doi:10.1006/jmbi.2001.4617 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **308**, 681–703

JMB

Dynamic Simulations of 13 TATA Variants Refine Kinetic Hypotheses of Sequence/Activity Relationships

Xiaoliang Qian, Daniel Strahs and Tamar Schlick*

Department of Chemistry and Courant Institute of Mathematical Sciences, New York University and the Howard Hughes Medical Institute, 251 Mercer Street New York NY 10012, USA

The fundamental relationship between DNA sequence/deformability and biological function has attracted numerous experimental and theoretical studies. A classic prototype system used for such studies in eukaryotes is the complex between the TATA element transcriptional regulator and the TATA-box binding protein (TBP). The recent crystallographic study by Burley and co-workers demonstrated the remarkable structural similarity contrasted to different transcriptional activity of 11 TBP/DNA complexes in which the DNAs differed by single base-pairs. By simulating these TATA variants and two other single base-pair variants that were not crystallizable, we uncover sequence-dependent structural, energetic, and flexibility properties that tailor TATA elements to TBP interactions, complementing many previous studies by refining kinetic hypotheses on sequence/activity correlations. The factors that combine to produce favorable elements for TBP activity include overall flexibility; minor groove widening, as well as roll, rise, and shift increases at the ends of the TATA element; untwisting within the TATA element accompanied by large roll at the TATA element ends; and relatively low maximal water densities around the DNA. These features accompany the severe deformation induced by the minor-groove binding protein, which kinks the TATA element at the ends and displaces local water molecules to form stabilizing hydrophobic contacts. Interestingly, the preferred bending direction itself is not a significant predictor of activity disposition, although certain variants (such as wild-type AdMLP, 5′-$TATA_4G$-3′, and inactive A29, 5′-TA_6G-3′) exhibit large preferred bends in directions consistent with their activity or inactivity (major groove and minor groove bends, respectively). These structural, flexibility, and hydration preferences, identified here and connected to a new crystallographic study of a larger group of DNA variants than reported to date, highlight the profound influence of single base-pair DNA variations on DNA motion. Our refined kinetic hypothesis suggests the functional implications of these motions in a kinetic model of TATA/TBP recognition, inviting further theoretical and experimental research.

© 2001 Academic Press

Keywords: TATA variants; TBP; transcriptional activity; sequence-dependent bending; flexibility

*Corresponding author

Introduction

The DNA/TATA-box binding protein (TBP) system is one of the most beautiful and important DNA/protein complexes known; the name TATA stems from the consensus octamer sequence of the DNA-binding site, TATA(t/a)A(t/a)⋆, where (t/a) indicates thymine or adenine, and ⋆ indicates any base[1] (see also http://www.epd.isb-sib.ch/promoter_elements/). As a member of transcrip-

Abbreviations used: AdMLP, adenovirus major late promoter; TBP, TATA-box binding protein; WB, AdMLP; WS, *S. cerevisiae cyc1* promoter; PC, principal component; PCA, principal component analysis.

E-mail address of the corresponding author: schlick@nyu.edu

0022-2836/01/040681–23 $35.00/0

© 2001 Academic Press

tion factor IID, TBP plays a central role in assembling the pre-initiation transcription complex in eukaryotes; see Burley & Roeder,[2] Figure 1, for the transcription complex assembly cycle. The structure of TBP with its DNA recognition site was solved in 1993 by two crystallographic teams.

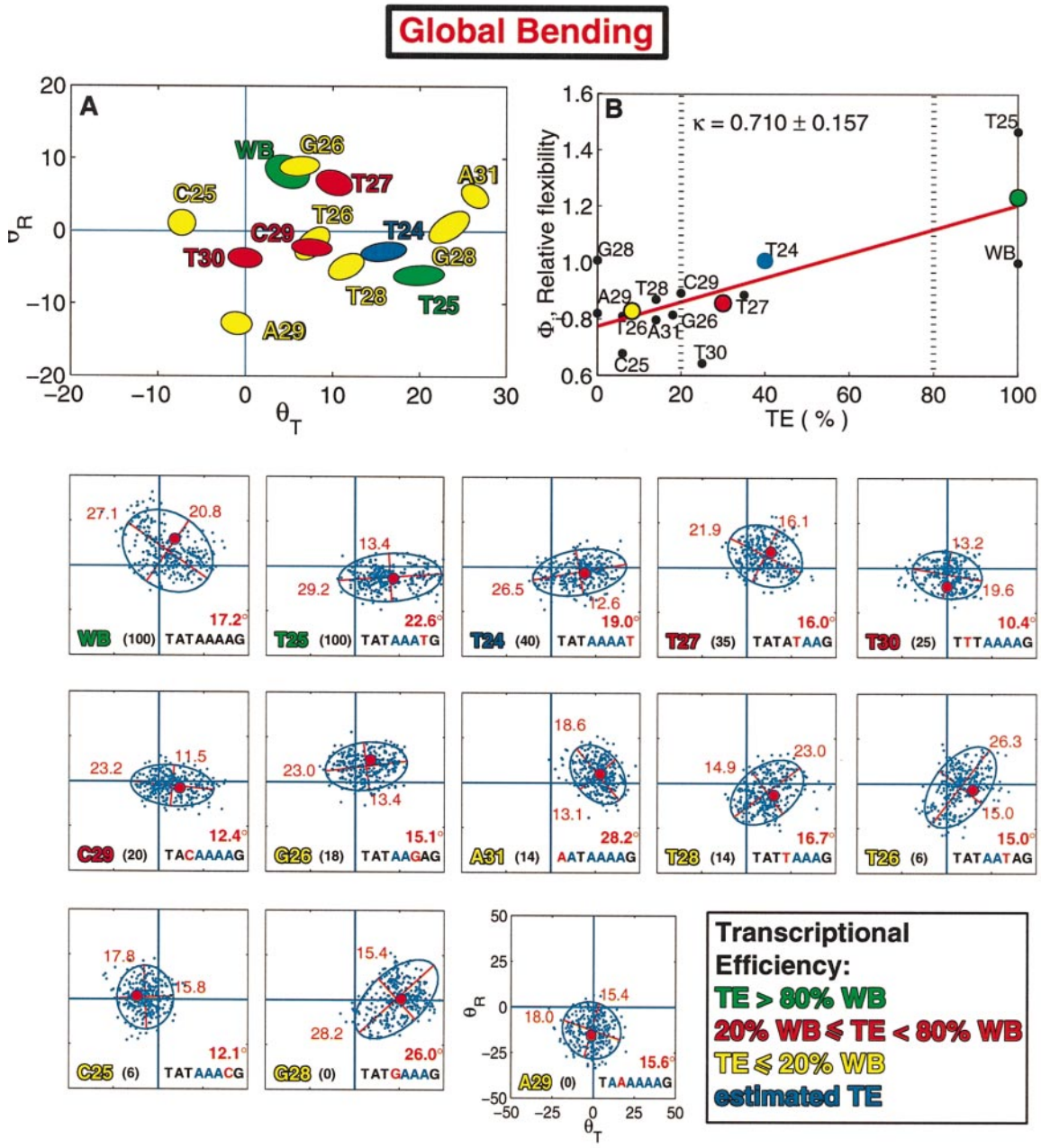

Figure 1. Bending propensities, relative flexibilities, and global tilt (θ_T) and global roll (θ_R) of 13 TATA variants over the production 1.8 ns period. (a) Global tilt and global roll of the 13 variants over the last 1.8 ns. Ellipses enclosing 90 % of the bending angles are drawn as described in Computational Methodology; the ellipse center is at the ensemble average $\langle\theta_T\rangle$, $\langle\theta_R\rangle$. The lengths of the major and minor ellipse axes are scaled to one-tenth size to show the relative positioning of all variants with minimal overlap; the full-scale ellipses are drawn for each variant below. The ellipses (and other figures) use a color-coding system to distinguish among variants with high TE (TE $\geqslant$ 80 % WB; green), medium TE (20% $\leqslant$ TE < 80%; red), low TE (TE $\leqslant$ 20 %; yellow), and estimated TE (for T24; blue). (b) Measured correlation between global bending flexibility of 13 TATA variants and TE. The flexibility Φ_i of sequence i is quantified from the global bending magnitude $\Theta_i = (\theta_T^2 + \theta_R^2)^{1/2}$ as the standard deviation of the global bending magnitude $\Theta_{\mathrm{sd},i} = (\langle\Theta_i^2 - \langle\Theta_i\rangle^2\rangle)^{1/2}$, normalized relative to the wild-type sequence WB: $\Phi_i = \Theta_{\mathrm{sd},i}/\Theta_{\mathrm{sd,WB}}$. The averages of measured property values for the three TE classes (low, medium, and high TE) are indicated by large circles. Linear least-squares best fits are indicated by red lines. Bottom panels: Global bending angles of 13 TATA variants. The lengths of the major and minor ellipse axes (α and β) are indicated in red next to the ellipses. The ensemble average bending magnitude $(\theta_T^2 + \theta_R^2)^{1/2}$ (large red dot) for each variant is indicated, along with with TE and sequence.

Although both TBPs and DNAs used in these studies are from different sources (yeast TBP/29-nucleotide yeast hairpin[3] and *Arabidopsis thaliana* TBP/14 base-pair adenovirus major late promoter (AdMLP)[4]), the structural similarity between these complexes is remarkable. This consensus structure now forms the basis for our understanding of transcription initiation in eukaryotes, and serves as a model system for appreciating the evolved partnership between proteins and DNA in regulatory processes.[5–7]

As experimental evidence has accumulated on the complex macromolecular transcriptional machinery in eukaryotes,[8] including TBP, its core TATA element promoter, and other transcription factors,[9] many theoretical studies have probed the relationship between DNA deformability and the molecular recognition/biological function of these complexes.[10–12] First, it is fundamentally appreciated that a unique coordination of sequence and structure compatibility has evolved between the DNA promoter and TBP (see Figure 4, center, for an illustration of the complex), as proposed in recent theoretical investigations (see Table 1). Specifically, the β-saddle shaped TBP protein, with a convex upper surface formed by α-helices, and a concave underside formed by an anti-parallel β-sheet, cradles the DNA through minor groove interactions and forms stabilizing hydrophobic contacts with the DNA along the complementary water-sparse surfaces. This close contact between the protein saddle with its framing ''stirrups'' (connecting loops) and the DNA results in a severely bent and distorted DNA. Second, the well-noted preference for AT base-pairs in TATA elements is a key aspect of minimizing the energetic cost of this deformation. Third, while AT base-pairs are preferred, many naturally occurring TATA elements exhibit sequence variations with respect to the consensus. Though tolerated in terms of binding to TBP, even single base-pair changes in TATA elements significantly affect the resulting transcriptional efficiency (TE) of the TBP/TATA complexes (Table 2).[13,14] Therefore, a better molecular-level understanding of DNA deformability/functional relationships will help analyze and extend these observations. Analyses of the equilibrium and dynamic properties exhibited by a larger ensemble of DNA variants than performed to date can help unify the findings reported to date (e.g. see Table 1).

A recent crystallographic investigation contrasted TBP co-crystallized with the AdMLP TATA element (5′-TATA·AAAG-3′, called WB here),

Table 1. Summary of TATA element theoretical studies

Study (Lab)	Methods	System	Result
1. Lebrun *et al.*[18] (Lavery)	EM/AM FLEX1 (in-house force-field)	WS and $(TA)_4$	Local stretching and unwinding of DNA leads to kinking deformations that mimic DNA distortion in complex
2. Flatters *et al.*[17] (Lavery/Beveridge)	MD/PME AMBER	WS	Large bending propensity observed in wild-type; *A*-DNA form observed during simulation suggested as possible intermediate to DNA in complex
3. Pastor *et al.*[25] 4. Pastor *et al.*[26] (Weinstein)	MD/PBC/PME CHARMM, AMBER	Poly(GC) and 6 TATAs: WB, T27, G26, $(TA)_4$, C31, T30/T28/T26	Consensus sequence of YRTATAYR suggested as a requirement for TBP binding/activity based on equilibrium, dynamic, and geometric local properties. Intrinsic sequence preferences correlate with observed deformations
5. Pardo *et al.*[19] (Weinstein)	MD/PME AMBER	WS	Forced alteration of only glycosyl angle transforms *A*-DNA to structure similar to DNA in complex
6. de Souza & Ornstein[23] (Ornstein)	MD/PME AMBER	4 TATAs: WB, T30, A29, T27	Intrinsic curvature and flexibility correlate with TBP binding activity. Curvature for all sequences is towards major groove (A29 too, though smaller in magnitude). This finding is in contrast to the present study
7. Flatters & Lavery[29] (Lavery)	MD/PME AMBER	WS-TA_7 variant	Bending of variant fluctuates significantly with respect to WS (Study 2 above); bending measured using extremal base-pairs only
8. Pastor *et al.*[51] (Weinstein)	MD/PBC CHARMM	WB and inosine variant (5′-TITIIIIG-3′)	Very different hydration and flexibility pattern noted in inosine variant (with respect to WB)

EM/AM, energy minimization/adiabatic mapping; MD, molecular dynamics; PME, particle mesh Ewald; PBC, periodic boundary conditions; QM, quantum mechanics; PMF, potential of mean force. The force-fields are FLEX1, AMBER, and CHARMM. The octamer sites (from −31 to −24) denoted by WB and WS are, respectively, AdMLP (5′-TATA·AAAG-3′) and *cyc1* (5′-TATA·TAAA-3′, crystallized by Sigler's laboratory.[3] Sequence notation follows Patikoglou *et al*. and refers to mutations relative to the respective WB or WS promoters. For example, A29 indicates the mutation in WB of T29 to A.

Table 2. Selected DNA sequences and their transcriptional efficiences[14]

Label		Sequence		Efficiency (%)
WB	GC	(−31)T A T A·A A A G(−24)	GGCA	100
A31	GCl	**A** A T A·A A A G	GGCA	14
T30	GC	T **T** T A·A A A G	GGCA	25
A29	GC	T A **A** A·A A A G	GGCA	1*
C29	GC	T A **C** A·A A A G	GGCA	20
G28	GC	T A T **G**·A A A G	GGCA	1
T28	GC	T A T **T**·A A A G	GGCA	14
T27	GC	T A T A·**T** A A G	GGCA	35
T26	GC	T A T A·A **T** A G	GGCA	6
G26	GC	T A T A·A **G** A G	GGCA	18
C25	GC	T A T A·A A **C** G	GGCA	6
T25	GC	T A T A·A A **T** G	GGCA	100
T24	GC	T A T A·A A A **T**	GGCA	40*

The TATA octamers are flanked by GC on the 5′-side and by GGCA on the 3′-side. The adenovirus 2 major late promoter (AdMLP) TATA element sequence serves as the control (or wild-type) sequence (WB). Single position variants (bold characters) are indicated relative to WB, and labeled according to the replaced base and position with respect to the transcription initiation site. Base complementarity is assumed for the opposite strand. Transcriptional efficiencies (TEs) for A29 and T24 (marked by asterisks) are based on Bernués *et al*.[48] and Wobbe & Struhl,[13] respectively. Additional TE data merging findings from three laboratories is available in Table 4.

against TBP complexes involving ten single base-pair DNA variants. It is intriguing that the high degree of structure conservation observed in the DNA/protein complexes did not translate into preservation in functionality.[14] Namely, TE values ranged from high (100 % with respect to WB) for T25 (TATA·AA**T**G) to moderate (around 40 %) for T24 (TATA·AAA**T**)† and T27 (TATA·**T**AAG), to very low (6%) for C25 (TATA·AA**C**G) and T26 (TATA·A**T**AG); see Table 2 for data and nomenclature. In terms of structure, A to T and T to A substitutions were found to be generally well tolerated, while substitutions from A or T to G or C were accommodated in some cases by rearrangements of specific interactions and alternative Hoogsteen base-pairing.[14] This crystallographic evidence suggested to Burley and co-workers that TATA/TBP recognition (and corresponding transcriptional activities) occur through mutually complementary motions expressed by sequence-dependent dynamics of each TATA variant.

The unusual distortion of DNA in the TBP-bound complex has attracted many theoretical and experimental investigators. Studies have explored the transition involved in deforming the DNA to the complex structure,[15–22] and the DNA sequence complementarity to TBP binding and function[23–26]‡ (see Table 1 for a summary table of theoretical studies). In particular, work has shown the importance of base-pair step flexibility at the phenylalanine intercalation positions −31 and −30 (step TA in WB),[20] bending motions in the complex,[27] and the role of solvation and internal DNA electrostatics in directing and stabilizing the complex deformations.[16]

Here, we report 13 nanosecond MD simulations on the 11 single base-pair variants of the TATA element co-crystallized with TBP[14] (WB plus ten variants), and on two TATA elements that resisted crystallization attempts (G28 and A29)§. Our analysis aims to dissect systematic thermodynamic and kinetic differences that are likely to affect TBP/DNA binding and, ultimately, the disparate activity values associated with these variants (Table 2); simulations of DNA/protein complexes represent the next step of our study and are underway. It is unfortunate that work to date cannot be unified easily, since it reflects different force-fields, various simulation protocols, as well as approximate functionality measurements, because studies were conducted prior to the Patikoglou *et al*. experimental work.[14] The different conclusions reached regarding bending direction and magnitude of TATA variants[17,23,24] are also sensitive to the pro-

† This TE estimate is based on the sequence TATA·TAAT (T24/T27 in our notation), with TE measured against a wild-type *his3* control.[13] The TE values using a wild-type *his3* control are directly comparable to a wild-type AdMLP control (see Table 4), indicating that a scaling factor of 0.4 may be applied to normalize the wild-type *his3* control to WB, as the authors suggest.

‡ Observations by Pastor *et al*.[25,26] were based on six TATA variants, of which T27, G26, and WB are also studied here.

§ G30 and C30 also resisted crystallization attempts; steric clashes with Leu163 explain this result.[28] A steric clash is presumed between Val119 and G28.[14]

cedures used to measure the global helical curvature; we attempt to reconcile these observations here by using our global bending analysis[29] that accounts for each base-pair step, not only extremal steps.[17,23,24,30,31] In addition, conclusions on general relationships between DNA sequence deformability and activity can appear oversimplified when reached from studies of a small group of variants; we show this here by the variability of bending preferences exhibited by high and low transcriptionally active TATA variants.

Our analysis delineates, over a large group of DNAs, a combination of equilibrium and dynamic factors that produces favorable elements for TBP activity in high TE variants. These factors include overall flexibility, increased roll and minor groove widths at the end base-pair steps, untwisting within the TATA element coupled with roll at the end base-pair steps, low maximal water densities close to the DNA, and an ''optimal'' ion density. The flexibility, local motions, and ionic environment trends facilitate the large-scale distortion of the DNA and modulate phenylalanine intercalation at TATA ends; the low maximal water densities facilitate TBP binding by lowering the solvation energy loss upon complexation. These general trends, though not applicable to all sequence variations, reinforce and extend the many works to date (Table 1) and provide further insights into the kinetic behavior of optimal TATA elements. Fundamentally, the results reinforce the common view that subtle, sequence-dependent DNA information and motions can direct protein binding and activity. However, our systematic results refine the notion[14] that sequence-dependent motions underly transcriptional differences between different TATA variant/TBP complexes by showing which intrinsic properties are more strongly correlated to activity than others, and which combinations are important. Quite ingeniously, single nucleotide changes can alter flexibility, as shown by Olson and co-workers;[32] local solvation patterns, as shown by Timsit and colleagues;[33] global solvation patterns within protein-DNA binding sites, as observed by Berman and colleagues;[34] and ionic patterns, as observed by Hud *et al*.[35]

Results

Models and overall analyses

Thirteen simulations were performed on 14-bp DNA duplexes, as shown in Table 2, with the AMBER PARM94 force-field[36] including water and ions. Setup, equilibration, force-field, and integrator details are described under Computational Methodology.

For structure analysis, we use the Curves program,[37,38] supplemented by our global bending analysis program Madbend,[29] and principal component analysis (PCA)[39–41] developed for this study.

The difficulty of quantifying bending in highly deformed DNA is widely appreciated,[42] given the sensitivity to local definition of base-pair parameters.[43,44] Various bending frameworks have thus been proposed; see Zhurkin,[30,31] Trifonov,[45] and Lavery,[24] for example. Our global bending description (program Madbend, http://monod.biomath.nyu.edu/, click on Software) extends procedures reported to date, essentially by summing accumulated roll and tilt projections onto a reference plane after adjusting for helical twist.[29]

PCA is a widely used tool for motion interpretation[39–41] that describes trajectory fluctuations by independent modes. The motions are hierarchically organized so that the first several modes describe most of the motion characteristics of the trajectory. Recently, Laughton and co-workers described the dominant global bending motions in A-tract DNA.[41] Here, in addition to PCA for analyzing individual trajectories, we develop an ''ensemble PCA'' protocol to analyze the merged trajectory of all 13 TATA element variants. These two PCA procedures are complementary: the individual ensemble PCA (performed separately for each variant trajectory) highlights the prominent motions of each variant (Figure 3); the uniform ensemble PCA rigorously identifies common motions and relates the significance of such motions among variants (Figures 2 and 7). Details can be found in Computational Methodology.

Global DNA bending

Figure 1 shows the relative (top) and entire (bottom) bending range of all variants in the framework of global tilt and global roll $\{\theta_T, \theta_R\}$, as described in Computational Methodology.[29] Our two high TE variants, WB and T25, are very flexible: WB bends towards the major groove (average bend of 17°), and T25 bends towards the backbone (23° bend). In contrast, the less flexible A-tract variant (A29) bends towards the minor groove (negative θ_R, 16° bend), in good agreement with previous A-tract simulations;[29,46,47] differences from other TATA element simulations[17,23,24] likely arise from the different bending analyses used in these works (bending angle measured between extremal base-pair step vectors). Though A29 has resisted crystallization attempts with TBP,[14] it forms transcriptionally inactive complexes with TBP *in vitro*.[48]

The relevance of global bending to the activity of TATA elements can be estimated from our uniform ensemble PCA. The two dominant independent motions (PC 1 and 2, Figure 2), capture 40% of the overall motion and describe global bending motions along the groove and backbone directions, respectively. Though certain variants display bend directions correlated with promoter activity, as indicated by previous studies of TATA elements

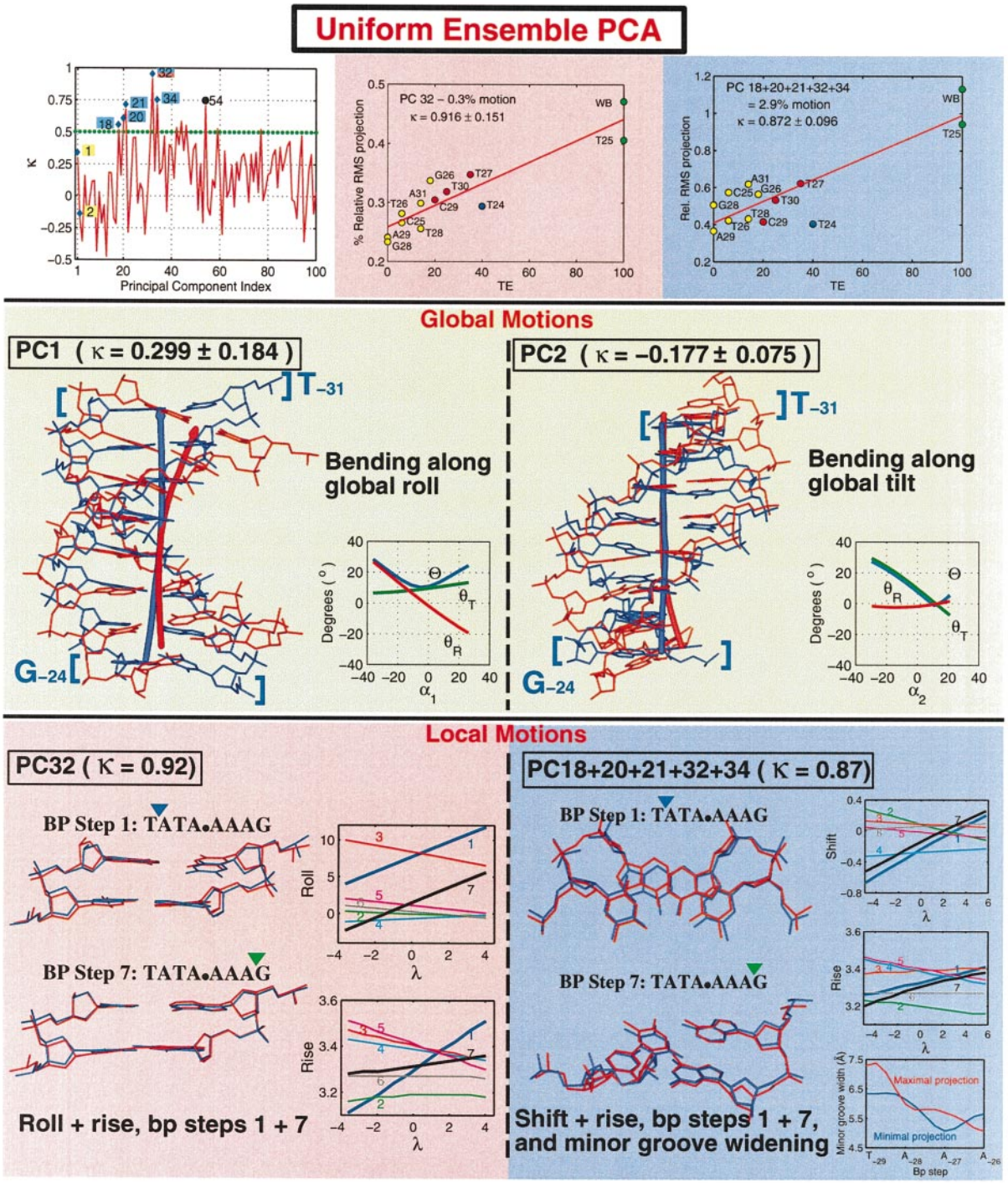

Figure 2. Analysis of the top 100 uniform ensemble PCs and associated motions. Top left: PCs and correlation coefficients. Top center and right: The correlation between each TATA variant's relative magnitude of motion along PC 32 (center) and combined PCs 18, 20, 21, 32, and 34 (right) with TE. The relative magnitude of motion $\alpha^2_{j,i}$ for each variant i in PC j (or combined PCs) is calculated as the mean-square projection, normalized by the trace (e.g. equation (2)). The linear least-squares fit of the relative magnitude of motion to TE is indicated by the red line. Center: Global motions in PCs 1 and 2. Images were generated by deforming the average structure of the merged ensemble of all 13 variants along the first or second PC. Values of $\{\theta_T,\theta_R\}$ and bending magnitude Θ corresponding to the full range of motion along these two PCs are indicated. The blue and red structures correspond to the minimal and maximal deformations of each PC, respectively (see Computational Methodology). The global bending angles were calculated from ten structures evenly spaced between the minimal and maximal projections. Bottom: The motions of PC 32 and combined PCs 18, 20, 21, 32, and 34 at base-pair steps 1 and 7 (left), and associated roll, rise, shift and major groove widening motions (right). Blue and red colors are used for the the minimal and maximal motions for base-pair steps 1 and 7, as viewed from the minor groove (PC 32) and from top and minor groove views (combined PCs 18, 20, 21, 32, and 34).

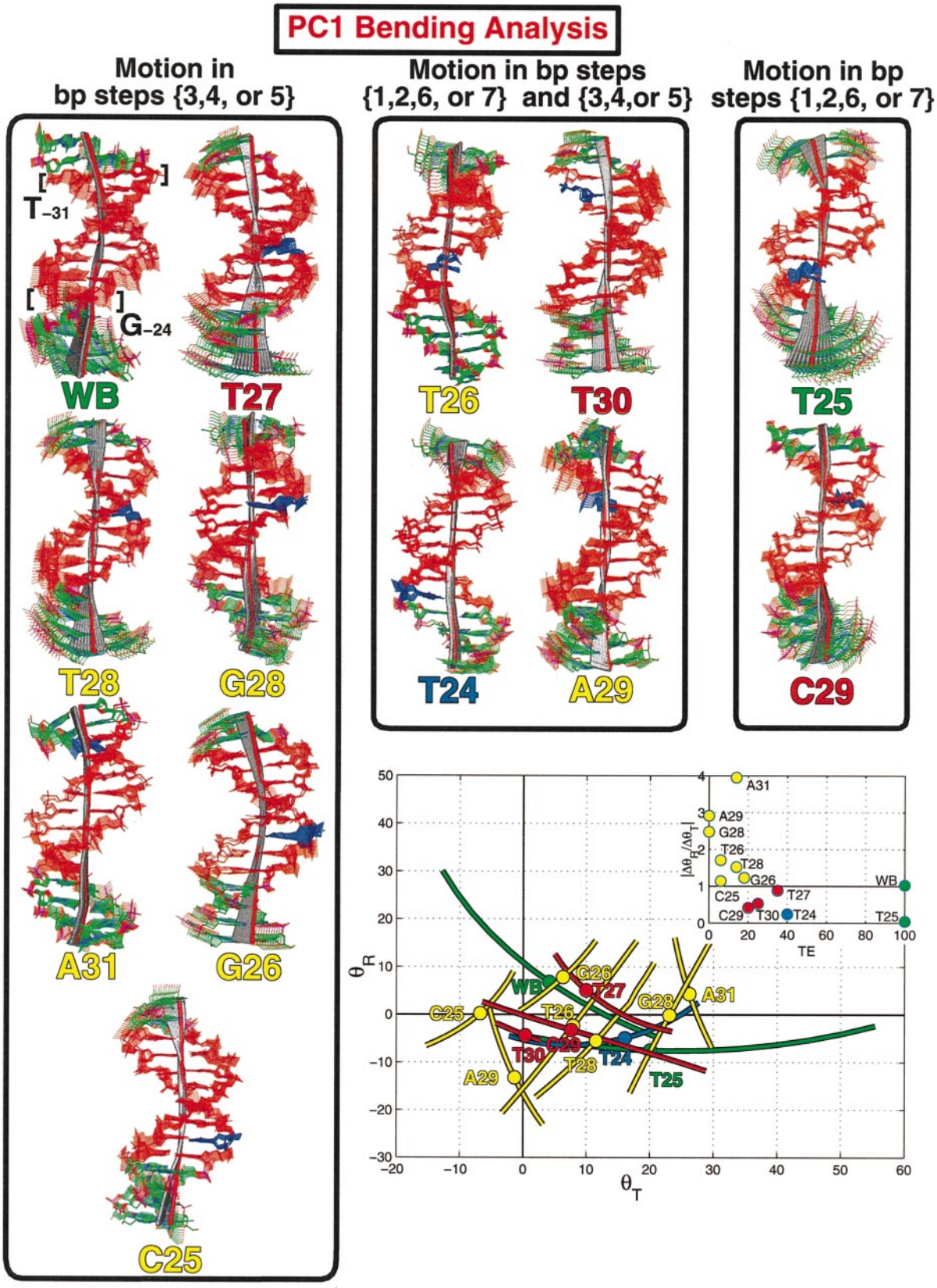

Figure 3. The motion of PC 1 from the individual PCA of 13 TATA variants and associated global bending motions. The structures generated by deforming the average structure of each variant's ensemble along the first PC are shown with characterizations of the base-pair step position of significant local motions. The TATA element is indicated in red; base-pairs changed in each variant are in blue. The bottom plot indicates the motion along global tilt (θ_T) and global roll (θ_R) in PC 1, with the bending of the average structure indicated by a large circle; the corresponding ratio $|\Delta\theta_R/\Delta\theta_T|$ is plotted in the inset.

Table 3. Motion description for the top ten PCs of the merged ensemble of 13 TATA trajectories

PC (%)	Motion	κ
1 (28)	Global rolling	0.30
2 (11.8)	Global tilting	−0.18
3 (9.5)	Wedge bending near steps 3 and 4	−0.35
4 (6.4)	Shifting at steps 2-4, sliding at steps 1, 4, 7	0.16
5 (4.6)	Sliding and tilting at steps 3 and 4	−0.17
6 (3.8)	Shifting at steps 3 and 7, sliding at step 7	−0.04
7 (2.8)	Tilting at steps 3 and 4	0.11
8 (2.7)	Shifting at steps 1 and 4, sliding at step 1	0.12
9 (2.5)	Rolling at step 1, twisting at steps 1 and 4, and shifting at steps 1-4	−0.43
10 (2.0)	Shifting at steps 1 and 2	0.03

Motions were identified by visualizing animations of the PCs and by calculating the changes to helical axis parameters as computed by Curves.[37] The relative percentage of the motion of the merged ensemble described by each PC is indicated in parentheses, as well as the correlation coefficient κ of the mean-square projection with TE for all 13 variants (equation (1)).

from the Lavery and Ornstein groups,[17,23,24] the correlation coefficient κ for PCs 1 and 2 with TE is weak ($\kappa = 0.3 \pm 0.18$ and $\kappa = -0.18 \pm 0.08$, respectively)†. This is apparent from the high TE variant T25, which does not display strong major-groove bending, and the low TE variant G26 (TE of 18%), which has preferential major-groove bending. This result underscores the importance of examining a large group of TATA variants under the same simulation protocol; selected subsets (such as the four variants WB, T27, T30, and A29 used by de Souza & Ornstein[23]) might suggest incorrect correlations between intrinsic bending and TE.

Still, we find that the overall flexibility (as measured by the standard deviation of the ensemble global bending magnitude relative to that of WB), rather than the bending direction *per se*, is highly correlated with activity ($\kappa = 0.71 \pm 0.16$): flexible sequences tend to have high TE values, as reported by de Souza & Ornstein,[23] while inactive variants are relatively stiff (Figure 1, top right). WB and T25, the high TE variants, are among the most flexible TATA sequences; low TE variants, such as A29, C25 and T26, are much less flexible. The overall fit is not perfect, indicating that other factors associated with bending combine to produce favorable propensities for TBP activity.

Local deformations

Such additional factors emerge in Figure 3 from our examination of the local deformations associated with the first PC (as deduced from the individual variant PCA); note that PC 1 is uniquely characterized by the bending axes ratio $\Delta\theta_R/\Delta\theta_T$. Based on visual and geometric analyses of PC 1, we define three deformation categories for global bending: large local motions such as twist, roll, and shift at the central TATA base-pair steps‡ 3, 4, or 5 (describing variants WB, T27, T28, G28, G26, C25, and A31); large motions at the 5′ end and/or 3′ end TATA base-pair steps 1, 2, 6, or 7 (T25 and C29); and large local motions within both central and end TATA regions (T30, T26, T24, and A29). Interestingly, low TE variants have $\Delta\theta_R/\Delta\theta_T$ ratios greater than 1 (average ≈1.9), generally indicating rolling at central base-pair steps (since θ_T and θ_R are measured with respect to the center of the TATA element; see Computational Methodology); medium and high TE variants have ratios below 1 (average ≈0.5, WB has a ratio of 1.02), generally indicating rolling at the end base-pair steps. This trend emphasizes the fact that active variants possess frequent, large motions at the TATA element ends. Indeed, Pastor & Weinstein have suggested that the flexibility of TA base-pair steps tailors TATA elements to typical TBP deformations[26] as defined by Suzuki *et al*.[49] based on several TBP/TATA complexes.

The significance of such local bend motions at the end base-pair steps is confirmed by our uniform ensemble PCA. We find that local motions, though accounting for a smaller portion of the overall motion amplitude, are much more strongly correlated with TE ($\kappa > 0.6$) than the global modes (the ten topmost PCs describe 74 % of the overall motion but have κ values ranging only from −0.43 to 0.30); see Table 3 and Figure 2. The highest correlation among the top 100 PCs ($\kappa \approx 0.91 \pm 0.15$) occurs for PC 32 (accounting for only 0.3 % of overall motion), visualized in Figure 2. It is significant that PC 32 acts locally near base-pair steps 1 and 7 of the TATA element, by increasing roll and rise (Figure 2), as well as by modestly increasing negative tilt and undertwisting at base-pair step 1 (data not shown).

In addition to PC 32, other PCs that have relatively large κ values (PCs 18, 20, 21, and 34) are associated with increased shift and rise at base-pair steps 1 and 7, and increased roll at base-pair step 7. The combined projection of these five PCs illustrated in Figure 2 (combined $\kappa = 0.87 \pm 0.1$) shows a widening of the minor groove by ≈1 Å. The relevance of these local PCs to transcription activity underscores the importance of the motion at the end base-pair steps, possibly through the development of the specific local deformations observed in the TBP/TATA complexes.[49]

The importance of rolling at the TATA element end base-pair steps is further demonstrated by our enumeration of the percentage of trajectory conformations that are underwound (less than 29° at each of the seven interior base-pair steps of the TATA element) and possess high roll (>10° at steps 1 and 7), as suggested by Lavery and coworkers.[17,24] Such structures may be preliminary substrates for TBP recognition, since DNA in the

† Error bars associated with κ are indicated in each Figure; the largest error for all correlations reported here is ±0.22.

‡ Base-pair steps are numbered from 1 through 7, where base-pair steps 1 and 7 correspond to T_{-31}/A_{-30} and A_{-25}/G_{-24}, respectively.

complex is underwound by less than 21° on average and sharply kinked more than 40° at the end base-pair steps;[2] higher probabilities of these specific deformations are likely to be correlated directly to the strength of TBP binding affinity. Figure 4, bottom, shows that variants that have the highest probability of adopting structures satisfying the bent, underwound probability (2% of snapshots or roughly three times per nanosecond) are T25 and WB, both with maximal TE. Though T26 (TE of <10%) has a probability comparable to WB, we find that the overall probability for the 13 sequences of adopting underwound, bent structures increases with TE ($\kappa = 0.71$).

Minor groove widths

The local motions discussed above at the TATA ends produce a widened minor groove. TBP binding widens the minor groove (to ≈12 Å), producing a shallow surface that interacts easily with the concave undersaddle of TBP.[3,4,50] It is intriguing that the minor groove width has been linked to TATA element activity in several theoretical studies,[17,23,24,51] although the significance has not been estimated.

Our measurements for the 13 TATA variants, as shown in Figure 5, indicate that many variants have widths of ≈6 to 7 Å along the 5′ and 3′ ends (near base-pairs −31/−30 and −25/−24), narrowing to less than 5.5 Å in the central region (base-pairs −29 to −26). The narrowing near the central adenine-rich segment of the TATA element agrees with prior observations on TATA elements,[24] as well as with other adenine-rich sequences;[29,52] the widening at the ends resembles structures in the DNA/TBP complexes.[2] Several variants deviate from this minor groove width pattern: the minor groove for C25 and T27 is approximately constant in width across the TATA element region, and the A-tract variant, A29, possesses an exceptionally narrow minor groove that widens at the 3′ end of the TATA element (Figure 5).

Analysis of the combined 5′-end width/3′-end width data in Figure 5 (top left) shows that the high TE variants WT and T25 have the widest minor grooves at both ends, while low TE variants (e.g. A29, C25, and T28) have narrow widths at both ends ($\kappa = 0.80 \pm 0.22$; Figure 5, top right). We can establish the significance of these minor groove widths at the 5′-end (mostly TA and AT base-pair steps) but not at the 3′-end (AA and AG steps) through a comparison to high-resolution crystallographic data from the NDB[53] (Figure 5, bottom right). This asymmetric widening at the 5′-end was observed in an active *cyc1* promoter variant (WS) by Lavery and co-workers,[17] but not in an A-tract variant.[24] The 5′-half of the TATA element is indeed recognized to be the more important half,[1,50] as it may control (with TFIIA/TFIIB) the directional assembly of the pre-initiation complex.[54]

Water interactions

Since the TBP/TATA complex desolvates a large region (approximately 3150 Å^2) during complex formation,[3,4,55] including many hydrophobic groups of TBP, the energy of TBP must overcome DNA's tendency to efficiently order nearby water molecules in the minor groove.[56] The intrinsic hydration patterns of TATA variants are likely to be a factor affecting the TBP/TATA element interaction; for example, Pastor *et al*. observed greater water coordination around a low TE inosine variant relative to WB.[51] The importance of water structure around protein-DNA-binding sites was recently re-examined by Berman and colleagues, who observed water molecules at free DNA positions corresponding to the protein-binding residues in two CAP-DNA complexes.[34]

Our analysis in Figure 4 (top) of the local water structure around TATA elements (i.e. number of water molecules per volume element at a 1 Å^3 volume resolution) shows that A29 and other low TE variants, such as T28, have significant local water structure near the DNA, whereas medium and high TE variants (such as WB and T27) have a much lower local density near the DNA, with the water localized near phosphate groups. The high local density of water for A29 (0.257 molecules per Å^3) is expected; A-tracts are known to stabilize a long-lived minor groove water spine.[29,57]

Our estimates of enthalpic solvation energies of each variant by two methods in Figure 6 show that solvation energies do not follow the trend identified in the measurements of maximal water density; instead, the solvation energies are very similar overall. Though the respective error bars are relatively large (data not shown), we suggest that entropic, rather than enthalpic, terms associated with the flexibility and inherent DNA motions may be responsible for the clear differences observed in our variants' local water density.

Ion atmosphere

The large-scale bending motions towards the major groove commensurate with TBP activity, as we observe in WB, may be facilitated by proper cationic shielding. This general principle was demonstrated by the classic studies by Mirzabekov, Rich, and Manning on the role of asymmetrically neutralized phosphate groups in promoting bending around nucleosomes[62,63] and recently highlighted by Maher and co-workers, who examined bending promoted by neutral phosphate analogs,[64] and by Williams, Dickerson, and co-workers, who have examined ion distribution and binding in crystallographic nucleic acid structures.[65,66] In the case of TATA/TBP complexes, this shielding may be particularly important, because the large bending into the major groove at each TATA end introduces an electronegative pocket near the DNA (Figure 4).

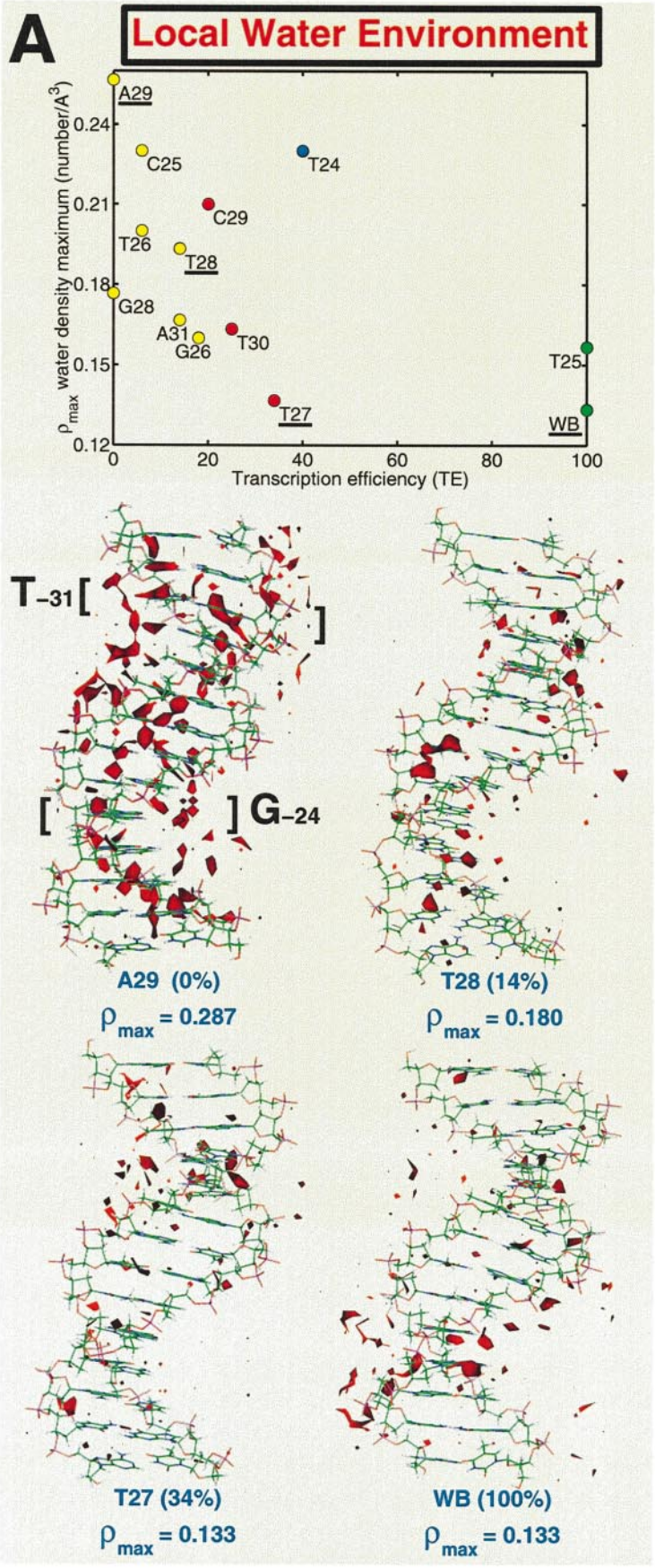

Figure 4. (*legend shown opposite*)

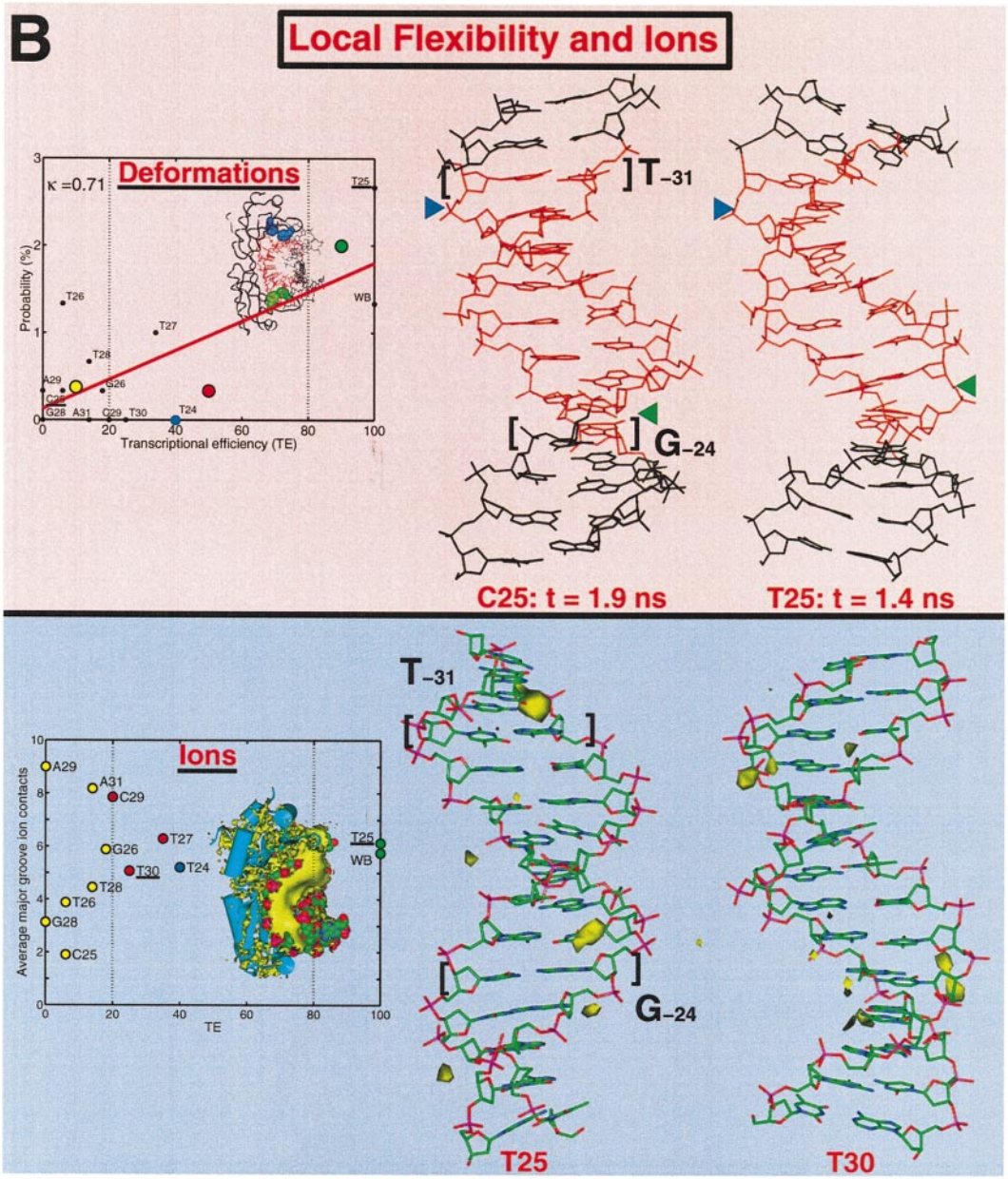

Figure 4. Various factors correlated with TE: solvation, ion atmosphere, and local deformations. (a) Top: Maximal water density for 13 TATA variants plotted against TE. The density for all water oxygen atoms (and ions in (b)) was accumulated on a 1 Å cubic lattice over 300 snapshots sampled at a frequency of 6 ps. (a) Bottom: illustration of the water oxygen density for the four TATA variants A29, T28, T27, and WB drawn at a contour density of 0.075 molecules per Å^3. (b) Top: the probability of adopting bent, underwound structures is calculated for 300 structures sampled over 1.8 ns at a frequency of 6 ps. Specifically, we count snapshots with cumulative twist less than 205° for the seven base-pair steps between, and including, −31/−30 and −25/−24; and roll angles greater than 10° at steps −31/−30 and −25/−24. The averages of the probabilities for the three TE classes are indicated (large circles). The TATA/WB complex (PDB code 1CDW) is illustrated in the center with the phenylalanine residues intercalated at the 5′ and 3′-ends indicated with blue and green CPK models, respectively. Structural examples of C25 and T25 are shown at the right side of the Figure with blue and green triangles indicating the 5′ and 3′-ends, respectively, of the TATA element (red). (b) Bottom: the average number of ion contacts per snapshot within a 5.5 Å radius of major groove base atoms is plotted against TE based on 300 snapshots spanning 1.8 ns. The electrostatic potential of the TBP/WB complex (as computed by Delphi[58]), illustrated in yellow at a contour of −5 k_BT/e, hides the WB TATA element (CPK model); secondary structural elements of TBP are illustrated in blue. The potential within the TBP atoms has been truncated for clarity. Note that the large roll at the TATA ends creates increased electrostatic potential. The right side of the Figure illustrates the ion density with a yellow contour at 0.02 molecule per Å^3 relative to the average structures of T25 and T30.

Our analyses of sodium ion contacts in Figure 4 in the TATA element major groove (enumerated within a cutoff radius of 5.5 Å) suggest that high TE sequences develop an optimal number of cation contacts during the simulations, differing from the network associated with medium and low TE variants. Furthermore, the attraction of ions is mildly asymmetric in high TE variants: we note a higher density of ions within a 7 Å sphere centered on major groove atoms for variants with larger bends at either the 5′-end base-pair steps (T_{-31}/A_{-30} and A_{-30}/T_{-29}) or the 3′ end (A_{-26}/A_{-25} and A_{-25}/G_{-24}) ($\kappa = 0.36$; data not shown).

Discussion

The intrinsic sequence-dependent properties of TATA variants are critical factors that modulate the interaction between TBP and its recognition element. The co-crystallization of TBP with ten different TATA variants[14] demonstrated extraordinary structural similarity in different complexes, despite sequence and activity differences. Motivated by this important experimental study, we have identified discerning dynamic, structural, and flexibility properties for a large collection of single base-pair TATA variants that underscore DNA's structural complementarity to TBP binding and deformation.

Though the TATA DNA in complexes with TBP exists in the novel, distorted structural form termed TA-DNA,[15] the substrate DNA that TBP initially recognizes is around the equilibrium *B*-DNA form, which our simulations have explored. Our ongoing DNA/protein simulations have been testing the reported properties of the unusual DNA forms seen in the complexes, as described below.

The form of the DNA in the complex is certainly unusual and relevant to understanding the properties of the transcription complex; however, TBP initiates complex formation by binding to TATA elements in their equilibrium *B*-DNA form. Therefore, our simulations are designed to explore the properties potentially regulating the initial association between TBP and DNA.

Bending preferences and overall flexibility

Our average bending analysis of TATA variants supports the hypothesis[67,68] that strong preferential bending may alternately promote or inhibit TBP binding. Bending towards the major groove, as noted for WB, accompanies TBP's binding and deformation, while a strong bending preference towards the minor groove, noted for the A-tract TATA variant, may inhibit complexation (Figure 1(a)). However, overall flexibility, rather than bending direction *per se*, is correlated to activity, in partial agreement with de Souza & Ornstein:[23] high-TE sequences (such as WB and T25) are significantly more flexible overall than the others. These bending preferences and flexibilities suggest alternately low and high deformational energy barriers involved in forcing bending towards the major groove. It is intriguing that early TBP/TATA intermediates form rapidly,[69] possibly with highly bent DNA.[70,71] This suggests that the bending and flexibility of TATA variants may be an important factor in the early association between TBP and TATA elements and can significantly accelerate complexation.[72]

Groove widening and local motions

The nature of local structures and bending motions, as identified by PCA, appears particularly important for TBP activation. Features like minor groove widening (Figure 5) at the phenylalanine intercalation sites, as first suggested by Lavery & Ornstein,[23,24] can initially facilitate TBP/TATA interactions, which in turn may reinforce the opened minor groove.[73] Our analysis of the first PC (Figure 3) also shows that bending motions at the 5′ and 3′-ends of the TATA element increase with activity of the TATA variant. The significance of localized, preferential motions also emerges from the PCA and trajectory analysis. Notable deformations in high TE variants include positive roll, shift, and untwisting motions (Figures 2 and 4). These preferential deformations are significant, since they predispose the conversion of TATA elements to the distorted form observed in the complexes.

Though the deformations required to form the TBP/TATA complex are much larger than those we observe in the TBP-free TATA elements, little motion is required of TBP as deduced from the similarity in crystal structures between free and DNA-bound TBP (C^{α} rmsd $\approx$0.5 Å).[2,14,74]

Water interactions and complex formation

The remarkable sensitivity we noted concerning the local hydration atmosphere to the DNA sequence (Figure 4), is compatible with observations that the minor groove is completely desolvated by TBP.[3,4,55] The inverse relationship we observe between the maximal water density and TE (Figure 4) may be correlated to the DNA flexibility; flexible variants have a more disordered water environment, while more rigid sequences tend to order water molecules (as in A-tracts). The increased flexibility may in turn inhibit water molecules near the DNA from maintaining steady spatial positions. Our finding of minimal enthalpic solvation energy differences among our variants suggests an entropic origin for the hydration patterns, with the larger motions observed in high TE variants leading to increased disorder. This hypothesis is supported by recent observations from an MD simulation of the TBP/WB complex interpreted in tandem with hydroxyl radical footprinting;[22] it was concluded that TBP uses enthalpic and entropic

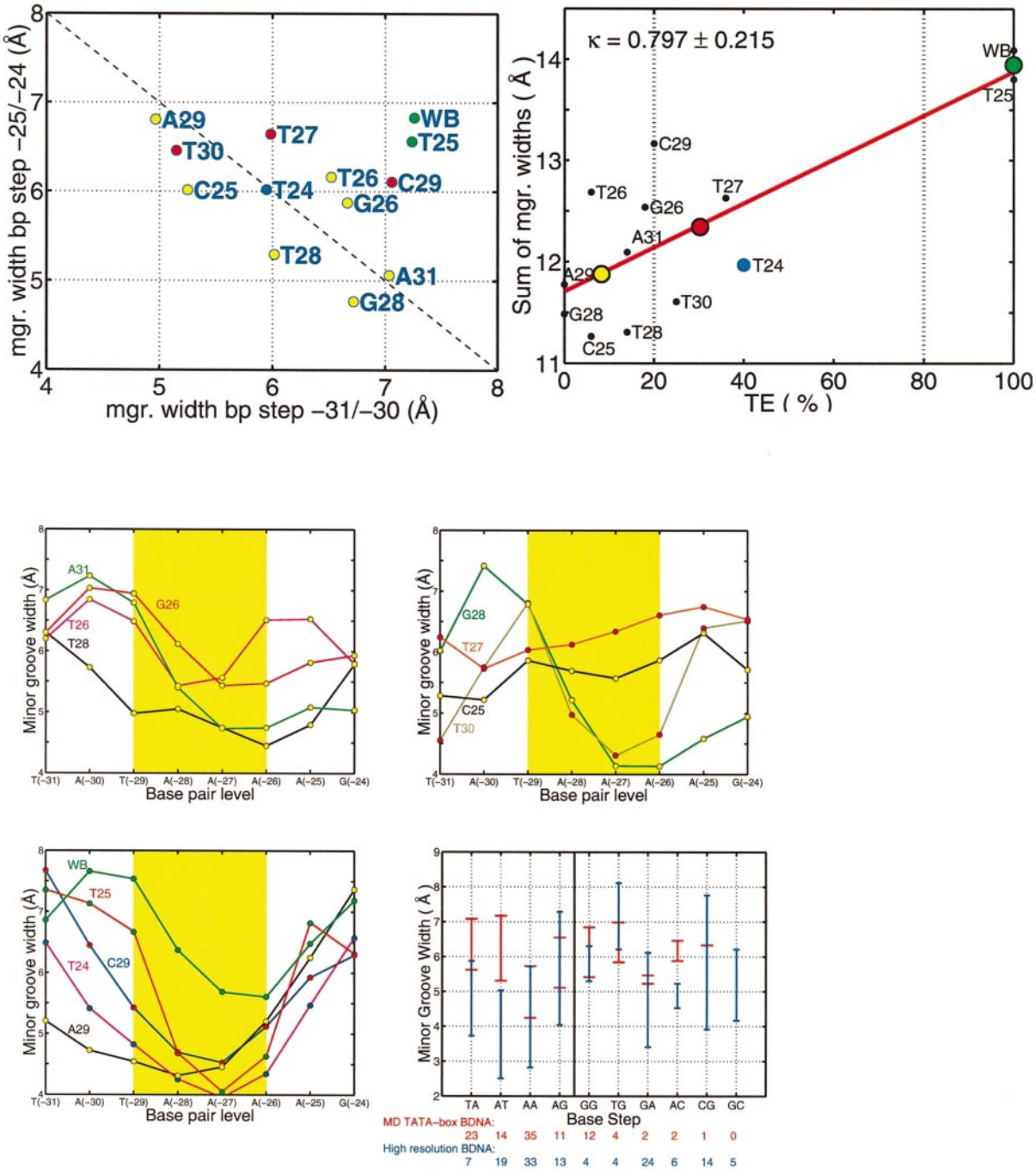

Figure 5. Minor groove analysis of TATA variants. Top left: average minor groove widths at base-pair steps −31/−30 and −25/−24. Minor groove widths of both the MD snapshots and the high-resolution NDB structures are calculated by the Curves program. Top right: correlation between selected minor groove widths at phenylalanine intercalation sites and TE. The averages for each of the three TE classes (see Figure 1) are also shown as large green, red, and yellow circles. Center left and right, and bottom left: ensemble average minor groove widths of the TATA element region over the last 1.8 ns for the 13 TATA variants. The average width at each step is labeled relative to the WB sequence. Widths (Å) are calculated as the distance between two base-pair points lying on spline curves generated on each strand from backbone phosphate atoms. The green (high TE), red (moderate TE), blue (estimated TE) and yellow (low TE) circles follow the color scheme in previous Figures. Bottom right: minor groove width analysis of high-resolution X-ray *B*-DNA crystal structures (blue) *versus* our MD results (red) for the TATA variants. A total of 24 *B*-DNA crystal structures with resolution $\leqslant$2.0 Å longer than 6 bp were obtained from the NDB (http://ndbserver.rutgers.edu); a list is provided in Computational Methodology. The MD data use all MD snapshots to calculate the mean and standard deviation of each step. The minor groove width per step for both the database and MD data is assigned to a step in the 5′ to 3′-direction of the DNA strand. For both sets of data, the minor groove widths are accumulated by the sequence of the base-pair step; the respective numbers of the accumulated steps for both the MD and high-resolution database structures are indicated below the plot.

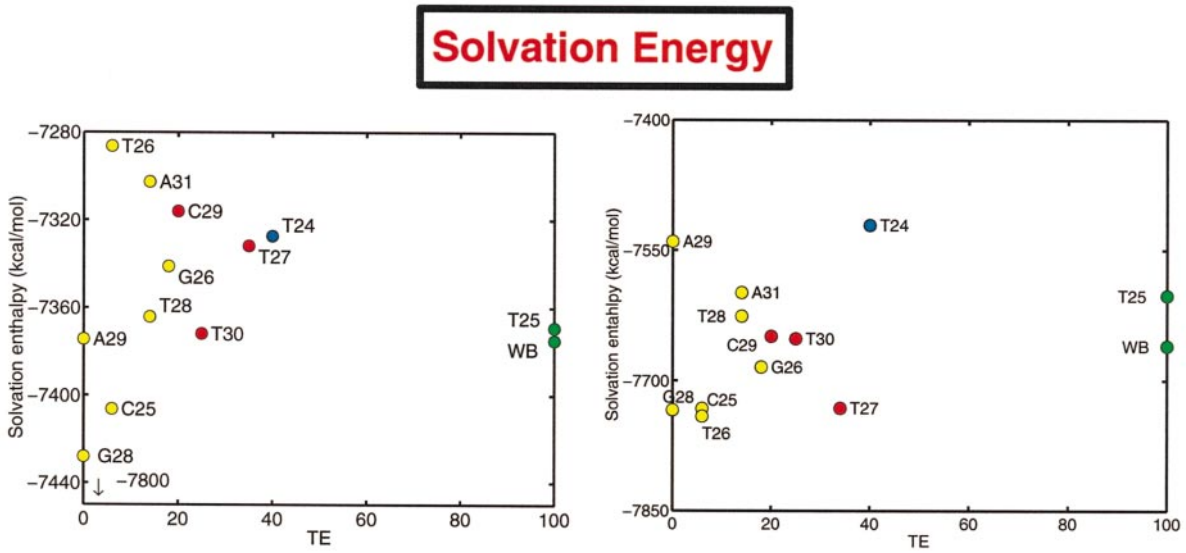

Figure 6. Solvation energy for 13 TATA variants plotted against TE. Left: the solvation enthalpy energy E was calculated as the sum of the electrostatic solvation energy ϕ computed by Delphi[58] and a surface area (A) dependent term representing the van der Waals energies: $E = R\phi + QA + b$, where R is equal to 0.593 (kcal/mol)/(k_BT/e), k_BT is the Boltzmann factor, Q is equal to 5.42 cal/(mol Å^2), and $b = 0.92$ kcal/mol.[59,60] Thirty snapshots per variant were analyzed due to computational constraints. Right: the solvation energies were calculated as the sum of the molecular mechanics force-field non-bonded, cavitation (a surface area-dependent term), and bonded energy changes of each DNA system with respect to a reference B-DNA structure.[61]

forces to stabilize solvent-exposed interactions with the TATA element.

TATA-DNA families

The combinations of factors required for optimal TBP interaction is echoed by our classification of variants by their cumulative motions (Figure 7). This classification is based on the pairwise correlations $\chi(i,j)$ (equation (3) in Computational Methodology) for two variants i and j involving the average percentage of the motion stemming from the first 100 ensemble PCs. We observe that our 13 DNAs fall into six classes: classes A (WB, G26, T27,

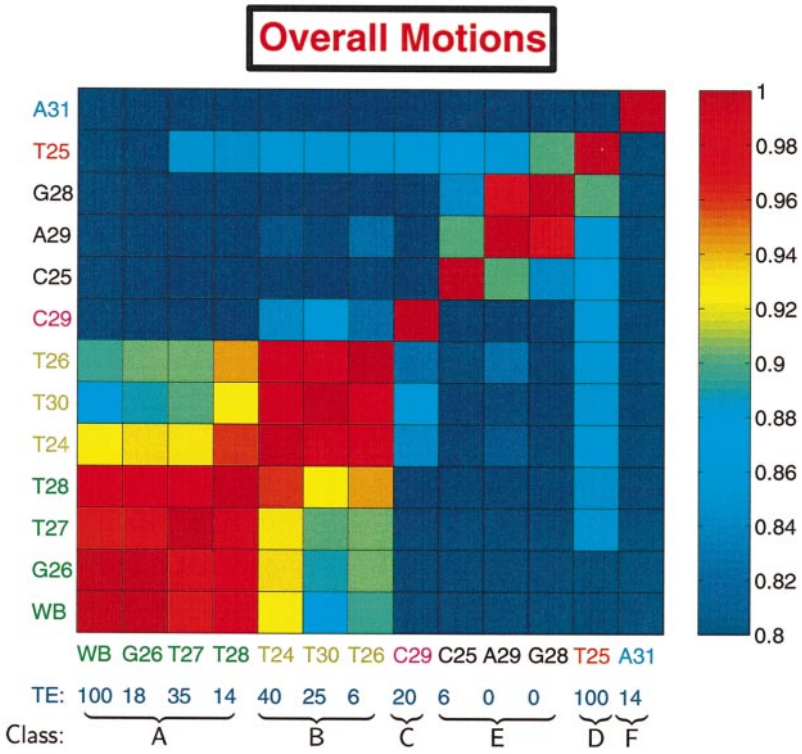

Figure 7. Grouping of TATA variants by overall flexibility factors, based on pairwise correlation coefficients $\chi(i,j)$ for the top 100 PCs of the TATA variants (equation (3)); the top 100 PCs cumulatively describe approximately 95 % of the total motion.

and T28) and B (T24, T30, and T26), which contain DNAs closely related to WB, with $\chi \approx 1$ (class A) and $\chi > 0.9$ (class B); class C (C29) with $0.8 < \chi < 0.9$, which resembles class B; class D (T25), also with $0.8 < \chi < 0.9$ but related to all other classes except F (A31); class E (G28, A29, C25), with minimal resemblance to classes A and B but resemblance to A29; and class F (A31), with no notable relation to any other class. Very roughly, the variants classified close to WB are associated with higher activity, while the classes close to A29 have lower activity (Figure 7). This combination of factors, manifested in properties such as flexibility (Figures 1, 3, and 7), local motions (as illustrated in Figures 2, 4 and 7), groove geometries (Figure 5), and local solvent and ion structure (Figure 4), contribute to TBP's selectivity. The influence of these structural factors and dynamic motions should further emerge in our continuing simulations of TBP/TATA complexes.

Conclusions

Understanding the structural, energetic, and dynamic aspects of eukaryotic transcription complexes has been an ongoing challenge.[9] The key eukaryotic transcriptional regulator TATA box binding protein (TBP) severely deforms the canonical *B*-DNA structure of the TATA recognition element, resulting in significant unwinding of the DNA and bending of more than 90°.[75] It is intriguing that while these structural distortions are conserved in single base-pair variations in the TATA elements, transcriptional activity can be greatly compromised by these mutations. The extensive series of nanosecond molecular dynamics simulations reported here, coupled with crystallographic and transcriptional activity data[14] and many other works to date on TATA elements (e.g. see Table 1) and DNA/protein complexes have highlighted a remarkable complementarity between DNA motion and TBP. Our results provide the basis of a refined kinetic hypothesis for TBP/TATA recognition and the interpretation of transcriptional activity. Namely, factors identified with increasing activity include enhanced global bending flexibility, local motions such as rise, roll, and shift at the ends of the TATA recognition element, and groove widening at these ends. These motions affect the local solvent and ion structure, establishing environmental trends contributing to TBP's selectivity. Together with structural and flexibility features of the final DNA/TBP complex, such findings highlight the role of the intrinsic DNA characteristics within the larger macromolecular assemblies associated with transcription. Our studies fit well with the large body of work on the fundamental influence of DNA sequence on biological activity[6,7,32] and may help the interpretation of other DNA/protein processes.

Our current hypothesis emerging from free TATA elements is now being examined closely in simulations of 13 analogous TATA/TBP complexes. Already, trends suggest that sequence-dependent motions in the complexes spread globally, leading to very different interface geometries for TFIIA/TFIIB recognition.[74,76,77] Moreover, subtle local motions lead to altered interactions within the DNA/TBP interface. Particularly important differences emerge for the roll values at the 5′ and 3′-phenylalanine intercalation sites, and the salt ions and water interactions. These details affect, in turn, the overall complex curvature and thus complex stability. Such sequence-dependent deformations of variant TATA/TBP complexes and their associated relevance to transcriptional activity will be detailed elsewhere.

Computational Methodology

System preparation and simulation method

Each 14-bp DNA duplex (see Table 2) was built in the standard *B*-DNA conformation using InsightII (Molecular Simulations, 1998). The 5′-terminal phosphate groups of each strand were replaced by hydroxyl groups. Ions were initially positioned 5 Å from the phosphate groups along the O-P-O bisector; a total of 26 sodium ions per duplex were included for charge neutralization. The initial water coordinates were generated by translation of a unit cell derived from the ice Ih hexagonal lattice, modified to increase the O-O distances to match the bulk density of liquid water.[78] This procedure was used to solvate regular hexagonal prisms (71.2 Å height with 28.8 Å side) containing the DNA and ions with ≈4800 TIP3P water molecules. Water molecules within 1.8 Å of the DNA heavy atoms were carved out of the system.

Periodic boundary conditions and the AMBER PARM94 force-field[36] converted for use in CHARMM version 26α2[79] are used for all energy minimizations and MD simulations. Non-bonded interactions are truncated at 12 Å, with force-shift for electrostatic and potential-switch for van der Waals interactions.

Energy minimization of the system was divided into three stages. Initially, the DNA and ion coordinates were fixed and only the water molecules in the system were minimized using an adopted-basis Newton-Raphson protocol for 2000 steps. In the second stage, the ion coordinates were subsequently released for 4000 steps of minimization. In the final minimization stage, all atom position constraints were released and the entire system was minimized for 4000 steps.

The resulting minimized systems were heated to 300 K over 2 ps with a Leapfrog integrator. Our multiple-timestep Langevin integrator, LN,[80,81] was used to equilibrate the systems for 4 ps with timesteps $\Delta\tau/\Delta t_m/\Delta t$ of 1/2/4 fs for fast/medium/slow force components (see below). A snapshot of the AdMLP system after 600 ps of MD simulation is shown in Figure 8. Each trajectory was simulated by LN for 2.4 ns with the timestep protocol of 1/2/120 fs. SHAKE constraints were applied to all bonds with hydrogen atoms. Coordinates were saved every 6 ps, and the last 1.8 ns of the trajectories used for data analysis. Each 2.4 ns trajectory took about 12 days (288 hours) on four 300-MHz R12000 processors of the NYU SGI Origin 2000 computer, or 280 hours on eight

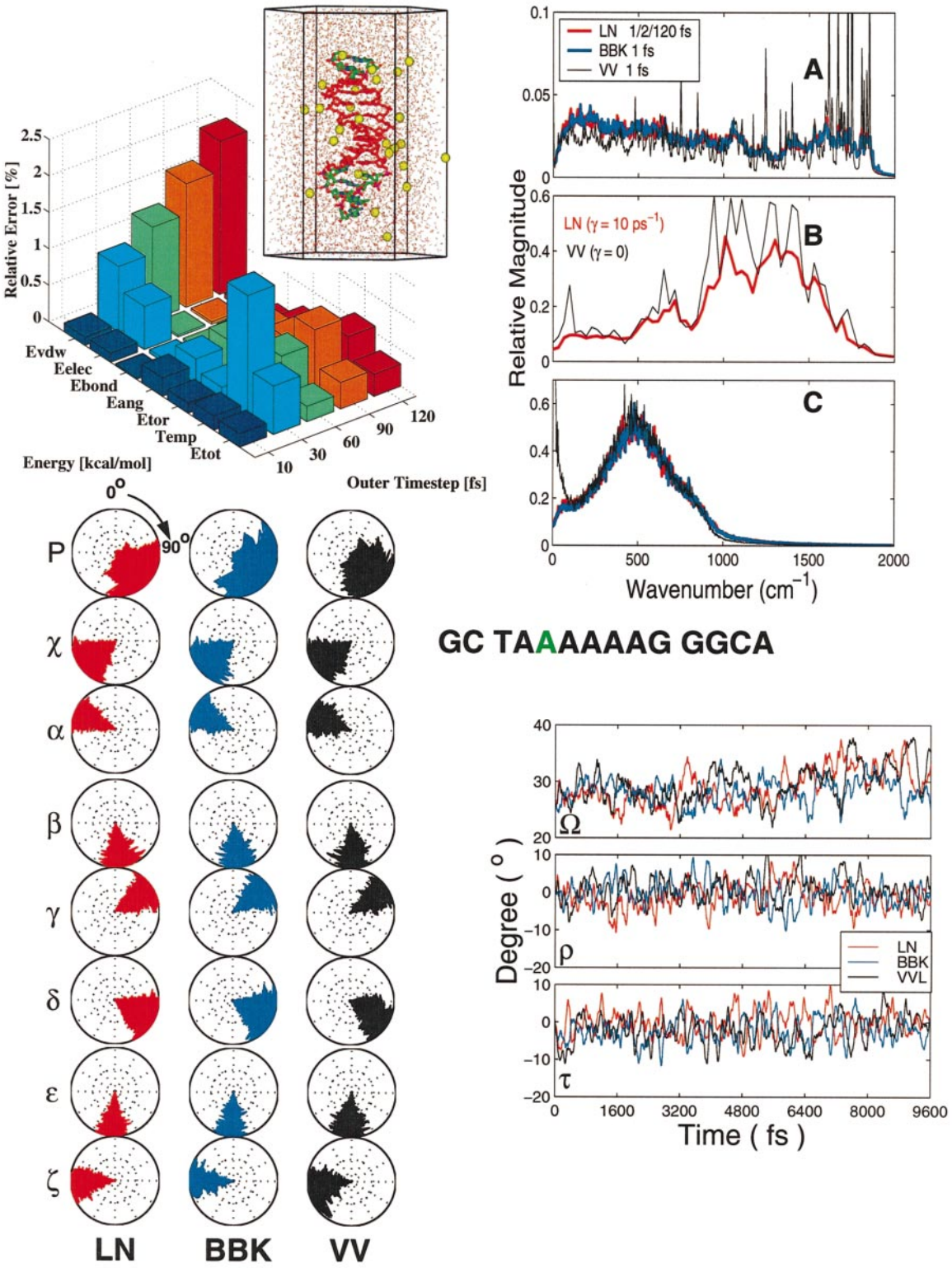

Figure 8. (*legend opposite*)

195-MHz R10000 processors of the NCSA Origin 2000 cluster. Performance details of the LN integrator (including speedup, error relative to single timestep methods, spectral densities, and geometry analyses) are discussed below.

Force-field and sampling limitations

We have chosen the AMBER force-field[36] over CHARMM[83] due to its general popularity and reliability for nucleic acid simulations.[20,52,84] Still, the AMBER94 force-field has been noted to prefer *B*-like DNA

Table 4. Relative transcriptional efficiences (TEs) for TATA variants collected from Starr *et al.*[67] (Hawley), Wobbe & Struhl[13] (Struhl), and Patikoglou *et al.*[14] (Burley)

Sequence	Reported TEs		
	Hawley	Struhl	Burley
A31		13	14
C31	41	12	
G31		2	
C30		2	
G30		1	
T30	68	25	25
A29	11	<1	
A29/T27	15		
C29		8	20
G29		1	
C28		1	
G28		<1	
T28	35	10	14
T28/T27	72	16	
C27		2	
G27		2	
T27	73	30	35
T27/C28		1	
T27/G28		1	
T27/A25		40	
T27/C25		14	
T27/T25		40	
C26		3	
G26		22	18
T26		4	6
C25	24		5
T25			100

The TE data from Hawley and Burley are normalized relative to WB (AdMLP), TATA·AAAG, whereas the TE data from Struhl are normalized relative to the *his3* promoter (TATA·AAGT; G25/T24 in the Burley notation); the scaling factor of 0.4 suggested by Wobbe & Struhl[13] is applied to their data to reflect the different intrinsic activities of WB and the *his3* promoter. Here, the Struhl variant T27 is actually T27/G25/T24 in the Burley and Hawley notation. Note that the TEs reported by Hawley are consistently larger than those reported in the Struhl or Burley studies by a factor of ≈2 or more.

structures,[85] to undertwist *B*-DNA, and to offset sugar pucker angles from the C2′-*endo* pucker;[52,86] this is being rectified with newer versions,[87] as well as other force-fields.[83] However, many researchers consider the AMBER94 force-field state-of-the-art despite its limitations,[84] since it reproduces structures in agreement with experimental predictions[29,46,52] and reasonably models transitions between *B* and *A*-DNA forms.[88] Moreover, correcting the force-field undertwisting tendency is non-trivial,[84] because of the many interdependent force-field terms and parameters. The force-field dependencies are less significant when results are analyzed within a large set of DNAs simulated under the same conditions, as done here.

In addition to the force-field limitations, even state-of-the-art biomolecular simulations are limited by the available computational resources, though the efficient LN integrator[80,81] allows us to complete 13 such trajectories

Figure 8. Solvated WB (5′-GC $TATA_4G$ GGCA-3′) with 4812 water molecules and 26 sodium ions (yellow) in a hexagonal prism domain and the performance of the LN integrator: errors in LN energy components, spectral densities for DNA and water, and selected dihedral angles relative to single timestep (1 fs) Langevin (BBK) and Velocity Verlet (VV) integrators for 4800 snapshots over 9.6 ps of the A29 variant. Top left: LN errors are shown for different energy components. The inner and medium timesteps are fixed at 1 and 2 fs, respectively, and the outer timestep was varied from 10 to 120 fs. Other conditions, such as non-bonded truncation distances, are the same as mentioned in the text. The energy components are van der Waals (Evdw), electrostatic (Eelec), bond (Ebond), angle (Eang), torsion (Etor), and total energy (Etot); the system temperature is also shown (Temp). Top right: the spectral densities of DNA (A and B) and water (C) are shown for the LN (red lines), BBK (blue), and VV (black) integrators. The spectral densities were computed for trajectories using SHAKE (A and C) or not using SHAKE (B). Bottom left: selected geometric (deoxyribose phase P) and dihedral quantities (χ, α, β, γ, δ, ε, and ζ) for base-pair step 3 of the TATA element (TA**A**AAAAG, indicated in green) computed with LN (red), BBK (blue), and VV (black) integrators. The quantities are plotted with Dials coordinates (with 0° at North and 90° at East) where the Dials coordinate pairs t and θ are time and angle, respectively. The dihedral angle nomenclature follows Saenger.[82] Bottom right: time evolution of twist Ω, roll ρ, and tilt τ for base-pair step 3 for the three integrators.

in several months on multiple processors, as discussed below. Despite these inherent drawbacks of current theoretical approaches, systematic analyses of complex systems have complemented experimental studies over the past decade.

LN integrator

The LN integrator for Langevin dynamics,[80,81,89,90] so called for its origin in a Langevin/Normal modes scheme, uses three force classes. The short timestep cycle updates the bond, angle, and dihedral energy terms every $\Delta\tau$ interval; the medium timestep cycle updates the non-bonded interactions within a spherical distance (7 Å is used here) every Δt_m interval; and the outer timestep Δt denotes the frequency of computing the remaining non-bonded interactions (up to the global non-bond interaction cutoff). Our LN simulations used inner/medium/outer timesteps of 1/2/120 fs and a medium-range cutoff of 7 Å with healing and buffer lengths of 4 Å each†.

Similar to the global non-bonded cutoff, a force shift function smoothes the transitions between medium-range and outer-range forces.[80] A damping constant $\gamma = 10\ \mathrm{ps}^{-1}$ couples the system to a 300 K heat bath. As shown by Barth & Schlick, smaller γ values generate trajectories closer to Newtonian trajectories; γ in the 10 ps^{-1} range also ensures stability by masking resonances sufficiently.[80,81,89] With this protocol, a speedup factor of 4.5 can be obtained compared to single timestep (1 fs) Verlet dynamics.

As shown in Figure 8, different energy components are consistently smaller than 2.5 % relative to a reference single-timestep Langevin integrator during a 9.6 ps trajectory with a timestep of 1 fs (Figure 8). Comparisons of the spectral densities of DNA and water, and the fluctuations of selected geometric quantities for 4800 snapshots sampled over 9.6 ps in Figure 8 (such as the deoxyribose puckering angle, glycosidic angle χ and phosphate-deoxyribose backbone dihedrals for a selected residue) indicate overall similarities between LN, the reference Langevin, and the Velocity Verlet trajectories.

The parallel version of the LN integrator was used to accelerate the simulations on four 300 MHz SGI R12000 processors of a 16 processor Silicon Graphics Origin 2000 system or eight 195 MHz SGI R10000 processors of an NCSA cluster of Origin 2000 systems (ranging from 32 to 128 processors). With the combined acceleration of LN and parallelization, a speedup factor of 18 is achieved over a single processor-single timestep Verlet simulation methodology. Each 2.4 ns trajectory takes about 48 days (1150 hours) on a single processor, about 12 days (288 hours) on four processors of the NYU Origin, or 280 hours on eight processors of an NCSA Origin.

Structure analysis

Nucleic acid structural parameters were derived from Curves, version 5.2.[37,38] Bending analysis described by global roll (θ_R) and global tilt (θ_T) is performed with our program Madbend[29] (http://monod.biomath.nyu.edu/, click on Software). These global angles incorporate tilt (τ), roll (ϱ), and twist (Ω) from each base-pair step as used in the prevalent models of DNA structure. Several global bending formulae have been described by the groups of Zhurkin, Trifonov, Lavery, Beveridge, and Ornstein;[17,23,24,30,31] these differ from Madbend, since they measure bending at a single angle between vectors associated with extremal base-pair steps. Our reference plane for measuring bending is the center of the TATA element (between base-pairs 4 and 5), corresponding to the appropriate 2-fold pseudo-rotation symmetry in the TBP/TATA complex.

Bending analyses in terms of ellipsoids generated by eigenvectors specifying bending direction are useful for comparison of bending trends in nucleic acids. Similar concepts have been advocated in the different context of local conformational analyses.[32] In our analysis of global bending, the major and minor axes of the ellipses bounding 90 % of the individual ensemble θ_R/θ_T data in Figure 1 were computed using PCA eigenvectors from the $\{\theta_T,\theta_R\}$ data. From the ranked eigenvectors ($\mathbf{V}_a$ and $\mathbf{V}_b$) and eigenvalues (α and β), we assign $\mathbf{V}_a$ to the major axis direction, with initial length α, and $\mathbf{V}_b/\beta$ to the minor axis direction/length. The ellipse size was determined by optimizing α and β in Matlab (The MathWorks, 1999) using a function $f(\alpha,\beta)$ of the ellipse area that requires 90 % of the data (M_e points) to be included: $\min_{\alpha,\beta} f(\alpha,\beta)$, where $f(\alpha,\beta) = [c_1(M_e/M - 0.9)^2 - c_2(M_e/(\pi ab))]$, and M is the total number of points (or snapshots); a and b are the major ($|\alpha\mathbf{V}_a|$) and minor ($|\beta\mathbf{V}_b|$) axis lengths, respectively; πab is the ellipse area; and c_1 and c_2 are adjustable constants (we use 800 and 5, respectively).

Water and ion probability densities were calculated on a 1 Å cubic lattice using procedures implemented in CHARMM (interested users are invited to contact us).

The correlation between the TE (TE_i) and a property (P_i) of variant i is calculated by evaluating the linear correlation coefficient $\kappa(P,\mathrm{TE})$ as:

$$\kappa(P,\mathrm{TE}) = \left(\sum_i P_i \cdot \mathrm{TE}_i\right) \Big/ \sqrt{\left(\sum_i P_i^2\right)\left(\sum_i \mathrm{TE}_i^2\right)} \qquad (1)$$

We estimate the error associated with $\kappa(P,\mathrm{TE})$ by standard error propagation techniques:[92]

$$\sigma_\kappa = \sqrt{\sum_i \left[\left(\frac{\partial \kappa}{\partial P_i}\sigma_{P_i}\right)^2 + \left(\frac{\partial \kappa}{\partial \mathrm{TE}_i}\sigma_{\mathrm{TE}_i}\right)^2\right]}$$

For this error, the standard deviation σ_{P_i} for each property P_i of variant i is obtained from the discrete sampling of the trajectories. The standard deviation σ_{TE_i} in TE of variant i is estimated from the multiple determinations of activity from different groups (see Table 4); if only one TE value is available, we assign the largest TE deviation of ≈8 % (obtained for the C29 variant). In the cases of A29 and G28 (no measurable activity) and WT (normalized to be 100 %), a zero deviation is assigned. Properties analyzed in this way include flexibility, minor groove width, and the normalized mean-square magnitude of a PC.

Database analysis

Twenty-four *B*-DNA structures with a resolution less than 2 Å longer than 6 bp were obtained from the Nucleic Acid Database (http://ndbserver.rutgers.edu).[53] The 24 structures are: BD0005, BD0007, BD0016, BD0018,

† LN simulations in combination with the particle mesh Ewald method[91] are now possible following the work of Batcho *et al*. (unpublished results and [90]).

BD0019, BD0023, BD0029, BD0037, BDF068, BDJ017, BDJ019, BDJ025, BDJ031, BDJ036, BDJ037, BDJ051, BDJ052, BDJ060, BDJ061, BDJ081(chains A/B), BDJ081(chains C/D), BDJ081(chains E/F), BDL001, BDL005. Minor groove width parameters were derived from Curves, version 5.2.[37,38]

Principal component analysis (PCA)

PCA decomposes the motions of a trajectory into independent modes, hierarchically organized so that the first several modes describe most of the motion characteristics of the trajectory. PCA has been widely used to study the intrinsic motions of both nucleic acids (including global bending)[41] and proteins,[40] as mentioned in Results. We use PCA applied to each variant's trajectory (an ensemble is the 300 frames collected over the 1.8 ns production run) and develop a new procedure termed uniform ensemble PCA to directly compare the PCs among our 13 variants. The latter is based on a merged trajectory of all variants' trajectories; additional details are available below.

A covariance matrix **C** is constructed using the average structure from the merged configurational ensemble as the following sum of outer products:

$$\mathbf{C} = \frac{1}{M}\sum_{k=1,M}(\mathbf{X}_k - \langle\mathbf{X}\rangle)(\mathbf{X}_k - \langle\mathbf{X}\rangle)^T$$

where $\mathbf{X}_k$ is the coordinate vector at the kth snapshot, and $\langle\mathbf{X}\rangle$ is the average structure from the dynamics simulation:

$$\langle\mathbf{X}\rangle = \frac{1}{M}\sum_{k=1,M}\mathbf{X}_k$$

The average structure used as a reference to develop the covariance matrices **C** is the unminimized coordinate average. Diagonalization of **C** produces the eigenvalues and eigenvectors as entries of Λ from the decomposition:

$$\mathbf{V}^T\mathbf{C}\mathbf{V} = \Lambda$$

or

$$\mathbf{C}\mathbf{V}_n = \lambda_n\mathbf{V}_n, \quad n = 1, 2, \ldots, 3N$$

where Λ is the diagonal matrix with eigenvalues $\{\lambda_i\}$: $=\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_{3N})$.

Each eigenvector $\mathbf{V}_n$ defines the direction of motion of N atoms as an oscillation about the average structure $\langle\mathbf{X}\rangle$. The normalized magnitude of the corresponding eigenvalue $(\lambda_n/\Sigma_{n=1,3N}\lambda_n)$ indicates the relative percentage of the trajectory motions along eigenvector $\mathbf{V}_n$.

Uniform ensemble PCA setup

To ensure comparable numbers of atoms between different variants, we set the WB sequence to be the reference and perform the following manipulation on each variant:

1A: if the variant → WB conversion is a pyrimidine to pyrimidine replacement (e.g. C to T, as in C29 to WB) or a purine to purine replacement (e.g. G to A, as in G26 to WB), the phosphate/deoxyribose backbone atoms and all non-hydrogen atoms of the pyrimidine and purine rings of the mutated base-pair are maintained. Exocyclic side-chains (such as the thymine methyl group) and hydrogen atoms are then built using standard geometries, and the nucleotides are accordingly renamed.

1B: if the variant → WB conversion is a pyrimidine to purine replacement (e.g. A to T, as in A31 to WB) or a purine to pyrimidine replacement (e.g. T to A, as in T27 to WB), the phosphate/deoxyribose backbone atoms are again maintained. Non-hydrogen atoms of the five-membered purine ring or the six-membered pyrimidine ring are used to replace the bases according to the superimposed positions of a purine and a pyrimidine in the standard *B*-DNA conformation. The remaining hydrogen atoms are built using standard geometries, and the nucleotides are accordingly renamed.

2: after the above base-pair replacement and adjustment, all atoms except those rebuilt from standard geometries are fixed. A short minimization (200 steps of adopted basis Newton-Raphson) is performed to optimize the exocyclic side-chain and hydrogen positions of the replaced bases.

This procedure introduces minimal perturbations to our trajectories: all base-pair geometries are maintained, and the average relative error between local base-pair step parameters before and after replacement is less than 2%. The average structure of the merged trajectory (3900 total frames from 13 variant trajectories of 300 frames each) is used to orient each frame of the merged trajectory to minimize the rmsd of the TATA element. The reoriented merged trajectory then produces a second average structure. The above process is repeated until the average structure converges between cycles and no rotation is necessary to minimize the rmsd.

Structure generation using PCs

An arbitrary structure **Y** can be generated from the average structure $\langle\mathbf{X}\rangle$ by a displacement **D** along the linear combination of all eigenvectors $\mathbf{V}_n$ with $3N$ scalars α_n, where:

$$\mathbf{Y} = \langle\mathbf{X}\rangle + \mathbf{D} = \langle\mathbf{X}\rangle + \sum_{n=1,3N}\alpha_n\mathbf{V}_n, \quad \alpha_n = \mathbf{V}_n^T\mathbf{D}$$

This basic method of generating structures from PCs has utility in several analysis procedures, which we describe below, such as measuring the motions of single and combined PCs (see Figures 2 and 3).

Namely, single PCs (e.g. PCs 1, 2, and 32, Figure 2) are analyzed by determining the structural deformations associated with the eigenvectors. A scalar α_n corresponding to the deformation **D** is computed by considering the minimal and maximal projection of individual PCs against the MD trajectory. The difference between the minimal and maximal projection is divided into ten equal segments. The resulting set of deformations **D** is used to generate 11 structures. The structures may then be analyzed with standard programs such as Curves or animated in visualization packages such as Insight.

This method can be applied to combined PCs; however, it requires that PCs be combined to form a single invariant eigenvector, though the instantaneous weight of each PC varies during an MD trajectory. If the weights are approximately comparable (as in our analysis of PCs 18, 20, 21, 32, and 34 in Figure 1), we use one weight. If the weights are significantly different in magnitude, we use filtering as described below.

PC analysis by trajectory filtering

An alternative method, filtering, was used to examine the combined motion of several PCs. For example, we applied this method to analysis of the top five PCs of the individual variant PCA, since the relative projections of the top five PCs changes rapidly and we could not combine the PCs to form a single invariant eigenvector.

In filtering, each trajectory snapshot $\mathbf{X}_k$ is defined as the deformation $\mathbf{X}_D$ from the average structure: $\mathbf{X}_k = \langle \mathbf{X} \rangle + \mathbf{X}_D$. The filtered snapshot $\mathbf{X}_k^{\text{filtered}}$, filtered to display only the motions of the PC or of several PCs, is generated by considering the projection of the deformation $\mathbf{X}_D$ against each PC:

$$\mathbf{X}_k^{\text{filtered}} = \langle \mathbf{X} \rangle + \sum_{\mathbf{V}} (\mathbf{X}_D \cdot \mathbf{V})\mathbf{V}$$

The snapshots of the filtered trajectory may be analyzed with Insight or Curves as described above. Although the filtering method accurately reports the inherent deformations along each PC during an MD simulation, the filtered trajectory does not represent a linear deformation along each PC and the results are consequently more difficult to interpret.

PC analysis by relative magnitude of motion

Following individual or uniform ensemble PCA, we compare for each variant i the normalized mean-square magnitude of the projection along PC n, $\overline{\alpha_{n,i}^2}$:

$$\overline{\alpha_{n,i}^2} = \frac{1}{\text{Tr}(\Lambda)} \frac{1}{M_i} \sum_{k=1,M_i} (\alpha_{n,i}^k)^2 \qquad (2)$$

where M_i is the number of trajectory frames of variant i, $\text{Tr}(\Lambda)$ is the sum of covariance matrix eigenvalues $\text{Tr}(\Lambda) = \Sigma_{n=1,3N} \lambda_n$, and $\alpha_{n,i}^k$ is the projection of sequence i Ton PC n at frame k.

For any two TATA variants i and j, we measure the similarities between the dynamic motions of the two variants by the correlation coefficient $\chi(i,j)$:

$$\chi(i,j) = \left(\sum_n \overline{\alpha_{n,i}^2} \cdot \overline{\alpha_{n,j}^2} \right) \Big/ \left[\left(\sum_n \overline{\alpha_{n,i}^2}^2 \right)^{1/2} \left(\sum_n \overline{\alpha_{n,j}^2}^2 \right)^{1/2} \right] \qquad (3)$$

where $\overline{\alpha_{n,i}^2}$ is the normalized mean-square magnitude of the projection of PC n for variant i and $n = 1, 2, 3, \ldots, 100$. Our cutoff value of 100 includes ≈95% of the ensemble motion. This χ analysis was used to generate Figure 7.

Our PCA procedures have been applied only to the heavy (non-hydrogen) atoms of the TATA element octamer; hydrogen and heavy atoms outside the TATA element are ignored. A total of 328 atoms are included in the analysis, resulting in 984 PCs (3×328). Snapshots are sampled from the last 1.8 ns of each trajectory at a frequency of $\Delta t = 6$ ps.

Acknowledgments

We are indebted to Dr Steve Burley for proposing, stimulating, and contributing to this exciting project though many discussions. We thank Dr Richard Lavery for use of the Curves program, and Dr Wilma Olson for reference suggestions. The work was supported by NIH grant GM55164, NSF grants BIR-94-23827EQ and ASC-9704681, and a John Simon Guggenheim Fellowship to T. S. Parts of the computations were performed on the NCSA Origin 2000 cluster at the University of Illinois Urbana-Champaign under NCSA grant MCA99S021N to T. S., who is an investigator of the Howard Hughes Medical Institute.

References

1. Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerise II promoter: elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212**, 563-578.
2. Burley, S. K. & Roeder, R. G. (1996). Biochemistry and structural biology of transcription factor IID (TFIID). *Annu. Rev. Biochem.* **65**, 769-799.
3. Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. (1993). Crystal structure of a yeast TBP/TATA-box complex. *Nature,* **365**, 512-520.
4. Kim, J. L., Nikolov, D. B. & Burley, S. K. (1993). Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature,* **365**, 520-527.
5. Tisne, C., Delepierre, M. & Hartmann, B. (1999). How NF-κB can be attracted by its cognate DNA. *J. Mol. Biol.* **293**, 139-150.
6. Dickerson, R. E. & Chiu, T. K. (1997). Helix bending as a factor in protein/DNA recognition. *Biopolymers,* **44**, 361-403.
7. Dickerson, R. E. (1998). DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucl. Acids Res.* **26**, 1906-1926.
8. Grove, A., Galeone, A., Yu, E., Mayol, L. & Geiduschek, E. P. (1998). Affinity, stability and polarity of binding of the TATA binding protein governed by flexure at the TATA box. *J. Mol. Biol.* **282**, 731-739.
9. Hampsey, M. (1998). Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbio. Mol. Biol. Rev.* **62**, 465-503.
10. Suzuki, M. & Yagi, N. (1995). Stereochemical basis of DNA bending by transcription factors. *Nucl. Acids Res.* **23**, 2083-2091.
11. Suzuki, M., Amano, N., Kakinuma, J. & Tateno, M. (1997). Use of a 3D structure data base for understanding sequence-dependent conformational aspects of DNA. *J. Mol. Biol.* **274**, 421-435.
12. Kosikov, K. M., Gorin, A. A., Zhurkin, V. B. & Olson, W. K. (1999). DNA stretching and compression: large-scale simulations of double helical structures. *J. Mol. Biol.* **289**, 1301-1326.
13. Wobbe, C. R. & Struhl, K. (1990). Yeast and human TATA-binding proteins have nearly identical DNA sequence requirements for transcription *in vitro*. *Mol. Cell. Biol.* **10**, 3859-3867.
14. Patikoglou, G. A., Kim, J. L., Sun, L., Yang, S.-H., Kodadek, T. & Burley, S. K. (1999). TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev.* **13**, 3217-3230.
15. Guzikevich-Guerstein, G. & Shakked, Z. (1996). A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein. *Nature Struct. Biol.* **3**, 32-37.

16. Elcock, A. H. & McCammon, J. A. (1996). The low dielectric interior of proteins is sufficient to cause major structural changes in DNA on association. *J. Am. Chem. Soc.* **118**, 3787-3788.
17. Flatters, D., Young, M., Beveridge, D. L. & Lavery, R. (1997). Conformational properties of the TATA-box binding sequence of DNA. *J. Biomol. Struct. Dynam.* **14**, 757-765.
18. Lebrun, A., Shakked, Z. & Lavery, R. (1997). Local DNA stretching mimics the distortion caused by the TATA box-binding protein. *Proc. Natl Acad. Sci. USA,* **94**, 2993-2998.
19. Pardo, L., Pastor, N. & Weinstein, H. (1998). Progressive DNA bending is made possible by gradual changes in the torsion angle of the glycosyl bond. *Biophys. J.* **74**, 2191-2198.
20. Pardo, L., Pastor, N. & Weinstein, H. (1998). Selective binding of the TATA box-binding protein to the TATA box-containing promoter: analysis of structural and energetic factors. *Biophys. J.* **75**, 2411-2421.
21. Pardo, L., Campillo, M., Bosch, D., Pastor, N. & Weinstein, H. (2000). Binding mechanisms of TATA box-binding proteins: DNA kinking is stabilized by specific hydrogen bonds. *Biophys. J.* **78**, 1988-1996.
22. Pastor, N., Weinstein, H., Jamison, E. & Brenowitz, M. (2000). A detailed interpretation of OH radical footprints in a TBP-DNA complex reveals the role of dynamics in the mechanism of sequence-specific binding. *J. Mol. Biol.* **304**, 55-68.
23. de Souza, O. N. & Ornstein, R. L. (1998). Inherent DNA curvature and flexibility correlate with TATA box functionality. *Biopolymers,* **46**, 403-415.
24. Flatters, D. & Lavery, R. (1998). Sequence-dependent dynamics of TATA-box binding sites. *Biophys. J.* **75**, 372-381.
25. Pastor, N., Pardo, L. & Weinstein, H. (1997). Does TATA matter? a structural exploration of the selectivity determinants in its complexes with TATA box-binding protein. *Biophys. J.* **73**, 640-652.
26. Pastor, N., Pardo, L. & Weinstein, H. (1998). How the TATA box selects its protein partner. In *Molecular Modeling of Nucleic Acids* (Leontis, N. B. & SantaLucia, J., Jr, eds), *ACS Symposium Series*, vol. 682, pp. 329-345, American Chemical Society, Washington, DC.
27. Miaskiewicz, K. & Ornstein, R. L. (1996). DNA binding by TATA-box binding protein TBP: a molecular dynamics computational study. *J. Biomol. Struct. Dynam.* **13**, 593-600.
28. Strubin, M. & Struhl, K. (1992). Yeast and human TFIID with altered DNA-binding specificity for TATA elements. *Cell,* **68**, 721-730.
29. Strahs, D. & Schlick, T. (2000). A-tract bending: insights into experimental structures by computational models. *J. Mol. Biol.* **301**, 643-663.
30. Ulyanov, N. B. & Zhurkin, V. B. (1984). Sequence-dependent anisotropic flexibility of *B*-DNA. a conformational study. *J. Biomol. Struct. Dynam.* **2**, 361-385.
31. Zhurkin, V. B., Ulyanov, N. B., Gorin, A. A. & Jernigan, R. L. (1991). Static and statistical bending of DNA evaluated by Monte Carlo simulations. *Proc. Natl Acad. Sci. USA,* **88**, 7046-7050.
32. Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA,* **95**, 11163-11168.
33. Mayer-Jung, C., Moras, D. & Timsit, Y. (1998). Hydration and recognition of methylated CpG steps in DNA. *EMBO J.* **17**, 2709-2718.
34. Woda, J., Schneider, B., Patel, K., Mistry, K. & Berman, H. M. (1998). An analysis of the relationship between hydration arid protein-DNA interactions. *Biophys. J.* **75**, 2170-2177.
35. Hud, N. V., Sklenar, V. & Feigon, J. (1999). Localization of ammonium ions in the minor groove of DNA duplexes in solution and the origin of DNA A-tract bending. *J. Mol. Biol.* **286**, 651-660.
36. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Jr, Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* **117**, 5179-5197.
37. Lavery, R. & Sklenar, H. (1997). *Curves 5.2: Helical Analysis of Irregular Nucleic Acids*, Laboratoire de Biochimie Theorique, CNRS URA 77, Institut de Biologie Physico-Chimique, Paris, France.
38. Stofer, E. & Lavery, R. (1994). Measuring the geometry of DNA grooves. *Biopolymers,* **34**, 337-346.
39. Sessions, R. B., Dauber-Osguthorpe, P. & Osguthorpe, D. J. (1989). Filtering molecular dynamics trajectories to reveal low-frequency collective motions: phospholipase A2. *J. Mol. Biol.* **210**, 617-633.
40. Caves, L. S. D., Evanseck, J. D. & Karplus, M. (1998). Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci.* **7**, 649-666.
41. Sherer, E. C., Harris, S. A., Soliva, R., Orozco, M. & Laughton, C. A. (1999). Molecular dynamics studies of DNA A-tract structure and flexibility. *J. Am. Chem. Soc.* **121**, 5981-5991.
42. Lavery, R. & Sklenar, H. (1989). Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct.Dynam.* **6**, 655-667.
43. Lu, X. J. & Olson, W. K. (1999). Resolving the discrepancies among nucleic acid conformational analyses. *J. Mol. Biol.* **285**, 1563-1575.
44. Lu, X.-J., Babcock, M. S. & Olson, W. K. (1999). Overview of nucleic acid programs. *J. Biomol. Struct. Dynam.* **16**, 833-843.
45. Ulanovsky, L. & Trifonov, E. N. (1987). Estimation of wedge components in curved DNA. *Nature,* **326**, 720-722.
46. Young, M. A. & Beveridge, D. L. (1998). Molecular dynamics simulations of an oligonucleotide duplex with adenine tracts phased by a full helix turn. *J. Mol. Biol.* **281**, 675-687.
47. Sprous, D., Young, M. A. & Beveridge, D. L. (1999). Molecular dynamics studies of axis bending in d(G_5-(GA_4T_4C)$_2$-C_5) and d(G_5-(GT_4A_4C)$_2$-C_5): effects of sequence polarity on DNA curvature. *J. Mol. Biol.* **285**, 1623-1632.
48. Bernués, J., Carrera, P. & Azorin, F. (1996). TBP binds the transcriptionally inactive TA_5 sequence but the resulting complex is not efficiently recognised by TFIIB and TFIIA. *Nucl. Acids Res.* **24**, 2950-2958.
49. Suzuki, M., Allen, M. D., Yagi, N. & Finch, J. T. (1996). Analysis of co-crystal structures to identify the stereochemical determinants of the orientation of TBP on the TATA box. *Nucl. Acids Res.* **24**, 2767-2773.
50. Juo, Z. S., Chiu, T. K., Leiberman, P. M., Baikalov, I., Berk, A. J. & Dickerson, R. E. (1996). How

proteins recognize the TATA box. *J. Mol. Biol.* **261**, 239-254.
51. Pastor, N., MacKerell, A. D., Jr & Weinstein, H. (1999). TIT for TAT: the properties of inosine and adenosine in TATA box DNA. *J. Biomol. Struct. Dynam.* **16**, 787-810.
52. Young, M. A., Ravishanker, G. & Beveridge, D. L. (1997). A 5-nanosecond molecular dynamics trajectory for *B*-DNA: analysis of structure, motions, and solvation. *Biophys. J.* **73**, 2313-2336.
53. Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. & Schneider, B. (1992). The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* **63**, 751-759.
54. Wang, Y. & Stumph, W. E. (1995). RNA polymerase II/III transcription specificity determined by TATA box orientation. *Proc. Natl Acad. Sci. USA,* **92**, 8606-8610.
55. Kim, J. L. & Burley, S. K. (1994). 1.9 Å resolution refined structure of TBP recognizing the minor groove of TATAAAAG. *Nature Struct. Biol.* **1**, 638-653.
56. Kopka, M. L., Fratini, A. V., Drew, H. R. & Dickerson, R. E. (1983). Ordered water structure around a *B*-DNA dodecamer: a quantitative study. *J. Mol. Biol.* **163**, 129-146.
57. DiGabriele, A. D. & Steitz, T. A. (1993). A DNA dodecamer containing an adenine tract crystallizes in a unique lattice and exhibits a new bend. *J. Mol. Biol.* **231**, 1024-1039.
58. Gilson, M. K., Sharp, K. & Honig, B. H. (1987). Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comput. Chem.* **9**, 327-335.
59. Srinivasan, J., Cheatham, T. E., III, Cieplak, P., Kollman, P. A. & Case, D. A. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *J. Am. Chem. Soc.* **120**, 9401-9409.
60. Sitkoff, D., Sharp, K. A. & Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **98**, 1978-1988.
61. Jayaram, B., McConnell, K. J., Dixit, S. B. & Beveridge, D. L. (1999). Free energy analysis of protein-DNA binding: the *ECO*R1 endonuclease-DNA complex. *J. Comput. Phys.* **151**, 333-357.
62. Mirzabekov, A. D. & Rich, A. (1979). Asymmetric lateral distribution of unshielded phosphate groups in nucleosomal DNA and its role in DNA bending. *Proc. Natl Acad. Sci. USA,* **76**, 1118-1121.
63. Manning, G. S., Ebralidse, K. K., Mirzabekov, A. D. & Rich, A. (1989). An estimate of the extent of folding of nucleosomal DNA by laterally asymmetric neutralization of phosphate groups. *J. Biomol. Struct. Dynam.* **6**, 877-889.
64. Maher, L. J., III (1998). Mechanisms of DNA bending. *Curr. Opin. Chem. Biol.* **2**, 688-694.
65. McFail-Isom, L., Sines, C. C. & Williams, L. D. (1999). DNA structure: cations in charge? *Curr. Opin. Struct. Biol.* **9**, 298-304.
66. Chiu, T. K. & Dickerson, R. E. (2000). 1 Å crystal structures of *B*-DNA reveal sequence-specific binding and groove-specific bending of DNA by magnesium and calcium. *J. Mol. Biol.* **301**, 915-945.
67. Starr, D. B., Hoopes, B. C. & Hawley, D. K. (1995). DNA bending is an important component of site-specific recognition by the TATA binding protein. *J. Mol. Biol.* **250**, 434-446.
68. Parvin, J. D., McCormick, R. J., Sharp, P. A. & Fisher, D. E. (1995). Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. *Nature,* **373**, 724-727.
69. Hoopes, B. C., LeBlanc, J. F. & Hawley, D. K. (1992). Kinetic analysis of yeast TFIID-TATA box complex formation suggests a multi-step pathway. *J. Biol. Chem.* **267**, 11539-11547.
70. Parkhurst, K. M., Brenowitz, M. & Parkhurst, L. J. (1996). Simultaneous binding and bending of promoter DNA by the TATA binding protein: real time kinetic measurements. *Biochemistry,* **35**, 7459-7465.
71. Parkhurst, K. M., Richards, R. M., Brenowitz, M. & Parkhurst, L. J. (1999). Intermediate species possessing bent DNA are present along the pathway to formation of a final TBP-TATA complex. *J. Mol. Biol.* **289**, 1327-1341.
72. Hoopes, B. C., LeBlanc, J. F. & Hawley, D. K. (1998). Contributions of the TATA box sequence to rate-limiting steps in transcription initiation by RNA polymerase II. *J. Mol. Biol.* **277**, 1015-1031.
73. Werner, M. H., Gronenborn, A. M. & Clore, G. M. (1996). Intercalation, DNA kinking, and the control of transcription (Published erratum appears in Science **272**, 19, (1996)). *Science,* **271**, 778-784.
74. Tan, S., Hunziker, Y., Sargent, D. F. & Richmond, T. J. (1996). Crystal structure of a yeast TFIIA/TBP/DNA complex. *Nature,* **381**, 127-151.
75. Burley, S. K. (1996). The TATA-box binding protein. *Curr. Opin. Struct. Biol.* **6**, 69-75.
76. Nikolov, D. B., Chen, H., Halay, E. D., Usheva, A. A., Lee, K. H. D. K., Roeder, R. G. & Burley, S. K. (1995). Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature,* **377**, 119-128.
77. Wu, J., Parkhurst, K., Powell, R., Brenowitz, M. & Parkhurst, L. (2001). DNA bends in solution TBP-TATA complexes are DNA sequence dependent. *J. Biol. Chem.* **In the press**.
78. Qian, X., Strahs, D. & Schlick, T. (2001). A new program for optimizing periodic boundary models of solvated biomolecules (PBCAID). *J. Comput. Chem.* **In the press**.
79. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187-217.
80. Barth, E. & Schlick, T. (1998). Overcoming stability limitations in biomolecular dynamics. I. Combining force splitting *via* extrapolation with Langevin dynamics in LN. *J. Chem. Phys.* **109**, 1617-1632.
81. Sandu, A. & Schlick, T. (1999). Masking resonance artifacts in force-splitting methods for biomolecular simulations by extrapolative Langevin dynamics. *J. Comput. Phys.* **151**, 74-113.
82. Saenger, W. (1984). *Principles of Nucleic Acid Structure*, Springer-Verlag, New York.
83. Foloppe, N. & MacKerell, A. D., Jr (2000). All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.* **21**, 86-104.
84. Beveridge, D. L. & McConnell, K. J. (2000). Nucleic acids: theory and computer simulation, Y2K. *Curr. Opin. Struct. Biol.* **10**, 182-196.

85. Feig, M. & Pettitt, B. M. (1997). Experiment vs. force fields: DNA conformation from molecular dynamics simulations. *J. Phys. Chem ser. B,* **101**, 7361-7363.
86. Olson, W. K. & Zhurkin, V. B. (2000). Modeling DNA deformations. *Curr. Opin. Struct. Biol.* **10**, 286-297.
87. Cheatham, T. E., III & Kollman, P. A. (1999). A modified version of the Cornell *et al*. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct Dynam.* **16**, 845-862.
88. Cheatham, T. E., III & Kollman, P. A. (1997). Insight into the stabilization of A-DNA by specific ion association: spontaneous *B*-DNA to *A*-DNA transitions observed in molecular dynamics simulations of d[ACCCGCGGGT]2 in the presence of hexaamminecobalt(III). *Structure,* **5**, 1297-1311.
89. Barth, E. & Schlick, T. (1998). Extrapolation versus impulse in multiple-timestepping schemes. II. Linear analysis and applications to Newtonian and Langevin dynamics. *J. Chem. Phys.* **109**, 1633-1642.
90. Schlick, T. (2001). Time-trimming tricks for dynamic simulations: splitting force updates to reduce computational work. *Structure,* **9**, R45-R53.
91. Darden, T., Perera, L., Li, L. & Pedersen, L. (1999). New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure,* **7**, R55-R60.
92. Bevington, P. R. (1969). *Data Reduction and Error Analysis for the Physical Science*, McGraw-Hill, New York.

Edited by B. Honig

(Received 6 December 2000; received in revised form 1 March 2001; accepted 1 March 2001)