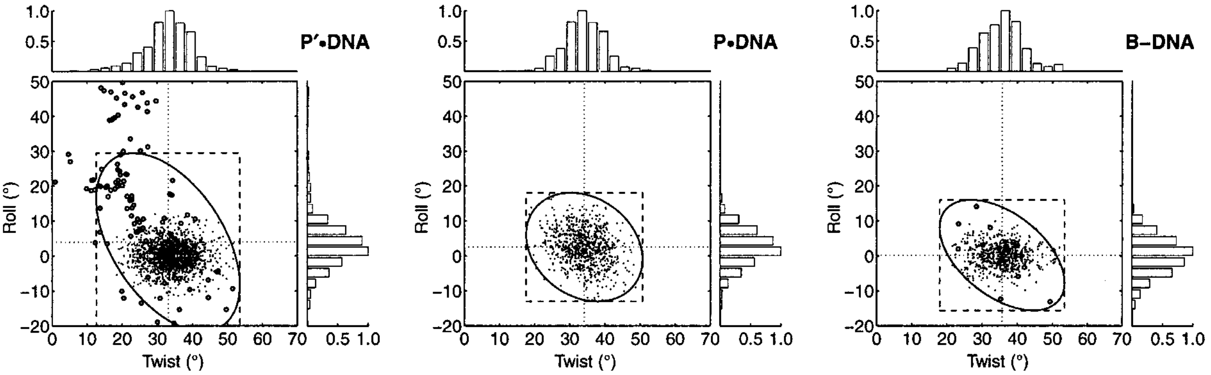

FIG. 1. Scatter plots in the Twist–Roll plane of base step parameters in protein-bound and B–DNA crystal complexes. Dots correspond to the "unperturbed" P·DNA sample used in averages (Table 1) and derived deformabilities (Table 2) and circles to states of extreme bending, twisting, and stretching (included in B– and P′·DNA samples). Rectangles enclose points lying within three rms ($3\Delta\theta$) deviations of ⟨Twist⟩ and ⟨Roll⟩. Ellipses are projections of the six-dimensional equi-potential surfaces on the Twist–Roll plane obtained from the 2 × 2 Twist–Roll covariance matrix (see text); these contours correspond to energies of 4.5 k_BT ("$3\Delta\theta$ ellipses"). Histograms on the edges of the scatter plots are scaled with respect to a value of unity for the most populated angular ranges (422–474 occurrences in P·DNA and 162–168 for B–DNA).

RESULTS AND DISCUSSION

Sequence-Dependent Deformabilities. The sequence-dependent characteristics of the protein–DNA complexes follow trends reported in earlier surveys of B–DNA structures (2). The average twisting of base pair steps increases in the same order within the three standard chemical classes: pyrimidine–purine, purine–purine, purine–pyrimidine. For example, among pyrimidine–purines, the order is Twist(AT) < Twist(AC) < Twist(GC) in both protein–DNA and B–DNA crystals; see Table 1 and ref. 2. The dispersion of base pair parameters is generally larger in the protein-bound sample than in the B–DNA data, which are poorly represented at some steps, e.g., GG. (Updated B–DNA data are given at http://rutchem.rutgers.edu/~olson/pdna.html). The distinctive dispersion of parameters for different dimers is another indication of the sequence-dependent deformability of the double helix.

The sequence-dependent force constants derived from the protein-bound DNA data (listed at the above Web site) reveal steeper changes in energy than expected from simpler potentials based solely on the dispersion of individual parameters, i.e., without cross terms in the stiffness matrix (18). For example, the CG twisting force constant, 0.047 $k_BT\mathrm{deg}^{-2}$, corresponds to ±4.6° fluctuations in Twist, i.e., angular changes that raise the energy by $k_BT/2$, whereas the observed dispersion of CG Twist is 5.5° (Table 1). This difference reflects the influence of other conformational variables, such as Roll and Slide. Furthermore, the positive value of the Twist–Roll force constant produces an energy pathway involving a decrease in one angle and an increase in the other, mimicking the well known variation of these parameters (2, 21). The negative Tilt–Shift constants, by contrast, point to correlations not usually highlighted in the x-ray literature (2) but easily understood at a mechanical level. That is, close base–base contacts engendered by tilting are partially relieved by translations of base pairs along their dyad axes.

Equi-Potential Surfaces. The equi-potential surfaces in Fig. 2 illustrate the sequence-dependent deformability and conformational interdependence in DNA dimer steps. The contours of the derived energy functions are reminiscent of the Ramachandran–Sasisekharan diagrams (22) long used to evaluate polypeptide conformation but closer in spirit to the knowledge-based potentials (23) developed to treat long range interactions in proteins. Unusual geometries, such as the states of extreme twisting and bending omitted from the statistical analysis, immediately stand out against the energy profiles. In addition, the derived potential functions provide a graduated scale of comparison that is potentially useful when checking the variation of conformational parameters in multi-dimensional space.

The contour plots reveal the unique conformational character of different dimer steps. For example, the anticorrelations of Twist and Roll depend on sequence: some dimer steps (CA, AG, GG) twist more easily than bend, others (TA, AA, GA) roll more easily than twist, and the remainder span comparable angular ranges. As noted above, the variation in Twist and Roll frequently is coupled to changes in one or more translational parameters, i.e., Slide, Shift. The Rise, which

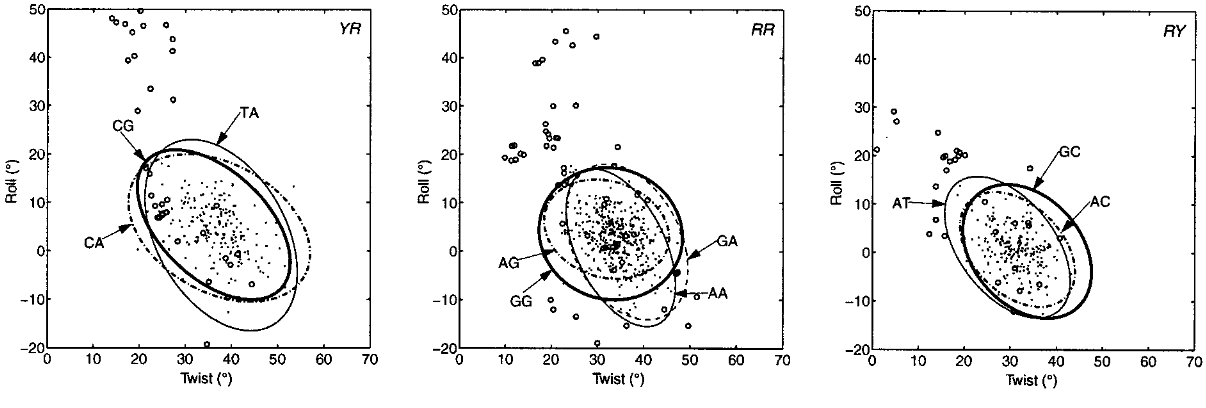

FIG. 2. Scatter plots in the Twist–Roll plane of observed parameters and derived energy contours of pyrimidine–purine (YR), purine–purine (RR), and purine–pyrimidine (RY) dimer steps. See legend to Fig. 1. Note the gradually decreasing areas of the $3\Delta\theta$ ellipses from left to right. Corresponding plots for B–DNA are found at the following URL: http://rutchem.rutgers.edu/~olson/pdna.html.

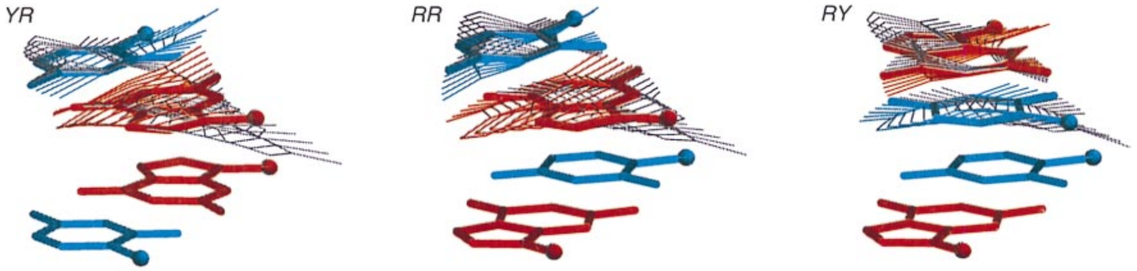

FIG. 3. Sequence-dependent motions along the longest principal axes of P·DNA dimer steps: pyrimidine–purine (YR), purine–purine (RR), and purine–pyrimidine (RY) steps. Nonequilibrium forms, corresponding to the parametric changes in Table 2, are superimposed on the average dimer structures (thickened bonds). Perturbed conformations correspond to states deformed along the longest principal axes of the derived energy functions, with increments equal to changes of $\langle\lambda_1^2\rangle^{-1/2}$ (Table 2) and energies ranging from 0 to 12.5 k_BT for $\pm5\langle\lambda_1^2\rangle^{-1/2}$ deviations. Opposing directions of fluctuations are distinguished by color-coded (Y, light blue; R, red) and gray images. Motions illustrated with respect to a reference frame in the 5′ base pair (i.e., the M in MpN dimers). Views from the leading strand. Base pairs represented as ideal Watson–Crick pairs with C1′ atoms of rest structures noted by circles. Note the decreasing range of motions from left to right.

spans a very narrow range of values, by contrast, varies almost independently of all other parameters in the x-ray sample, the only notable correlation involving Tilt at AG steps.

Intrinsic Motions. The molecular images in Fig. 3 illustrate the pathways of preferred conformational changes in base pair steps. The sets of low energy movements, which lie along the longest principal axes of the derived energy surfaces, i.e., in the direction of most probable conformational change, are reminiscent of the normal modes of vibration of small molecules. The natural deformations are a composite of the parameters conventionally used to describe them. The illustrated motions involve combinations of Twist and Roll plus varying degrees of residue translation. As is clear from the precise angular and translational changes reported in Table 2 for the lowest energy modes of the 10 dimers, correlations between Twist and Roll dominate the preferred movements of all steps. Some steps (CA, TA, AG), however, incorporate significant translational changes in the deformations whereas others (CG, AT, AC) involve essentially no base pair displacement. A few steps (CA, AG, GG, AA, GA) also incorporate changes in Tilt in their lowest energy modes, although the observed variations are consistently lower than those of Roll. The smallest conformational changes (along the shortest principal axes of the energy functions) always involve Rise.

The volumes of conformation space, V_{step}, obtained from the eigenvalues of the stiffness matrix **F** (17) offer a measure of dimer deformability (Table 2). When combined with the average step parameters in Table 1, the set of volumes complete the "fingerprint" of each DNA dimer, making it possible to distinguish all 10 bp steps, some of which, e.g., AG and GG, are nearly identical on the basis of their average geometries. Also apparent from the volumes and from Figs. 2 and 3 is the flexibility of AA steps (comparable to that of other steps), which persist even when all highly distorted protein-complexed dimers, e.g., steps from the TATA-box binding protein (TBP) complex (24, 25), are removed from the data set. The broad range of AA states argues against notions of AA stiffness based on the resistance of poly(dA)·poly(dT) to nucleosome formation at low temperature (26, 27) and supports observed nucleosome formation at higher temperature (28). Surprisingly, division of the AA steps into A-runs (i.e., the 70 steps within A_n sequences, where $n \geq 3$) and isolated AA dimers suggests that the former steps may be slightly more deformable than the latter: Twist, Roll = (34.8 ± 4.2°, 0.6 ± 6.3°) vs. (35.4 ± 3.7°, 0.9 ± 4.2°).

Pyrimidine–Purine "Hinges" and Protein Binding. DNA binding proteins clearly take advantage of the intrinsic conformational mechanics of the double helix. First, extreme states tend to lie along the pathways of slowest energy ascent (compare data points outside the low energy ellipses in Figs. 1 and 2 with energy contours in Fig. 2). For example, large protein-induced bends in the omitted steps entail significantly more Roll than Tilt, i.e., both "unperturbed" and "distorted" dimer steps exhibit bending anisotropy (20), but the effect is more pronounced in the latter case (*cf.* P·DNA and P′·DNA in Table 1). Second, the high energy states adhere closely to the sequence-dependent rules described by the derived energy functions. In particular, the highly bent and stretched steps in TATA-box binding protein–DNA co-crystal complexes (24, 25), which are distinctly separated from the working data, follow the general trends of the energy maps. The principal axes of the omitted data roughly coincide with those for the reference TA, AA, and AT points.

Table 2. Contributions to base pair flexibility in protein-bound DNA dimers

	V_{step}*, deg^{3}Å^3							
Step	P·DNA	B–DNA‡	±ΔTwist†, deg	±ΔTilt†, deg	±ΔRoll†, deg	±ΔShift†, (Å)	±ΔSlide†, (Å)	±ΔRise†, Å
CG	12.1	1.3	4.9	0	−4.3	0	−0.08	0.05
CA	9.8	1.8	6.2	1.0	−3.1	0.03	0.52	0.03
TA	6.3	1.7	3.9	0	−6.1	0	0.27	0.06
AG	2.1	0.3	4.4	−0.4	−1.1	−0.28	−0.02	0.04
GG	6.1	—	4.8	1.6	−1.8	0.10	0.10	0.08
AA	2.9	0.7	2.8	−0.4	−5.2	−0.08	−0.12	0.02
GA	4.5	0.3	2.9	0.2	−5.0	−0.05	0.02	0.14
AT	1.6	0.4	3.8	0	−4.3	0	0.01	0.06
AC	2.3	—	3.6	0	−3.2	0.02	0.02	0.07
GC	4.0	3.6	3.8	0	−3.7	0	0.29	0.05
MN	9.2	—	4.7	0	−3.6	0	0.18	0.06

*Volumes of conformation space within common energy contours given by the product of the eigenvalues of the stiffness matrix **F**.

†Parametric changes that contribute to $k_BT/2$ energy deformations along the longest principal axes obtained from the corresponding eigenvectors and eigenvalues. These deformations constitute the longest dimensions of the equi-potential surfaces.

‡Volumes of B–DNA samples in Table 1. Steps with <25 examples are not considered.

The pyrimidine–purines (YR) are the most easily deformed of all dimer steps, especially when highly distorted conformations are considered (Table 2, Fig. 2). This tendency is consistent with empirical energy predictions (29) and interpretations of gel mobilities (30). The heightened response of YR dimers to protein binding presumably is related to their relatively weak stacking interactions (31). Remarkably, the sharp bends in DNA observed in known protein–DNA crystal and NMR complexes occur almost exclusively at YR steps (10–12). These steps apparently act as flexible "hinges" fitting the duplex to the protein surface. The nonrandom positioning of YR dimers in protein binding sites also facilitates DNA loop formation (32), revealing how such steps may serve as long range signals for protein binding. Periodicity in the occurrence of flexible dimers (and trimers), if further confirmed, can be used to predict sites of protein binding in anonymous genomes.

The increased deformabilities of YR steps in protein–DNA complexes compared with free DNA (see Table 2) potentially can affect the entropy of complex formation. If the protein–DNA interface does not prevent local base pair movement, the placement of flexible YR "hinges" could be entropically advantageous. Although binding-induced enhancement of DNA mobility at first may seem counter-intuitive and contradictory to conventional "lock and key" mechanisms of ligand–biomolecule interactions and entropy/enthalpy compensation, the extra conformational space encompassed by some protein-bound steps, especially the distorted YR steps, is so profound that these dimers may reveal a real-time flexibility in selected protein–DNA complexes in solution. If so, the increased motions would add to the entropy of the complex along with other factors, such as solvation effects (33). On the other hand, crystal packing may reduce B–DNA variability compared with P·DNA deformations. The hypothesized entropic role of flexible YR hinges is, nevertheless, worthy of further inquiry and may be tied to published calorimetric (34) and NMR (35) properties of nucleoprotein complexes.

Translational Parameters: Rise, Shift, and Slide. Although now limited to the six base pair step parameters, these knowledge-based energy functions provide useful tools for understanding the nucleic acid machinery. The deformations along the principal axes of the potential surfaces reveal the natural response of DNA to external forces. Enzymatically "activated" states of DNA and physically stretched double helices presumably take advantage of these intrinsic properties (8). Because the variation in Rise is so restricted, proteins such as the catabolic activator protein (36) and TATA-box binding protein (24, 25) make use of the natural coupling of Twist and Roll with Slide and/or Shift to stretch DNA at selected base pair steps, particularly at the deformable YR steps noted above. The energy surfaces obtained here incorporate this sequence dependence (see Table 2; Figs. 2–4).

The above correlations reveal the complex interplay between the two different kinds of interactions in DNA: the sequence-dependent stacking of bases (which is primarily responsible for the directionality of dimer bending and sliding) and the largely sequence-independent interactions stabilizing the sugar–phosphate backbone conformation (which tends to retain its optimal conformation and thereby serves as an elastic string that controls the couplings between the three angular and three translational parameters). The specific ways in which these interactions reveal themselves at the local level obviously depend on the so-called "morphology" of bases and thus on the base sequence (2).

Among these correlations there are several that appear to be "almost universal" and thus applicable to various sequences. In addition to the Twist–Roll and Tilt–Shift correlations noted above, both the Roll–Slide and Twist–Slide couplings deserve mention. In seven of 10 dimers, the Roll–Slide force constants are positive, i.e., the correlation is negative. This is evident in Fig. 4, where the variations of Roll and Slide in the three YR dimers are presented. By contrast, the Twist–Slide constants are predominantly negative and the correlations are positive.

The interdependence between DNA bending and lateral displacements of base pairs is likely to have functional implications. When the DNA is bent in a tight loop, e.g., in a nucleosome or in the transcription initiation complex, such base pair displacement would regulate its superhelical handedness. Detailed geometric considerations (29, 37) show that, to facilitate left-handed superhelix formation, the DNA axis should be sheared such that Slide is positive when DNA bends into the minor groove and negative when the duplex bends into the major groove. In other words, Slide and Roll must be anti-correlated, exactly as observed here for the CA:TG and TA dimers (Fig. 4).

In addition to the thoroughly studied connection between DNA twisting and superhelical sense, both Slide and Shift regulate the DNA spatial trajectory. Because the variabilities of Roll and Slide are larger than those of Tilt and Shift, only the former are considered here. Displacements of the duplex axis via Shift also can be operative in the nonplanar packaging of DNA in chromatin.

As is well known (38), the penetration of a protein α-helix into the major groove induces a series of interdependent movements of DNA base pairs: dimer steps roll into the major groove, narrowing the groove and tightening the embrace of the α-helix; the duplex becomes slightly underwound; and base pairs translate (shift/slide) toward the minor groove, deepening the major groove and enhancing α-helix access. That is, protein-bound DNA takes advantage of the molecular mechanism involved in the B $\leftrightarrow$ A transformation—bending, unwinding, and displacing neighboring base pairs (21, 38, 39) through concerted changes in backbone and glycosyl torsions (40).

Polymer Applications. The current DNA structural database is still restricted, and our knowledge-based energies are subject to limitations of the available data (see above). The number of available structures limits queries to relatively simple questions. In particular, although protein-bound DNA structures are represented in greatest numbers, there are still not enough examples to analyze the nonharmonic, i.e., "bimodal," behavior of highly distorted dimer steps (Fig. 1). The precise numerical standards are likely to change as new and better resolution data accumulate, but the general trends in the present data are expected to persist. The knowledge-based patterns could thus prove useful in the refinement of future crystal and NMR structures of DNA, particularly if the analysis were extended to higher dimensions of conformation space, e.g., base pair parameters such as Buckle and Propeller Twist (16) and/or chain torsions. Our energy functions also should facilitate modeling the overall architecture of extremely large DNA–protein complexes for which only partial x-ray or NMR information is available, e.g., the global folding of DNA brought about by the multimeric binding of transcription

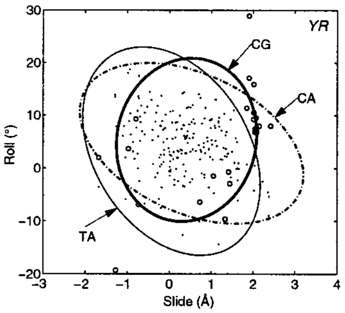

FIG. 4. Scatter plots in the Slide–Roll plane of observed parameters and derived energy contours of individual pyrimidine–purine (YR) dimer steps. Note the negative coupling of CA and TA parameters. See legend to Fig. 1.

factors, and *vice versa*, the allosteric changes in proteins caused by DNA binding (41). The "inverse harmonic analysis" introduced here can be used as well to evaluate the elastic force constants of base pair steps and to estimate the influence of environmental factors on dimer flexibility in DNA simulated at the all-atom level (42, 43).

The six base pair step parameters (16) are closely related to the representation of the double helix used in elastic models of DNA polymers (44). The bending, twisting, and stretching found in crystal structures are therefore potentially useful in simulations of long supercoiled chains. The present results, however, point to limitations in conventional elastic treatments of DNA based on an ideal homogenous isotropic rod. For one, DNA bending is anisotropic (20). Furthermore, fluctuations in base pair parameters are highly sequence-dependent and correlated, and the intrinsic equilibrium structure varies with sequence. In addition, stretching is taken up preferentially through shearing motions in the base pair plane, i.e., Slide and Shift, as opposed to changes in the axial direction, i.e., Rise (45). The incorporation of such "chemical" features may help account for discrepancies between observed properties of long DNA and predictions of the simple rod model (46).

Our simple energy functions are the first practical realization of a flexible dimeric model of the double helix (4, 8, 18, 42). The set of empirical expressions provides a way to monitor the global deformations of DNA in terms of realistic sequence-dependent conformational features. Typical static representations of DNA, by contrast, ignore the intrinsic deformability of individual base pair steps; see refs. 8 and 18 for further discussion. Energies derived from the protein–DNA crystal set, however, must be scaled to account for known properties of DNA in solution. We are currently modeling random sequence DNA in terms of the fluctuations of an average dimer based on all possible dimer steps, with the temperature adjusted to reproduce the persistence length (9) and cyclization tendencies (5). Later, as more data accumulate, we anticipate extending the present work to extract the energetic behavior of longer DNA fragments, such as trimers and tetramers. We expect that such higher order conformational data will enable us to account for the cooperative interactions that are thought to give rise to the observed curvature of DNA in solution (5).

Perspectives on Genomic Analysis. Finally, the sequence-dependent energy functions add a new perspective to traditional "literal" analyses of genomic sequences. The intrinsic structure and deformability of individual dimer steps could prove useful in detecting signals, such as sites of protein interaction (47), in anonymous DNA sequences and in positioning nucleosomes (48). The stage is set for using the new rules to understand the dynamic organization of the genome; that is, what are the sites of DNA where protein is more likely to bind, and how do long loops of several hundred base pairs rearrange in response to the "sliding" of nucleosomes during transcription (49) and/or replication? In other words, the deformability of DNA encoded in the base sequence determines whether widely separated parts of the long chain molecule come into close contact and whether one part of the DNA potentially can affect actions at other sites. The energies extracted here offer an important step toward understanding the mechanical properties of DNA and "fusing" this knowledge with the analysis of genetic sequences.

This work has benefited from valuable discussions with H. Berman, M. Clore, A. Gronenborn, N. Gō, V. Ivanov, R. Jernigan, S. Leikin, D. Rau, and N. Ulyanov. We are also grateful to A. Aggarwal, R. Brennan, S. Burley, K. Ely, K. Kamada, P. König, C. Pabo, T. Richmond, T. Steitz, and C. Wolberger for sharing crystallographic data in advance of public distribution and to A. Gelbin for assistance with NDB access. Support of this work through U.S. Public Health Service Grant GM20861 and National Cancer Institute Contract N10-CO-74102 is acknowledged gratefully. L.M.H. was supported by REU National Science Foundation Grant CCR 9119999.

1. Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R. & Schneider, B. (1992) *Biophys. J.* **63,** 751–759.
2. Gorin, A. A., Zhurkin, V. B. & Olson, W. K. (1995) *J. Mol. Biol.* **247,** 34–48.
3. Breslauer, K. J., Frank, R., Blöcker, H. & Marky, L. A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 3746–3750.
4. Hagerman, P. J. (1992) *Biochim. Biophy. Acta* **1131,** 125–132.
5. Crothers, D. M., Drak, J., Kahn, J. D. & Levene, S. D. (1992) *Methods Enzymol.* **212,** 3–29.
6. Trifonov, E. N. (1991) *Trends Biochem. Sci.* **16,** 467–470.
7. Dickerson, R. E., Goodsell, D. & Kopka, M. L. (1996) *J. Mol. Biol.* **256,** 108–125.
8. Olson, W. K. & Zhurkin, V. B. (1996) in *Biological Structure and Dynamics*, eds. Sarma, R. H. & Sarma, M. H. (Adenine, Schenectady, NY), Vol. 2, pp. 341–370.
9. Hagerman, P. J. (1988) *Annu. Rev. Biophys. Chem.* **17,** 265–286.
10. Suzuki, M. & Yagi, N. (1995) *Nucleic Acids Res.* **23,** 2083–2091.
11. Werner, M. H., Gronenborn, A. M. & Clore, G. M. (1996) *Science* **271,** 778–784.
12. Dickerson, R. E. (1998) *Nucleic Acids Res.* **26,** 1906–1926.
13. Drew, H. R. & Travers, A. A. (1985) *Nucleic Acids Res.* **13,** 4445–4467.
14. Otwinowski, Z., Schevitz, R. W., Zhang, R.-G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F. & Zigler, P. B. (1988) *Nature (London)* **335,** 321–329.
15. Holbrook, S. R., Dickerson, R. E. & Kim, S.-H. (1985) *Acta Crystallogr. B* **41,** 255–262.
16. Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. N., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H. C. M., Olson, W. K., *et al.* (1989) *J. Mol. Biol.* **208,** 787–791.
17. Gō, M. & Gō, N. (1976) *Biopolymers* **15,** 1119–1127.
18. Olson, W. K., Marky, N. L., Jernigan, R. L. & Zhurkin, V. B. (1993) *J. Mol. Biol.* **232,** 530–554.
19. Karplus, M. & Kushick, J. N. (1981) *Macromolecules* **14,** 325–332.
20. Zhurkin, V. B., Lysov, Y. P. & Ivanov, V. (1979) *Nucleic Acids Res.* **6,** 1081–1096.
21. Calladine, C. R. & Drew, H. R. (1984) *J. Mol. Biol.* **178,** 773–782.
22. Ramachandran, G. N., Ramakrishnan, C. R. & Sasisekharan, V. (1963) *J. Mol. Biol.* **7,** 95–99.
23. Miyazawa, S. & Jernigan, R. L. (1985) *Macromolecules* **18,** 534–552.
24. Kim, J. L., Nikolov, D. B. & Burley, S. K. (1993) *Nature (London)* **365,** 520–527.
25. Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. (1993) *Nature (London)* **365,** 512–520.
26. Simpson, R. T. & Kunzler, P. (1979) *Nucleic Acids Res.* **6,** 1387–1415.
27. Rhodes, D. (1979) *Nucleic Acids Res.* **6,** 1805–1816.
28. Puhl, H. L. & Behe, M. J. (1995) *J. Mol. Biol.* **245,** 559–567.
29. Ulyanov, N. B. & Zhurkin, V. B. (1984) *J. Biomol. Struct. Dynam.* **2,** 361–385.
30. Bolshoy, A., McNamara, P., Harrington, R. E. & Trifonov, E. N. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 2312–2316.
31. Ornstein, R. L. & Fresco, J. R. (1983) *Biopolymers* **22,** 1979–2000.
32. Barber, A. M. & Zhurkin, V. B. (1990) *J. Biomol. Struct. Dynam.* **8,** 213–232.
33. Spolar, R. S. & Record, M. T., Jr. (1994) *Science* **263,** 777–784.
34. Berger, C., Jelesarov, I. & Bosshard, H. R. (1996) *Biochemistry* **35,** 14984–14991.
35. Wisniowski, P., Karslake, C., Piotto, M., Spangler, B., Moulin, A.-C., Nikonowicz, E. P., Kaluarachchi, K. & Gorenstein, D. G. (1992) in *Structure and Function*: *Proteins*, eds. Sarma, R. H. & Sarma, M. H. (Adenine, Schenectady, NY), Vol. 2, pp. 17–54.
36. Schultz, S. C., Shields, G. C. & Steitz, T. A. (1991) *Science* **253,** 1001–1007.
37. Olson, W. K. (1979) *Biopolymers* **18,** 1235–1260.
38. Nekludova, L. & Pabo, C. O. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6948–6952.
39. Shakked, Z., Guzikevich-Guerstein, G., Frolow, F., Rabinovich, D., Joachimiak, A. & Sigler, P. B. (1994) *Nature (London)* **368,** 469–473.
40. Olson, W. K. (1976) *Biopolymers* **15,** 859–878.
41. Lefstin, J. A. & Yamamoto, K. R. (1998) *Nature (London)* **392,** 885–888.
42. Zhurkin, V. B., Ulyanov, N. B., Gorin, A. A. & Jernigan, R. L. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 7046–7050.
43. Cheatham, T. E., III, & Kollman, P. A. (1996) *J. Mol. Biol.* **259,** 434–444.
44. Olson, W. K. (1996) *Curr. Opin. Struct. Biol.* **6,** 242–256.
45. Lebrun, A., Shakked, Z. & Lavery, R. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 2993–2998.
46. Baumann, C. G., Smith, S. B., Bloomfield, V. A. & Bustamante, C. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 6175–6190.
47. Cardon, L. R. & Stormo, G. D. (1992) *J. Mol. Biol.* **223,** 159–170.
48. Widlund, H. R., Cao, H., Simonsson, S., Magnusson, E., Simonsson, T., Nielsen, P. E., Kahn, J. D., Crothers, D. M. & Kubista, M. (1997) *J. Mol. Biol.* **267,** 807–817.
49. Studitsky, V. M., Kassavetis, G. A., Geiduschek, E. P. & Felsenfeld, G. (1997) *Science* **278,** 960–1963.