Daniel Strahs

Danny Barash

Xiaoliang Qian

Tamar Schlick

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

Sequence-Dependent Solution Structure and Motions of 13 TATA/TBP (TATA-Box Binding Protein) Complexes

Abstract: *The TATA element is a well-known example of a DNA promoter sequence recognized by the TATA box binding protein (TBP) through its intrinsic motion and deformability. Although TBP recognizes the TATA element octamer unusually (through the minor groove, which lacks the distinctive features of the major groove), single base-pair replacements alter transcriptional activity. Recent crystallographic experiments have suggested that TATA/TBP complexes differing by a single base pair retain substantial structural similarity despite their functional differences in activating transcription. To investigate the subtle role of sequence-dependent motion within the TATA element and certain aspects of its effect on assembly of the transcriptional complex, we examine 5-ns dynamics trajectories of 13 variant TATA/TBP complexes differing from each other by a single base pair. They include the wild-type (WT) adenovirus 2 major late promoter (AdMLP) TATA element, TATAAAAG (the octamer specifies positions -31 to -24 with respect to the transcription initiation site), and the variants A31 (i.e., AATAAAAG), T30, A29, C29, G28, T28, T27, G26, T26, C25, T25, and T24. Our simulated TATA/TBP complexes develop sequence-dependent structure and motion trends that may lead to favorable orientations for high-activity variants (with respect to binding TFIIA, TFIIB, and other transcription factors), while conversely, accelerate dissociation of low-activity TATA/TBP complexes. The motions that promote favorable geometries for preinitiation complexes include small rotations between TBP's N- and C-terminal domains, sense strand DNA backbone "slithering," and rotations in TBP's H2 and H2' helices. Low-activity variants tend to translate the H1 and H1' helices and withdraw the intercalating phenylalanines. These cumulative DNA and protein motions lead to a spatial spread of complex orientations up to 4 Å; this is associated with an overall bend of the variant TATA/TBP complexes that spans 93° to 110° (107° for the crystal reference). Taken together, our analyses imply larger differences when these local structural and bending changes are extended to longer DNA (upstream and downstream) and suggest that specific local TATA/TBP motions (e.g., shifts in TBP helices and*

INTRODUCTION

The TATA box binding protein (TBP) is the central protein component stimulating the assembly of the preinitiation complex (PIC).[1,2] TBP binds to the consensus DNA octamer sequence TATA(t/a)A(t/a)N, where (t/a) indicates thymine or adenine, and N indicates any base.[3] The complex between the TATA element and TBP, determined by crystallography for several different organisms,[4–9] is unusual: the DNA is bent more than 90° while the protein interacts with the minor groove of the DNA primarily through nonspecific hydrophobic contacts with the bases. The deformation of this complex brings distant DNA segments closer to each other, erecting a scaffold for other transcription factors to establish and maintain simultaneous upstream and downstream contacts. Several ternary structures of TATA/TBP with additional transcription factors, such as TFIIA,[10,11] TFIIB,[12] and NC2,[13] reveal that different surfaces of TBP and the bent TATA promoter are exploited for contacting these other transcription elements.

While these tightly tailored structures have been insightful in formulating the basic model of PIC assembly, the modulation of transcriptional activity by the intrinsic promoter variation within the genome continues to be an active area of investigation.[14–18] Although the minor groove of DNA lacks specific functional groups to distinguish transversion variations (i.e., Pyr ⇔ Pur such as TA to AT), TATA/TBP complex activity is substantially affected by single base-pair (bp) variations.[19,20] The substantial alteration of TBP activity through DNA variations that do not alter minor groove functional groups suggests that TBP recognizes TATA elements indirectly though dynamic aspects (such as the energy of bending deformation), rather than average structure *per se.*[21–24] Indeed, a recent structural study by Burley and co-workers examining the details of TBP bound to 11 different single-bp DNA variants reveal a conserved structure (including similar overall bend) despite functional variation[14] (see Table I). Data from a recent fluorescence study examining the overall size and bending of a subset of these variant TATA/TBP complexes suggest that large sequence-dependent bending patterns exist, and that these differences might in turn modulate activity and PIC assembly.[15] In addition,

recent measurements of the binding constants between TBP and these variant TATA elements showed the tendency of variants with higher transcriptional efficiencies (TE) to have lower binding constants to TBP (than low-TE variants)[17]; some differences in the solvent accessibility of the TATA DNA between variant complexes were noted.[17,18] Taken together, the works suggest that sequence-dependent modulation of promoter activity depends in a complex way on a combination of factors such as specific motions, bending trends, binding energy/stability, and subtle structural variation of the complexes.

Molecular dynamics (MD) simulations have become a fundamental utility for complementing crystallographic and biochemical investigations.[25–27] The late Peter Kollman, to which this article is dedicated, developed not only the highly successful program AMBER,[28] but pioneered stable and accurate long-time nucleic acid simulations,[29] as well as free energy calculations that have important implications for drug design and the synergy of experiment and theory.[27]

Motivated by Kollman's works, we analyze here 13 5-ns trajectories of the variant TATA/TBP complexes examined experimentally[14,15,17] (Table I) with the state-of-the-art AMBER force field[30] to explore the potential role of DNA bending, motion, solvent, and structural changes in variant TATA/TBP complexes. This study complements our earlier investigation—the 13 variant TATA sequences without the protein[24]—which related characteristic DNA motions (such as increased roll/rise and minor groove widening at TATA ends), flexibility, and conformational preferences (such as reduced twist and increased bending into the major groove) to intrinsic activity.

Here we delineate local motion/flexibility trends that may promote favorable PIC complex formation in high-activity variants, and conversely, facilitate dissociation in low-activity variants. Though the TATA/TBP system is highly complex and simulations are inherently limited in their temporal and spatial sampling, the collective motions we analyze—rotation of TBP's helices and between TBP N- and C-terminal domains, DNA bending, DNA backbone "slithering," and "ratcheted" bp stacks—produce differing geometries that are likely to affect the binding of other transcription factors, such as TFIIA, TFIIB, TAFs, and RNA polymerase.[1,2]

Table I   DNA Sequences, Their Transcriptional Efficiencies, and the TBP[14] Used in These Studies[a]

Label (PDB)	Sequence			Efficiency (%)
WT (1QNE)	GC	(−31)T A T A · A A A G(−24)	GGCA	100
A31 (1QNC)	GC	[A] A T A · A A A G	GGCA	14
T30 (1QNA)	GC	T [T] T A · A A A G	GGCA	25
A29 (nc)	GC	T A [A] A · A A A G	GGCA	<1*
C29 (1QN9)	GC	T A [C] A · A A A G	GGCA	20
G28 (nc)	GC	T A T [G] · A A A G	GGCA	<1
T28 (1QN8)	GC	T A T [T] · A A A G	GGCA	14
T27 (1QN7)	GC	T A T A · [T] A A G	GGCA	35
T26 (1QN6)	GC	T A T A · A [T] A G	GGCA	6
G26 (1QN5)	GC	T A T A · A [G] A G	GGCA	18
C25 (1QN3)	GC	T A T A · A A [C] G	GGCA	6
T25 (1QNB)	GC	T A T A · A A [T] G	GGCA	100
T24 (1QN4)	GC	T A T A · A A A [T]	GGCA	40*
TBP		(16)K H P S G I V P T L Q N I V S T V N L D C K L D L K A I A L Q A R N A E Y N P K R F A A V I M R I R E P K T T A L I F A S G K M V C T G A K S E D F S K M A A R K Y A R I V Q K L G F P A K F K D F K I Q N I V G S C D V K F P I R L E G L A Y S H A A F S S Y E P E L F P G L I Y R M K V P K I V L L I F V S G K I V I T G A K M R D E T Y K A F E N I Y P V L S E F R K I(198)		

[a] 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 (WT). The four-letter PDB entry codes[33] are indicated in the first column; variants that were not crystallizable (nc) are also indicated. Single-position variants (boxed characters) are indicated relative to WT, and labeled according to the replaced base and position with respect to the transcription initiation site. Transcriptional efficiencies (TEs) for A29 and T24 (marked by asterisks) are based on Bernués *et al.*[20] and Wobbe and Struhl,[19] respectively. The 183 residue fragment of *A. thaliana* TBP (residues 16–198) used in these simulations indicated at bottom is common to all eleven crystallographic structures.[14]

We find that, in high- and medium-activity variants, the two domains of TBP rotate relative to each other, establishing a possibly optimal interdomain orientation. Specific helices of TBP (H2/H2′; see Figure 5) are also rotated in high-activity variants to potentially affect interactions with TBP associated factors (TAFs)[31,32] and TFIIA.[10,11] Several other subtle motions emerge as likely important for activity, including deoxyribose/phosphate backbone slithering within the TATA octamer and associated stretching within the TATA element bases.

In low-activity variants, the H1 helix tends to translate and rotate, altering an interaction site for TFIIA.[10,11] Low-activity variants reveal a dissociation pathway, consisting of increased DNA twist and withdrawal of intercalated phenylalanines. Other conformational changes within low-activity TATA elements lead to unique bp rearrangements and ratcheted structures with slipped bp stacks. Although the DNAs in all systems remain highly bent, bending differences lead to large overall changes in the modeled positions of the +1 initiation site and upstream transcription factor binding sites. Throughout the simulations, we observe unusual water interactions within the TATA/TBP interface and salt interactions with the TATA bp stack; these solvent and ionic interactions may be particularly important to the dissociation pathways in low-activity variant structures.

METHODS

Background

The 13 trajectories we analyzed are 14-bp duplexes containing single-bp variants of the TATA octamer complexed with TBP as defined in Table I. These simulations complement an earlier study of the same duplexes without the protein[24] from which sequence-dependent motions that tailor DNA for interaction with TBP were identified. We employ both simulation sets to compute hydroxyl radical footprint differences between the TATA DNA and the TATA/TBP complexes[17] to validate the theoretical structures (as described in the accompanying appendix).

System Preparation and Simulation Method

Crystallographic structures of the duplexes complexed to TBP were obtained from Dr. Stephen K. Burley[14] and the Protein Data Bank[33]; only one of the two complexes was retained from each asymmetric unit. For two variants (A29 and G28), crystallographic structures were not available; initial models were prepared by replacing the appropriate bp of the wild-type AdMLP structure by the variant bp. Hydroxyl groups were added to the 5′ and 3′ termini of the DNA as necessary. The protein was truncated to a 183-residue contiguous fragment present in all crystal structures (Table I, bottom). Various residue sidechains not resolved in the crystallographic structure (typically surface lysines and arginines) were attached to the corresponding C^α positions in an all-*trans* conformation using InsightII (Molecular Simulations, 2000). Hoogsteen conformations at G_{-24}–$C_{-24'}$ reported for several crystallographic complexes were judged to be artifactual[14] (unlike the Hoogsteen conformation at $G_{-24'}$ in the C25 variant), stabilized by interactions from an adjacent asymmetric unit. When establishing an initial conformation for these complexes, the Hoogsteen conformations at G_{-24} were replaced by a standard Watson–Crick geometry by changing the glycosidic angle χ using InsightII from $\approx 70°$ to $\approx -120°$. These models were then imported into CHARMM[34] version $26\alpha2$ and briefly optimized *in vacuo* with a $1/r$ dielectric function in two stages with the AMBER PARM94 force field[30]: first, only the bases of $G_{-24'}$ and $C_{-24'}$ were allowed to move during 1500 cycles of steepest descent minimization; second, only the bases and deoxyribose backbone of G_{-24} and $C_{-24'}$ were allowed to move during 15,000 cycles of adopted basis Newton–Raphson optimization.

The initial orientations of the TATA/TBP models were optimized using our program PBCAID[35] within a rectangular prism periodic domain to increase the distance between the model system and its periodic images. The initial water coordinates were also generated within PBCAID 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. Water molecules within 2.0 Å of either the TATA/TBP heavy atoms or the crystallographic waters were removed using CHARMM, leaving a net total of $\approx 12,300$ TIP3P water molecules (including crystallographic waters). The rectangular prisms of each of the 13 solvated complexes measure 77.4 Å in length, 74.1 Å in width, and 69.7 Å in height. The number of ions included to establish a neutral cell with an ionic strength of 150 mM are 37 sodium and 26 chloride ions; the ions replace water molecules at regions of high negative (or positive) potential as calculated by a finite difference solution to the nonlinear Poisson–Boltzmann equation using the Delphi package;[36] each ion was placed so that it is more than 9 Å distant from other ions and the complex. The final system sizes totaled $\approx 41,000$ atoms (WT:41011; A31: 40903; T30: 40858; A29: 41014; C29: 40938; G28: 41010; T28: 40891; T27: 40891; T26: 40888; G26: 40896; C25: 40944; T25: 40942; T24: 40937).

Periodic boundary conditions and the AMBER PARM94 force field[30] converted for use in CHARMM[34] version $26\alpha2$ are used for all energy minimizations and MD simulations. Nonbonded interactions are truncated at 12 Å, with force shift for electrostatic and potential-switch for van der Waals interactions.[37]

Energy minimization of the system was divided into three stages: minimization of only bulk water, minimization of all waters and ions, and minimization of the entire system. Initially, the coordinates of the heavy atoms of the DNA, protein, ions, and crystallographic waters were fixed; only bulk waters and hydrogen atoms were minimized using 1000 steps of steepest descent and an adopted-basis Newton–Raphson protocol for 4000 steps. In the second stage, the ions were released for 6000 steps of adopted-basis Newton–Raphson minimization. Each system was then briefly relaxed during 24 ps of our multiple-timestep Langevin integrator LN based on a normal-mode conceptual framework[38,39] with a timestep protocol $\Delta\tau/\Delta t_m/\Delta t$ of 1/2/120 fs for fast/medium/slow force components using SHAKE constraints, short-range nonbonded cutoffs of 7 Å (used to define the medium class), and healing and bookkeeping lengths of 4 and 4 Å, respectively; as mentioned above, the long-range cutoff was maintained at 12 Å. The LN integrator[38,39] allows simulations with greater computational efficiency; see Ref. 40 for detailed verification of LN on large systems. LN uses large outer timesteps (120 fs in this study) to compute the slow force components, thereby reducing the number of nonbonded interaction updates and the corresponding computational cost. Medium-range nonbonded interactions are handled by a smaller timestep (2 fs), while bonded forces are updated on a short inner timestep (1 fs). The long-range nonbonded interactions in the TATA/TBP simulations were calculated using periodic boundary conditions; recent work[41,42] has extended our algorithms to include the particle-mesh Ewald method pioneered by Kollman, Darden, and others.[29,43] The LN integrator is available in both CHARMM[34] and AMBER.[28]

After the brief relaxation, the entire unconstrained system was then optimized with 6000 steps of adopted-basis Newton–Raphson minimization. The systems are then equilibrated with 3 ps of Newtonian dynamics using a position Verlet integrator with a timestep of 0.5 fs without SHAKE constraints, followed by 24 ps of Langevin dynamics with our LN integrator with SHAKE constraints using the timestep protocol indicated above. The final coordinate set was used as the initial structure for the production cycle of each of our systems.

For the production dynamics runs, SHAKE constraints were applied to all bonds with hydrogens. Coordinates were saved every 1.2 ps, and the last 4.7 ns of the trajectories was used for data analysis. Each 0.36-ns segment of the 5.04 ns trajectory took 11 days (244 h) on four 300-MHz R12000 processors of the NYU SGI Origin 2000 computer, or 4 days (96 hours) on sixteen 195-MHz R10000 processors of the NCSA Origin 2000 cluster.

Structure Analysis

Nucleic acid structural parameters were derived from the analysis program RNA of Olson and co-workers.[44–46] Although we have used Curves[47,48] in previous publications,[24,49] Curves was unable to analyze the unusual stacking geometries we encountered in many complexes here, such as Hoogsteen conformations and slipped bp stacks.

Bending and Flexibility of DNA.   We were not able to compute bending using our program MadBend and the angles global roll and global tilt,[24,49] because the standard bending angles [roll (ρ), tilt (τ), and twist (Ω)] are not solely sufficient to describe the bending of unusual structures such as the ratcheted bp stacks. As an alternative, we computed bending as an angle between normals to the three-dimensional curve described by the TATA DNA helical axis. Namely, long pieces of *B*-form DNA (29 and 24 bp, respectively) were matched to the 5′ and 3′ ends of the TATA DNA through overlapping sequences at each end and superimposing the long pieces to minimize the RMS difference between the overlapping sequences. Specifically, at the 5′ end, 2 bps (-33 and -32) were used for the overlapping large ends; at the 3′ end, 3 bp (-23 through -21) were used. The resulting 61-bp DNA has both the highly bent TATA/TBP complex at the central bps 28–41 with idealized straight DNA extending to the downstream $+1$ initiation and the upstream -60 promoter sites. For each snapshot in each trajectory, we computed the center of mass of five regions: (1) the first four bps of the long 5′ segment; (2) bps -33 through -30 of the TATA complex; (3) bps -29 through -26 of the TATA complex; (4) bps -25 through -22 of the TATA complex; and (5) the last four bps of the long 3′ segment. We defined two planes and two normals to the planes using regions $\{1, 2, 3\}$ and regions $\{3, 4, 5\}$. The bending angle (a dihedral angle) was computed as the scalar product of the normals. With this algorithm, the bend of the WT cocrystal complex is $\approx 106.5°$, while our program MadBend computes a bend angle of $\approx 117.0°$. The bending flexibility of variant i (ϕ_i) is computed as the standard deviation of the bending angle:

$$\phi_i = \sqrt{\langle (\text{bend}_i - \langle \text{bend}_i \rangle)^2 \rangle}.$$

Analysis of Salt, DNA/Protein Interface, and Water Interactions.   Ion interactions, protein–DNA interactions, and water interactions were monitored using standard utilities available in CHARMM.[34] Ion binding events were computed by monitoring ions approaching the TATA DNA major groove within 4 Å, i.e., one water diameter; the end bps of the TATA DNA were not included. Protein–DNA interactions were monitored as distances between specific side-chain and DNA functional groups; no distance cutoff was used. Lifetimes of water molecules located in the interface between the TATA element and TBP were computed as a joint probability: by monitoring distances between the water molecules and selected TBP side-chain and TATA minor groove atoms, waters simultaneously within

cutoff distances to both atom sets were determined to be in the interface. The cutoff distances used depended on the specific geometry of nearby TATA and TBP atoms, but ranged from 4.5 to 6.0 Å, with an average of 5.2 Å; in contrast, the bulk waters in the TATA major groove are ≈ 8–9 Å from TBP. Since water molecules are moving around the interface, coordination events involving the same water molecule in the same site separated by less than 5 ps were counted as a single coordination event.

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

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

Confidence limits at a 95% significance level ($P < 0.05$) are estimated by linear regression. Properties analyzed this way include global bending and flexibility (Figure 2), rotation of domains and helices (Figure 5), and the mean square magnitude of motion along a principal component (Figures 8–10), water lifetimes (Figure 11), and ion contacts (Figure 12).

Motion analysis by Principal Component Analysis

Principal Component Analysis (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)[24,50] and proteins.[51] We use PCA applied to a merged trajectory of all variants (a procedure that we term "uniform ensemble PCA") to directly compare the different magnitude of each motion among our 13 variants; additional details are available below.

Our PCA procedures have been applied only to heavy (nonhydrogen) atoms of the TATA/TBP complex; hydrogens are ignored. A total of 2020 atoms are included in the analysis, resulting in 6060 PCs (3×2020). Snapshots are sampled from the last 4.68 ns of each trajectory at a frequency of $\Delta t = 1.2$ ps. In this study, our cutoff value of 100 PCs includes $\approx 89\%$ of the ensemble motion; in the earlier TATA simulations without TBP,[24] the top 100 PCs included $\approx 95\%$ of the ensemble motion.

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: $\langle \mathbf{X} \rangle = \frac{1}{M} \Sigma_{k=1,M} \mathbf{X}_k$. The average structure used as a refer-

ence to develop the covariance matrices $\mathbf{C}$ is the unminimized coordinate average. Diagonalization of $\mathbf{C}$ produces the eigenvalues and eigenvectors as entries of $\mathbf{\Lambda}$ from the decomposition:

$$\mathbf{V}^T\mathbf{C}\mathbf{V} = \mathbf{\Lambda}, \quad \text{or} \quad \mathbf{C}\mathbf{V}_n = \lambda_n\mathbf{V}_n, \quad n = 1,2,\ldots,3N,$$

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

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 compare motions between different TATA/TBP complexes, we must ensure comparable numbers of atoms between different variants and then merge the different trajectories. To make the trajectories comparable, we set the WT sequence to be the reference and convert each variant to resemble the WT sequence:

1A. If the variant $\rightarrow$ WT conversion is a *pyrimidine to pyrimidine* replacement (e.g., C to T, as in C29 to WT) or a *purine to purine* replacement (e.g., G to A, as in G26 to WT), the phosphate/deoxyribose backbone atoms and all nonhydrogen atoms of the pyrimidine and purine rings of the mutated bp 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 $\rightarrow$ WT conversion is a *pyrimidine to purine* replacement (e.g., A to T, as in A31 to WT) or a *purine to pyrimidine* replacement (e.g., T to A, as in T27 to WT), the phosphate/deoxyribose backbone atoms are again maintained. Nonhydrogen atoms of the 5-membered purine ring or the 6-membered pyrimidine ring are used to replace the bases according to the superimposed positions of a purine and a pyrimidine, maintaining the planar orientation of the original base. The remaining hydrogen atoms are built using standard geometries, and the nucleotides are accordingly renamed.

2. After the above bp replacement and adjustment, all atoms except those rebuilt from standard geometries are fixed. A short minimization (10 steps of steepest descent and 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 bp geometries are maintained, and the average relative error between local bp step parameters before and after replacement is less than 2%. To eliminate translational and rotational motion in the merged trajectory, the average structure of the merged trajectory (50700 total frames from 13 variant trajectories of length 4.68 ns with 3900 frames each) is used to orient each frame of the trajectory to minimize the RMSD of the TATA/TBP complex. The reoriented merged trajectory then produces a second average structure. The above process is repeated approximately 7 times until the average structure converges between cycles and no rotations or translations are necessary to minimize the RMSD.

Structure Generation Using PCs. An arbitrary structure $\mathbf{Y}$ can be generated from the average structure $\langle\mathbf{X}\rangle$ by a displacement $\mathbf{D}$ 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 the analysis procedures that we describe below, such as measuring the motions of single and combined PCs. Namely, single PCs (e.g., PCs 1–11 in Figure 9) are analyzed by determining the structural deformations associated with the eigenvectors. A scalar α_n corresponding to the deformation $\mathbf{D}$ 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 $\mathbf{D}$ is used to generate 11 structures. The structures may then be analyzed with standard programs such as RNA,[44–46] Curves,[47,48] or animated in visualization packages such as Insight.

PC Analysis by Relative Magnitude of Motion. Following uniform ensemble PCA, we compare for each variant i the normalized mean square magnitude of the projection along PC n, $\overline{\alpha_{n,i}^2}$:

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

where M_i is the number of trajectory frames of variant i, $\text{Tr}(\mathbf{\Lambda})$ is the sum of covariance matrix eigenvalues $\text{Tr}(\mathbf{\Lambda}) = \Sigma_{n=1,3N}\lambda_n$ and $\alpha_{n,i}^k$ is the projection of sequence i on PC n at frame k. This use of this method is illustrated in Figure 8 and in the selection of significant PCs for discussion.

RESULTS AND DISCUSSION

Overall Structural Deformations

Analyses of RMS deviations of each complex (average structure from final 0.36 ns segment of each variant's trajectory) reveal no large structural changes in either the protein or DNA components of the com-
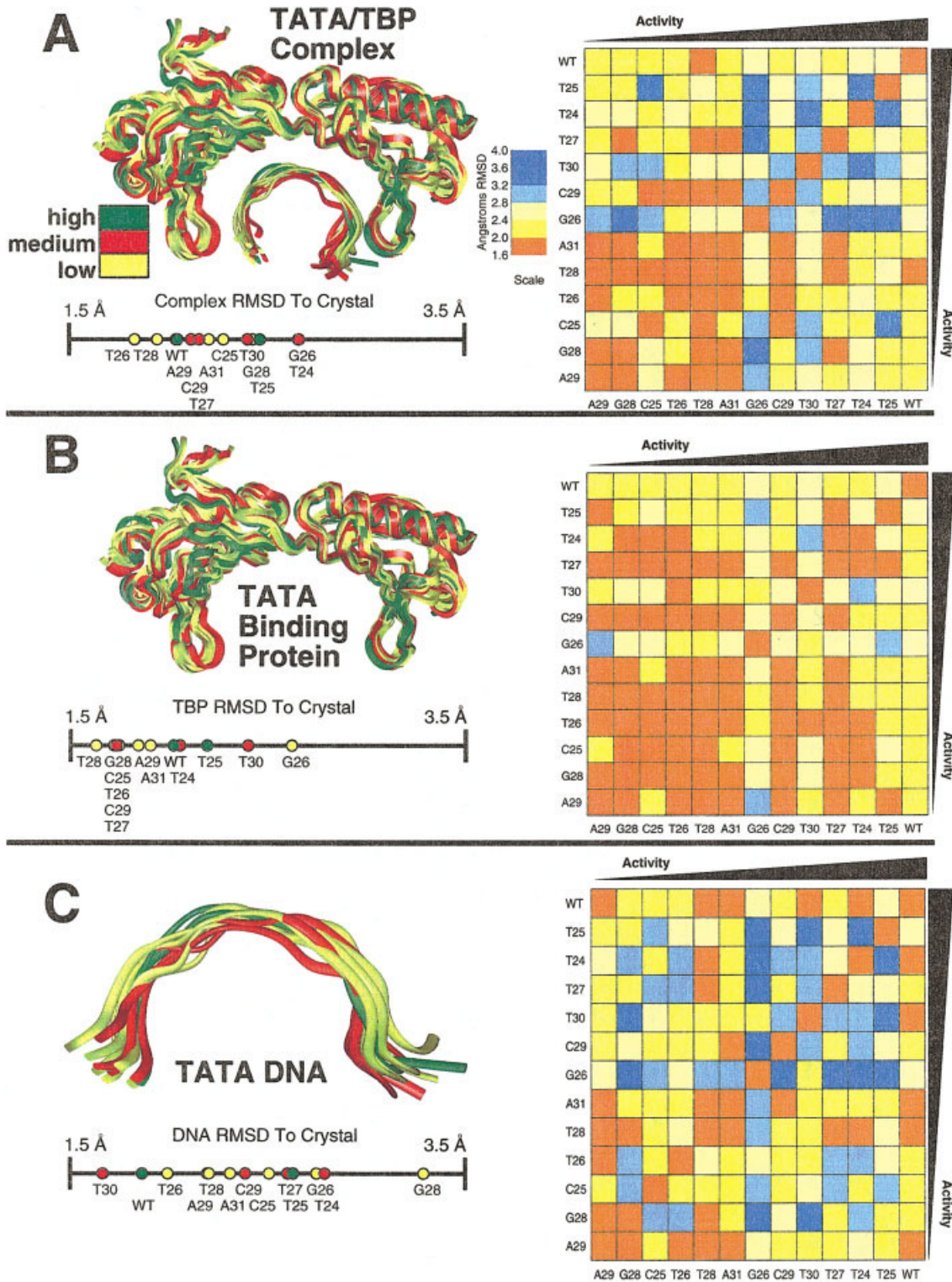
FIGURE 1 Average structures of the simulated TATA/TBP complexes. (**A**) Superimposed average structures of the TATA/TBP complexes computed over the final 0.36 ns. Protein structures are illustrated with ribbons drawn following the C_α trace; the nucleic acid strand is illustrated with a ribbon intersecting the purine N1 and pyrimidine N3 atoms of the sense strand. The ribbons (and other figures) use a color-coding system to distinguish among variants with high-TE (TE $\geq$ 80% WT; green), medium-TE (80% $<$ TE $\geq$ 20%; red), low TE (TE $\leq$ 20%; yellow). The RMS

plex (Figure 1). The RMS values for the superimposed complex, the protein alone, and the DNA alone, show that the average conformations are within 2–3 Å RMSD of the initial crystallographic structures.

Still, the pairwise RMS analysis suggests that complexes with minimal activity (such as A29, G28, C25, T26, T28, and A31 with TEs ranging from 0 to 14%) exhibit more similarity with each other compared to the more diverse family of higher-activity complexes (such as G26, C29, T30, T27, T24, T25, and WT with TEs ranging from 18 to 100%).

We have separated our results below into two groupings focusing on the activity of the TATA/TBP complexes: structures and motions associated with low-activity variants, and structures/motions associated with high- and medium-activity variants. Within these groupings, we discuss the different DNA, protein, and interface structures and motions associated with activity. Since specific structural details are determined to be important for either high-activity or low-activity variants (such as H2/H2′ rotation for high-activity variants), the discussion sections are not precisely parallel. Following these two separate sections, we analyze the water and ion interactions associated with these structural and dynamic features.

Low-Activity Complexes

DNA Structure.

Local DNA Bending. Overall, we find no large changes in the bending geometry of the TATA/TBP complex: the TATA DNAs all remain highly bent, with average bends over all snapshots in the 13 ensembles decreasing only 10° from the initial value of $\approx 107°$. In Figure 2, we show the average bending difference from the initial structure computed over the final 4.7 ns of the simulations, and the standard deviation of the bending as a measure of the complex flexibility. The correlation coefficient κ (defined in Methods) indicates that the slight increase in bending is not significantly correlated with activity ($\kappa = -0.13 \pm 0.1$). The spread in bend angles of the variant complexes ($\approx 15°$) is comparable to the spread in bend angles measured for the free TATA DNAs,[24] highlighting the intrinsic bending proclivities of TATA sequences.

The overall bend is important because a recent study of the end-to-end distance in fluorophore-labeled TATA/TBP complexes by Parkhurst *et al.*[15] suggested that activity and the bend angle are correlated: low-activity variants were found to bend less than high-activity variants. This fluorophore study used 6 TATA variants, also studied here: A29, C25, G26, T26, T27, and WT. Our study does not confirm this bend/activity relationship; and, given the limited sampling of all nanosecond-scale MD simulations, our data neither support nor refute the FRET experimental findings.[15]

The bending flexibility ϕ, calculated as the standard deviation of the bending angle, is notably higher for some low and medium-TE variants, such as C29 (5′-TA**C**AAAAG-3′, $\phi_{C29} = 8.3°$) and A29 (5′-TA-**A**AAAAG-3′, $\phi_{A29} = 6.4°$) than for the high-TE WT and T25 (WT:5′-TATAAAAG-3′, $\phi_{WT} = 4.9°$; T25: 5′-TATAAA**T**G-3′, $\phi_{T25} = 3.8°$); overall, the bending flexibility of the TATA DNA is moderately significant to low-activity variants ($\kappa = -0.42 \pm 0.07$). This result agrees with observations from fluorophore studies that low-activity variants have larger standard deviations of the end-to-end DNA distance.[15]

We interpret this increased flexibility in low-TE variants as a tendency of these variants to dissociate from TBP, since low- and medium-TE variants may have binding constants smaller than high-TE variants.[17] We also conjecture that the increased flexibility and bending of low and medium-TE variants promotes the release of TATA DNA from the complex, possibly through an intermediate form that is more bent than the cocrystal.

DNA Bending Within the Larger PIC Region. These local bending trends are quite significant if extrapolated to larger DNAs approaching transcriptionally important regions. This can be seen in models that extend the TATA/TBP complex DNA from the +1 initiation site to the −60 upstream position by adding large idealized DNA segments to each end (Figure 3). If the DNA between the 3′ end of the TBP complex and initiation site is approximately straight, our modeling suggests that the cumulative effect of the local bends can spread over a 60–70 Å range in the upstream promoter and downstream initiation sites.

These extrapolated models also suggest that the

FIGURE 1. (*Continued from the previous page.*) difference between the final average structure and the initial crystal structure is plotted below the superimposed structures. The pairwise variant/variant RMS difference is plotted to the right of the superimposed structures. (**B**) Superimposed average structures of TBP component of the complexes, RMS difference to the initial crystal, and pairwise variant/variant RMS difference (**C**) Superimposed average structures of TATA component of the complexes, RMS difference to the initial crystal, and pairwise variant/variant RMS difference.
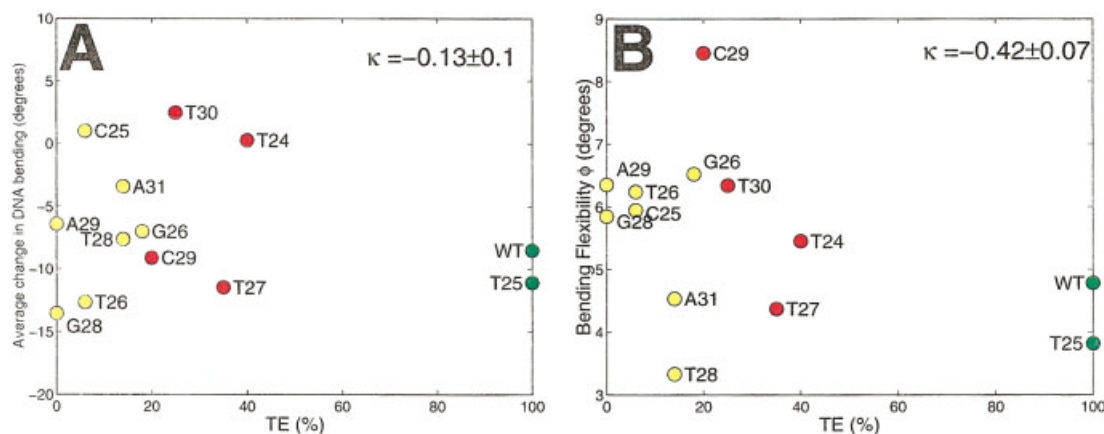
FIGURE 2 Average bending and bending flexibility in TATA/TBP complexes. (**A**) The average change in bending from the WT cocrystal for each variant is plotted against the transcriptional efficiency (TE) of each variant. The average bending change for each variant is computed over 3900 snapshots sampled during the final 4.68 ns of each trajectory. (**B**) The bending flexibility ϕ is plotted against each variant's TE. The flexibility is computed over each variant's trajectory as described above.

upstream and downstream regions of low-activity variants converge to a relatively common spatial region spanning less than 14 and 21 Å, respectively (Figure 3), in accordance with the small RMSD values among the group of low-activity variants (Figure 1). In contrast, the average positions of the upstream/downstream regions of high and medium-activity variants are spread over wider ranges (Figure 3). This convergence of the modeled upstream and downstream regions in the low-TE variants suggests the presence of a common conformation essential to either inactivity or complex dissociation.

Unusual DNA Structures. Although the TATA element is very distorted in the complex,[4,5,14] we observe additional unusual DNA features such as Hoogsteen bps and "ratcheted" bp stacks during the simulations of several low-activity variants.

Hoogsteen Base Pair in C25. The cocrystal structure of the C25 variant (5′-TATAAACG-3′) adjusts to the N2 amine group of guanine at position −25′ by transforming the guanine to a Hoogsteen geometry.[14] This Hoogsteen bp was preserved during the simulations (although in *an unprotonated state* to test its stability); we note that $G_{-25'}$ is stable, although it participates in a strong bifurcated hydrogen bond network with A_{-26} (Figure 4A).

Slipped Base-Pair Stacks and "Ratcheted" Structures. An interesting feature of the cocrystal structures is the strong bifurcated hydrogen-bond interactions between adjacent bp steps in the TATA box,[4,5,14] as shown in Figure 4B for the WT cocrystal. In four of our simulations, some of the hydrogen

bonds between adjacent bp steps (Figure 4B, red bonds, normally greater than 4 Å) shorten to a distance typical of normal Watson–Crick basepaired geometries ($\approx$3 Å), while the Watson–Crick hydrogen bonds (Figure 4B, blue bonds) lengthen to more than 3 Å. This pattern extends for several bps, and can be explained by a slipped or ratcheted bp (by one base), such that the adenine bp *j* normally partnered with thymine bp *j*′ now has Watson–Crick hydrogen bonds with bp (*j* − 1)′ One base is left at each end of the slipped region with no complementary partner.

The overall conformational transition leading to this slipped arrangement only requires motions of magnitude 1–2 Å for of the bases, well within the thermal range accessible to the bifurcated hydrogen bond geometries. Significantly, the two complexes that develop the largest ratcheted structures, namely T24 (5′-TATAAAA**T**-3′) and A29, have medium- and low-activity values of 40 and 0%, respectively. In both A29 and T24, the ratcheted structure stretches over four bps, from A_{-30} to A_{-27} in the A29 variant and from A_{-28} to A_{-25} in the T24 variant. A third medium-activity variant, C29, develops the smallest possible ratcheted bp stack, stretching from A_{-28} to A_{-26}. The low-activity C25 variant develops a smaller but similar bp rearrangement, localized around a Hoogsteen bp (described above). While changes in the structure of the bp stack might affect transcription factors like TFIIB[12] and NC2[13] that contact the TATA major groove, the observed pattern of DNA structural changes in only a few low- and medium-activity variants suggests that this is likely to be
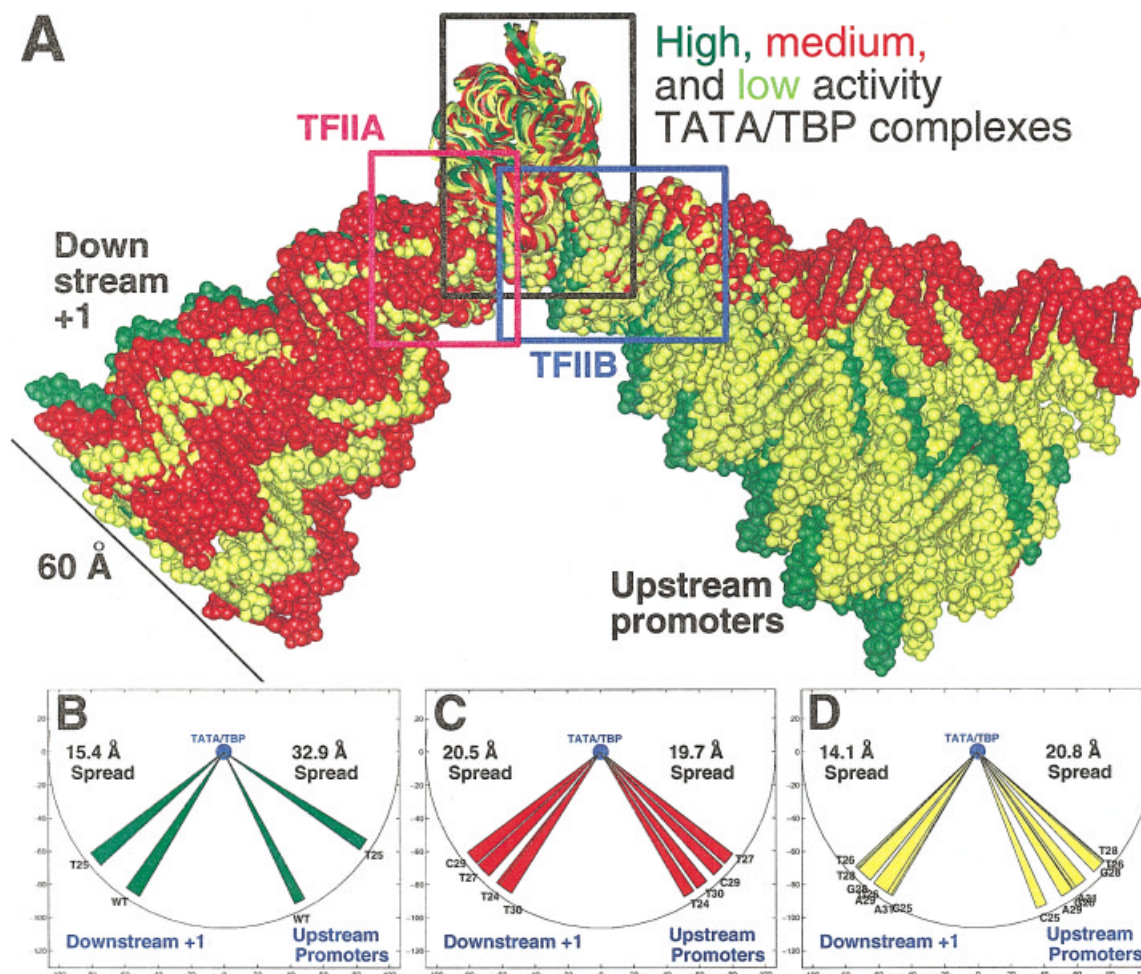
FIGURE 3   Models of long DNAs with variant TATA/TBP promoter complexes from the +1 initiation site to the −60 upstream promoter region. (**A**) Models of long DNAs from the upstream promoter to the downstream +1 region. Idealized straight *B*-DNA segments of length 29 and 24 bp were superimposed on the 5′ and 3′ ends, respectively, of the average structure from the final 360 ps of each variants 5.04 ns trajectory. (**B**) The mean distance from the average position of the modeled downstream +1 region and the upstream −60 promoter region over the final 360 ps of the high-activity variant trajectories. The fans illustrates the range of distances (scaled by a factor of 5) from the average position for each variant. (**C**) The mean distance from the average position of the modeled downstream +1 region and the upstream −60 promoter region over the final 360 ps of the medium-activity variant trajectories. (**D**) The mean distance from the average position of the modeled downstream +1 region and the upstream −60 promoter region over the final 360 ps of the low-activity variant trajectories.

a promoter-specific mechanism of either transcription regulation or dissociation from TBP.

Protein Structure. Since TBP serves as a central locus for the gathering of transcription factors and PIC assembly,[1,2] large rotations and translations of α-helices and of the N- and C-terminal domains may modulate the transcriptional activity by locally altering binding and stability propensies for complexes between TBP and other transcription factors. The structure of TBP is dominated by a broad β-sheet, which forms the N- and C-terminal saddles (Figure 5A). Regions where this β-sheet narrows to just two β-strands define the the C-terminal stirrup, the central hinge region, and the N-terminal stirrup. We monitor the rotation between TBP's N- and C-terminal domains using these five β-sheet regions; in addition, independent motions of the H1/H1′ and H2/H2′ α-helix subdomains, which respectively interact with TBP's stirrup and saddle subdomains, are followed
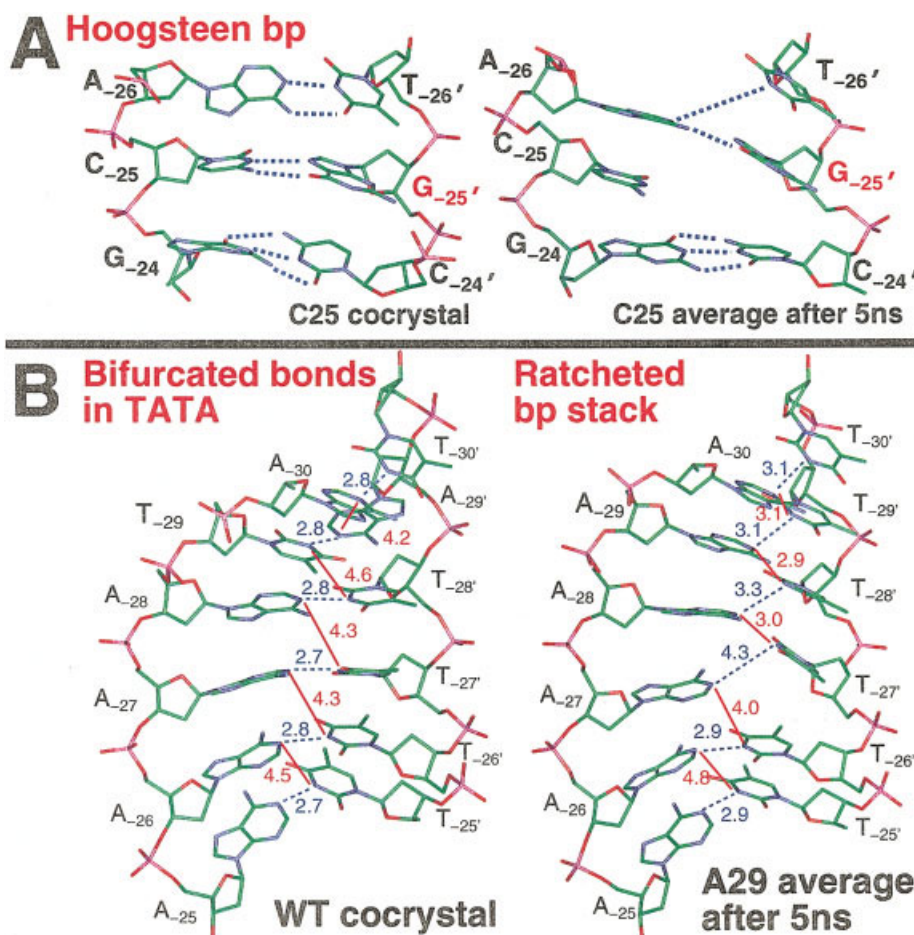
Strahs *et al.*

FIGURE 4 Structural changes in the TATA DNA component of simulated complexes. (**A**) The Hoogsteen bp at position −25 (red) in the cocrystal (left) and after 5 ns of MD simulation (right). (**B**) A typical bp stack observed in most of the TATA/TBP cocrystals and a bp stack from the modeled A29 TATA/TBP complex after 5 ns of MD simulation. The lengths in Ångstroms of normal Watson–Crick (blue) and bifurcated (red) hydrogen bonds are indicated. Note that the bifurcated hydrogen bonds of the A29 bp stack (red) have lengths typical of normal Watson–Crick hydrogen bonds, while the Watson–Crick bonds (blue) are longer.

through the orientation of the respective helical axes to these β-sheet regions. In Figure 5, we illustrate TBP domain and subdomain conformational changes in turn: rotations and lateral translations of the H1 and H1′ helices as observed in low-activity variants; and rotation between the N-terminal and C-terminal domains and of the H2 and H2′ helices in high/medium-activity variants.

H1/H1′ Rotation/Translation. The H1 and H1′ helices, important interaction sites for transcription factors such as TFIIA[10,11] and TFIIB,[12] are exposed to solvent at the ends of the TBP domains and are sensitive to motions like domain rotation and DNA twisting. We note from Figure 5C, which analyzes the H1/H1′ helices' translation and rotation (relative to

each variant's initial cocrystal structure), that motions in the H1 helix increase with decreasing activity (κ_{rot} = −0.4 ± 0.1, κ_{trans} = −0.5 ± 0.01). These motions are larger than the corresponding motions associated with the H2/H2′ helices (discussed below). While the H1′ helix also displays comparably large rotations and translations, these motions have weaker correlations with variant activity (κ_{rot} = 0.2 ± 0.08, κ_{trans} = −0.07 ± 0.01). The structural changes in the H1 helix likely affect TFIIA interactions[10,11] and may be linked to the formation of the TFIIA/TATA/TBP complex, while the structural changes in the H1′ helix alter the TBP/TFIIB interface.[12]

Domain Rotation and H2/H2′/Rotation. We find that the specific motions of TBP's N- and C-terminal
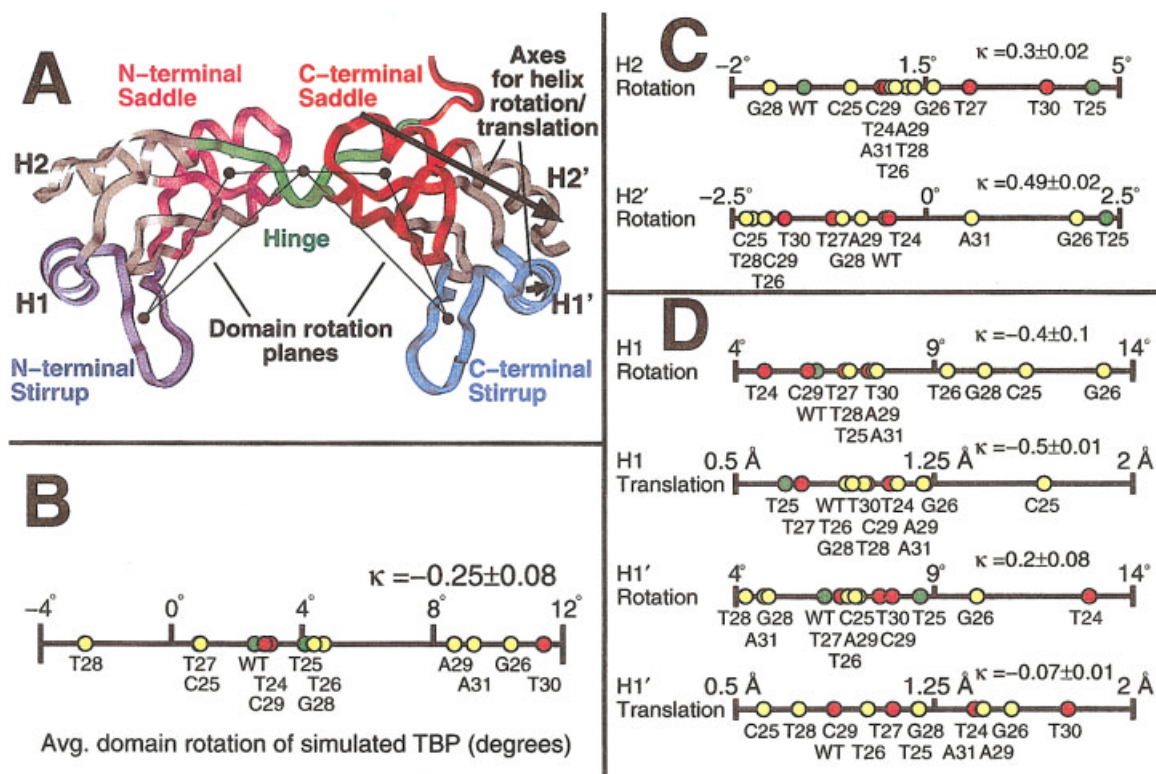
FIGURE 5 Structural changes in the TBP protein component of simulated complexes. (**A**) Crystal structure of TBP. The five regions and their associated centers of mass used to compute the overall rotation between the N-terminal and C-terminal domains are indicated by color and the accompanying black dots, respectively; the two planes determined by the centers of mass used to compute the overall domain rotation are indicated by cross-hatched areas. An earlier study used a comparable torsional measure[52]; however, our definitions exclude the H2 and H2′ helices, which shift their orientation independent of the larger N-terminal/C-terminal domains. The four α-helices (H1, H1′, H2, and H2′) of TBP and an example of their helical axes are also shown. The details of the calculation of the structural changes are discussed in the text. (**B**) Average domain rotation between the N-terminal and C-terminal domains over the final 4.68 ns of the trajectories. The domain rotation is measured as a dihedral angle defined between two planes: one plane is defined by the centers of mass of the N-terminal stirrup, saddle, and hinge region, and the other plane by the centers of mass of the C-terminal stirrup, saddle, and hinge region. (**C**) Average rotation of helices H2 and H2′ over the final 4.68 ns of the trajectories. Helix rotation is measured as the angle between the H2 (or H2′) helical axis and a vector determined from the coordinates of the supporting saddle β-sheet underneath that helix. (**D**) Average rotation and translation of helices H1 and H1′ over the final 4.68 ns of the trajectories. Helix rotation is measured as the angle between the H1 (or H1′) helical axis and a vector determined from the coordinates of the supporting stirrup β-sheet underneath that helix; translation is measured as the projection of the translation vector of the helix's center of mass against the initial cocrystal's helical axis.

domains and of the H2/H2′ helices are only significant for high- and medium-activity variants, as discussed below.

DNA/TBP Interface. The interface between the TATA element and TBP is characterized by an unusual set of specific protein/DNA interactions. Four phenylalanines intercalate into the DNA[4,5] (two each into the first and last TATA bp steps), the minor

groove of the DNA is primarily recognized by hydrophobic protein side chains through van der Waals contacts, and the DNA backbone interacts with a set of cationic and hydrophilic residues. For reference, Figure 6 shows the location and types of interactions (i.e., hydrophobic, polar, or electrostatic) between side chains and the DNA. In Figure 7, we illustrate the mean distance between the interacting protein and DNA functional groups with corresponding standard
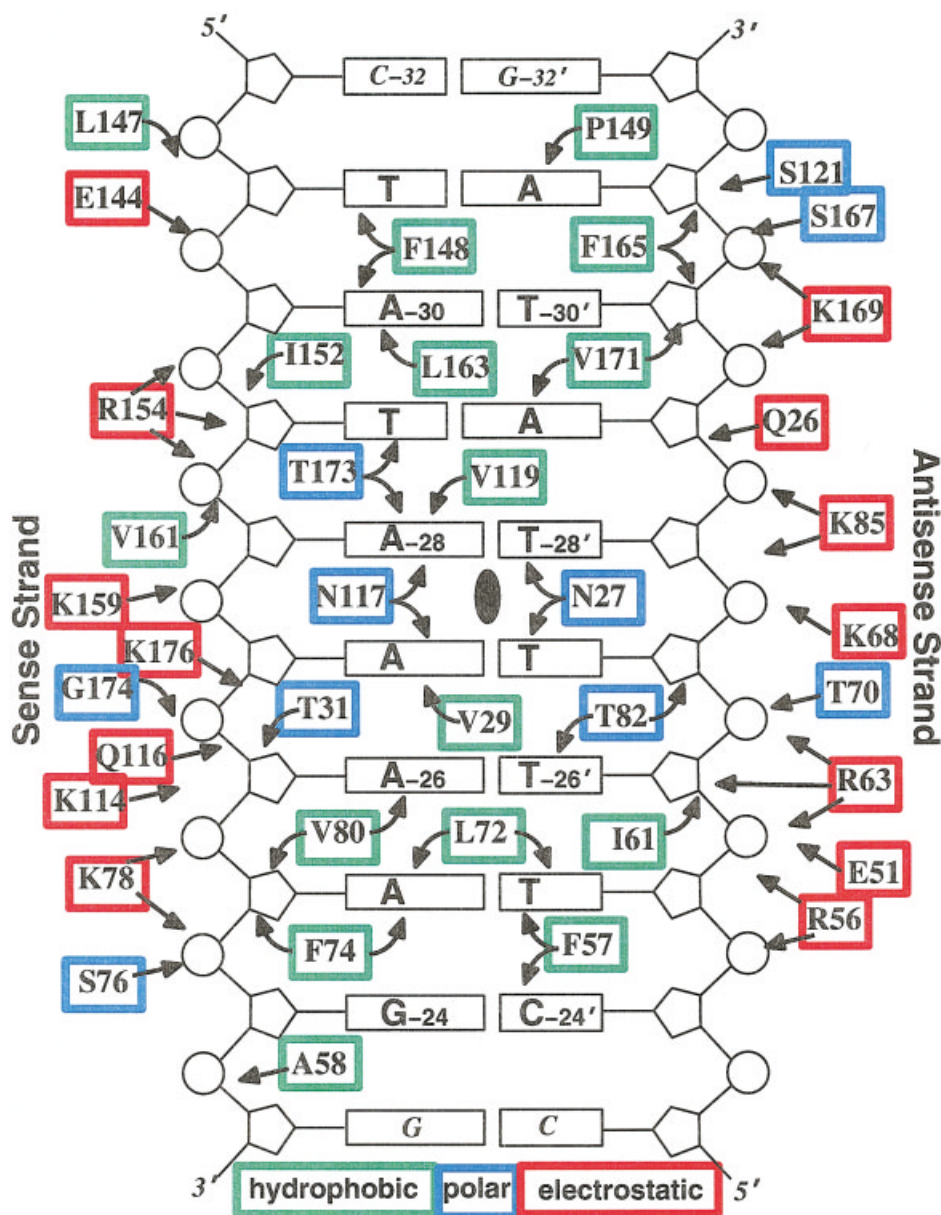
FIGURE 6   Schematic of interactions between the TATA element and TBP. The interactions (arrows) are classified as either hydrophobic (green), polar (blue), or electrostatic (red) based on the AdMLP/TBP complex 1QNE.[14] Sense strand and antisense strand interactions are placed on the left and right of the TATA element schematic; base interactions are placed within the schematic. The pseudo-dyad center of symmetry is indicated by the black ellipse.

deviations from the mean as a measure of each interaction's flexibility.

The trajectory analyses reveal that most interactions to the bases are well-maintained, i.e., hydrophobic interactions < 4.5 Å and polar interactions < 3.5 Å (Figure 7A), in good agreement with the crystallographic structures.[14]

However, in several low-activity variants (A29, T26, C25, C29, and T24), the flexibility of the

Phe 148 and Pro 149 residues increases (Figure 7A), suggesting an overall instability of the 5′ end of the TATA element interactions and withdrawal of Phe 148 from DNA in these variants. No such instability is observed at the 3′ end phenylalanines (Phe 57 and Phe 74). We associate this withdrawal motion with PC 11 (discussed below), a proposed dissociation pathway prominent in low-activity variants.

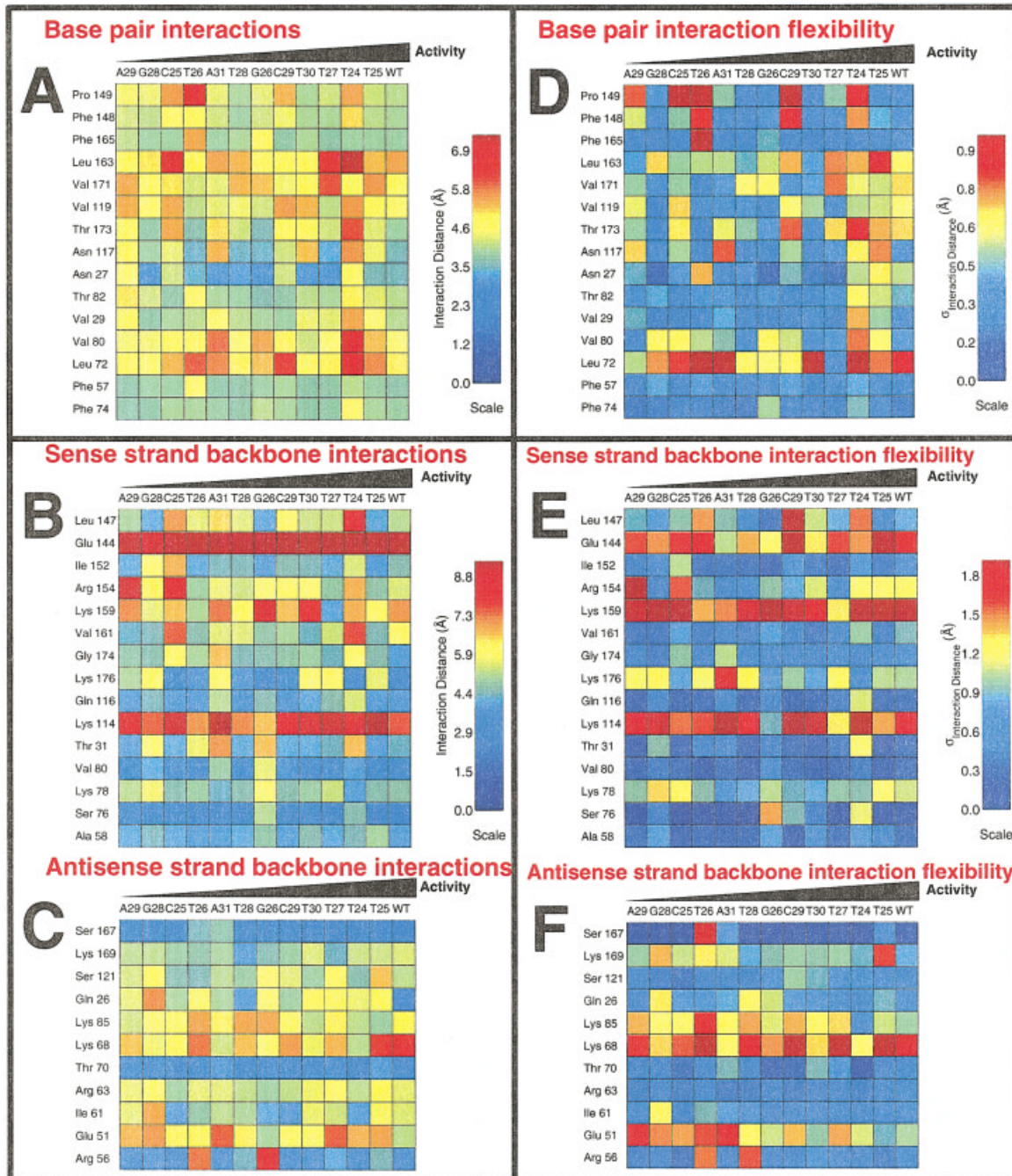Interactions in the central 6 bp of the TATA oc-

FIGURE 7 Interaction matrices of TATA/TBP variants. (**A**) Mean interaction distances between bases and TBP. (**B**) Mean interaction distances between the sense strand phosphate-deoxyribose backbone and TBP. (**C**) Mean interaction distances between the antisense strand phosphate-deoxyribose backbone and TBP. Each interaction matrix shows either the mean interaction distance (A–C). (**D**) Flexibility of interactions between bases and TBP. (**E**) Flexibility of interactions between the sense strand phosphate-deoxyribose backbone and TBP. (**F**) Flexibility of interactions between the antisense strand phosphate-deoxyribose backbone and TBP. Each interaction matrix shows either the mean interaction distance (A–C) or flexibility (D–F), i.e., the standard deviation around the mean, measured over 3900 snapshots sampled over 4.68 ns. The average distance was computed by measuring distances between TBP's side chains and the relevant interaction sites in the TATA DNA. Note that interactions involving several sidechain atoms or several DNA groups (such as intercalated phenylalanines or bifurcated asparagines) were averaged over all relevant distances. The average distance and standard deviations are color coded according to the scale shown at right.
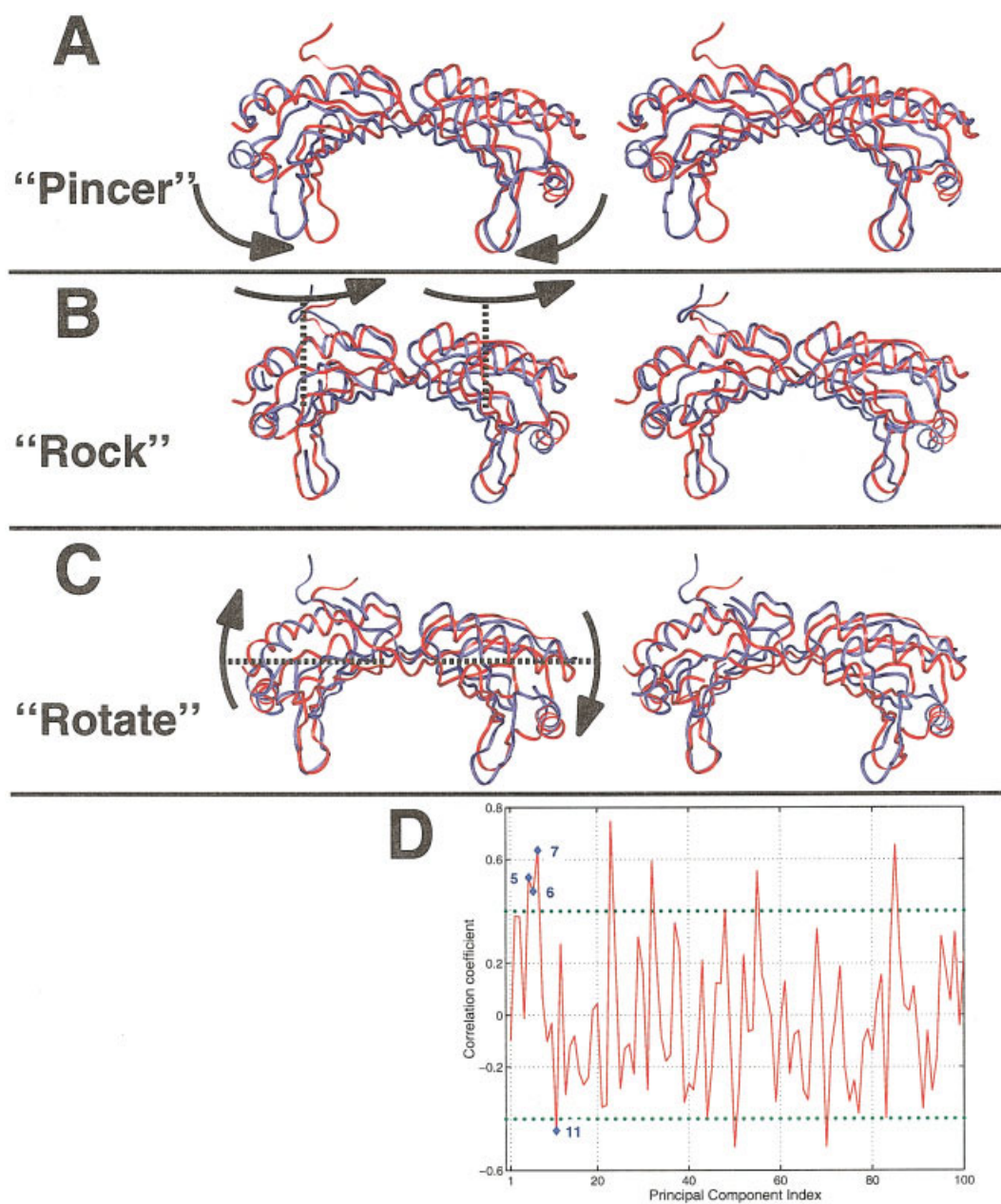
FIGURE 8 Definitions of domain motions of TBP and correlation coefficients for the top 100 PCs. Each domain motion is a rotation about one of three orthogonal axes, designated "pincer," "rock," or "rotate." Note that each domain (N- and C-terminal) rotates independently around its own set of axes, since the domains have few interdomain contacts. Motions are illustrated by deforming the average structure of the merged ensemble of all 13 variants along a given PC. The blue and red structures correspond to the minimal and maximal deformations of the associated PC, respectively (see Methods). (**A**) Pincer domain motion. (**B**) Rock domain motion. (**C**) Rotate domain motion. (**D**) PCs and correlation coefficients are plotted as a red line; PCs discussed in the text are labeled with blue diamonds. The 95% confidence limit on the correlation coefficient is approximately 0.09 for the top 11 PCs. Note that the confidence limit increases in higher order modes since they represent minor motions; however, PC 50 and lower order PCs have 95% confidence limits less then 0.2.

tamer (between Leu 163 and Leu 72) are tighter in many medium and low-activity variants than the high-activity variants WT and T25. The decreased motion and the smaller mean interaction distance of these central bp interactions reflect the increased overall curvature observed in low-activity variants (Figure 2). Ratcheted bp stacks, such as observed in the A29, T24, and C29 simulations, lengthen interactions at the TATA octamer center to 5 Å or greater in low and medium activity variants; this lengthening and the formation of voids accessible to water and ions next to the unpartnered bases at each end of the slipped region suggest that the slipped bp stacks may be a part of dissociation pathways.

Several interactions contacting the deoxyribose backbone are altered in an activity-dependent manner. The interaction between Lys 68 and bps $-27'/-28'$ becomes tighter in low- and medium-activity variants, losing the water molecule bridge indicated by the crystal structures.[14] TFIIA contacts the antisense strand[10] near the center of the TATA element where Lys 68 forms this interaction; thus, a competition between Lys 68 and TFIIA may destabilize TFIIA in low-activity complexes.

Global Motion. Results of the motion analysis by PCA are illustrated in Figure 8 as selected independent motions. We term the orthogonal rotation axes of TBP's N-terminal and C-terminal domain motions "rock," "rotate," and "pincer," respectively. In Figures 9 and 10, we show the minimal (blue) and maximal (red) motion associated with each PC; for PCs associated with TATA/TBP activity, additional analyses illustrate the motion. Below we discuss PC 11, that we find to be associated with low-activity variants ($\kappa \leq -0.4$); PCs correlated with high-activity variants ($\kappa \geq 0.4$), namely, PCs 5, 6, and 7, are discussed later. We are now posting animation sequences of selected PCs on our group website: http://monod.biomath.nyu.edu/.

PC 11 is the sole mode among the top 50 PCs that collectively describes more than 2% of the overall motion and is associated with low-activity variants ($\kappa = -0.45 \pm 0.1$). The dominant motion in PC 11 for TBP is a "pincer" motion, which withdraws the intercalated phenylalanines. This pincer motion is particularly large in the C-terminal domain, withdrawing Phe 148 and Phe 165 by ≈ 2.5 Å from the first TATA bp step; in the N-terminal domain, Phe 57 and Phe 74 are withdrawn only 1 Å from the last bp step. The withdrawal of the phenylalanines decreases the roll at both ends of the TATA element, and increases the twist outside the TATA element (Figure 9D). Essen-

tially, PC 11 suggests a possible dissociation pathway for low-activity variants.

High and Medium-Activity Complexes

DNA Structure. Our simulated TATA/TBP complexes remain highly bent as discussed above, close to the active crystallographic structure.[14] The changes in DNA structure we observe (such as bending flexibility, Hoogsteen bps, or "ratcheted" bp stacks) are either associated with low-activity variants (as discussed in the prior section) or not significant (and hence not discussed).

Protein Structure.

Domain Rotation. During the simulations, the N- and C-terminal domains twist relative to each other (Figure 5B). To characterize the relative rotation between the N- and C-terminal domains, recall that we use the five regions defined above to measure the rotation as a dihedral angle between the two planes defined by the domains (illustrated in Figure 5A). This mean torsional change from the cocrystal structure between the TBP domains varies from $-4°$ to $12°$. Most of the high- and medium-activity variants exhibit relatively low domain rotations (between $1°$ and $4°$), whereas most of the low-activity variants (yellow) have larger values. These results generally suggest that high-activity variant complexes preserve a favorable spatial relationship between the N- and C-terminal domains, while low-activity variants generate a torque between the TBP domains ($\kappa = -0.25 \pm 0.08$). Transcription factors that bind simultaneously to both domains (either directly like TFIIB[12] and the histone-like factor NC2[13] or indirectly through other factors) are likely to be affected by this domain rotation.

H2/H2′ Rotation. The H2 and H2′ α-helices occupy the upper surface of TBP opposite the DNA binding domain and expose a large surface area available for binding of other transcription factors, such as TFIIA,[10,11] NC2,[13] and TAFs.[31,32] The H2 and H2′ helices display small rotations ranging from $-2.5°$ to $5°$ (Figure 5C). Both the mean rotation of the H2 helix and the H2′ helix increase slightly with activity ($\kappa_{H2} = 0.31 \pm 0.02$, $\kappa_{H2'} = 0.49 \pm 0.02$). TBP thus adjusts to binding of high-activity variants by rotating upper surface α-helices, and modulating activity of factors interacting with this site.

H1/H1′ Rotation/Translation. No significant motions of the H1 and H1′ helices are found for medium- and high-activity variants. The motions associated with low-activity variants were discussed in the prior section.
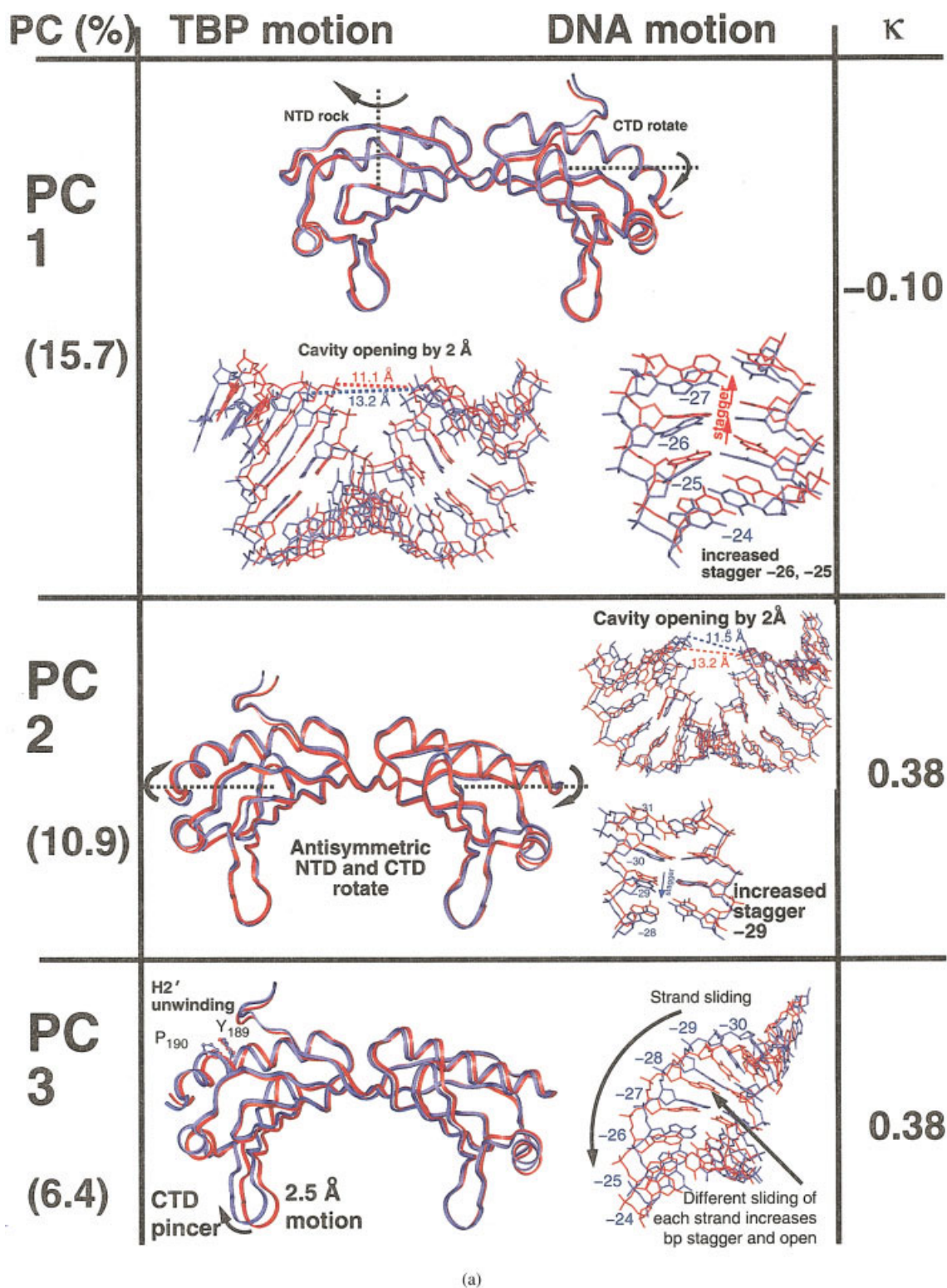
FIGURE 9   Motions in 7 of the top 11 uniform ensemble PCs. (**A**) PCs 1, 2, and 3. (**B**) PCs 4, 8, and 9. (**C**) PC 10. Structures were generated by deforming the average structure of the merged ensemble of all 13 variants along the corresponding PC. The blue and red structures correspond to the minimal and maximal deformations of the associated PC, respectively (see Methods).
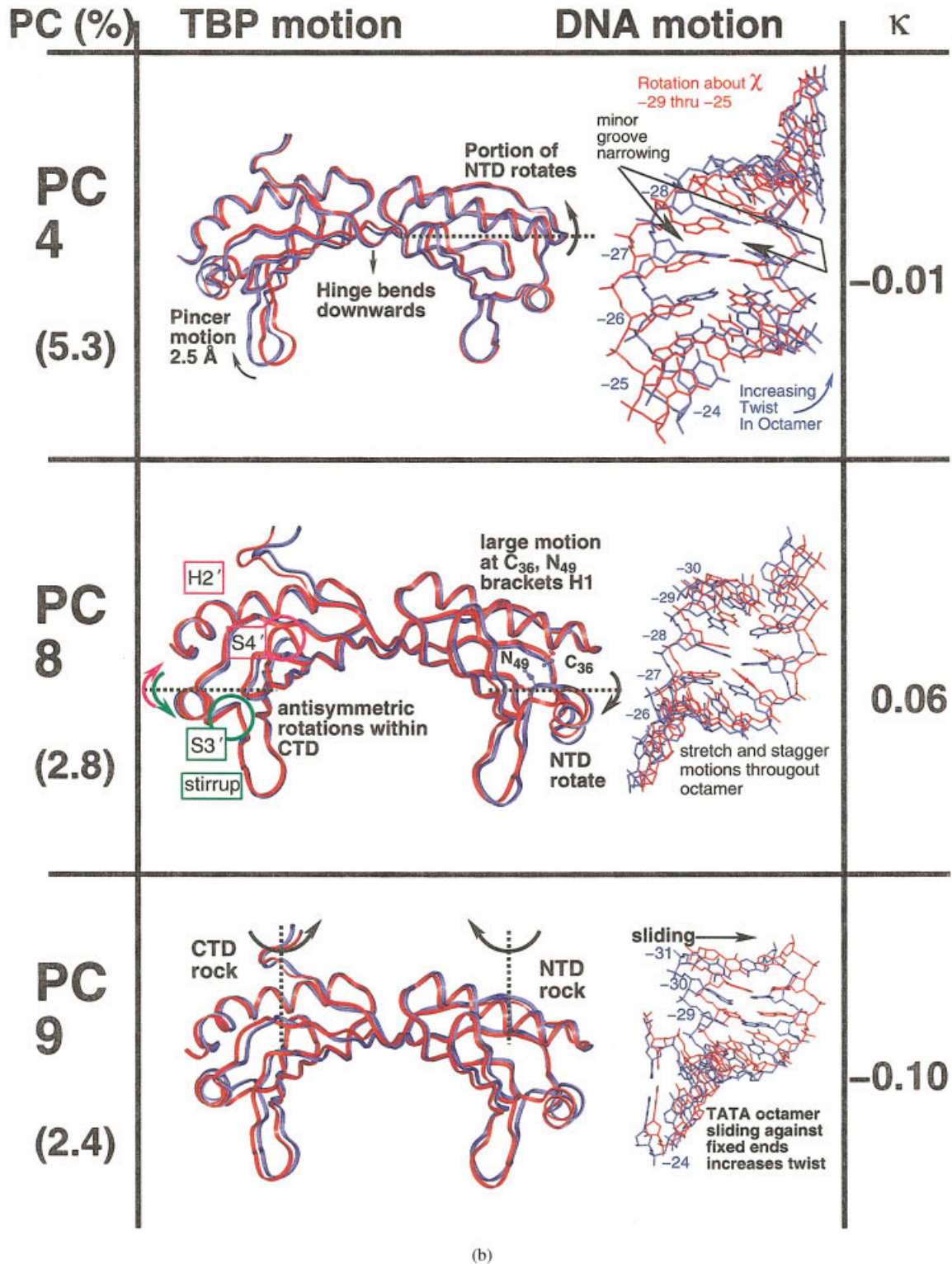
FIGURE 9 (*Continued from the prevous page.*)

DNA/TBP Interface. In high- and medium-activity variants, Arg 154 contacts the sense strand backbone near bp $-29/-28$; in low-activity variants, this inter-

action is broken (Figure 7B). Arg 154 contacts several different functional groups; this variability may entropically stabilize Arg 154 against excess solvent, as
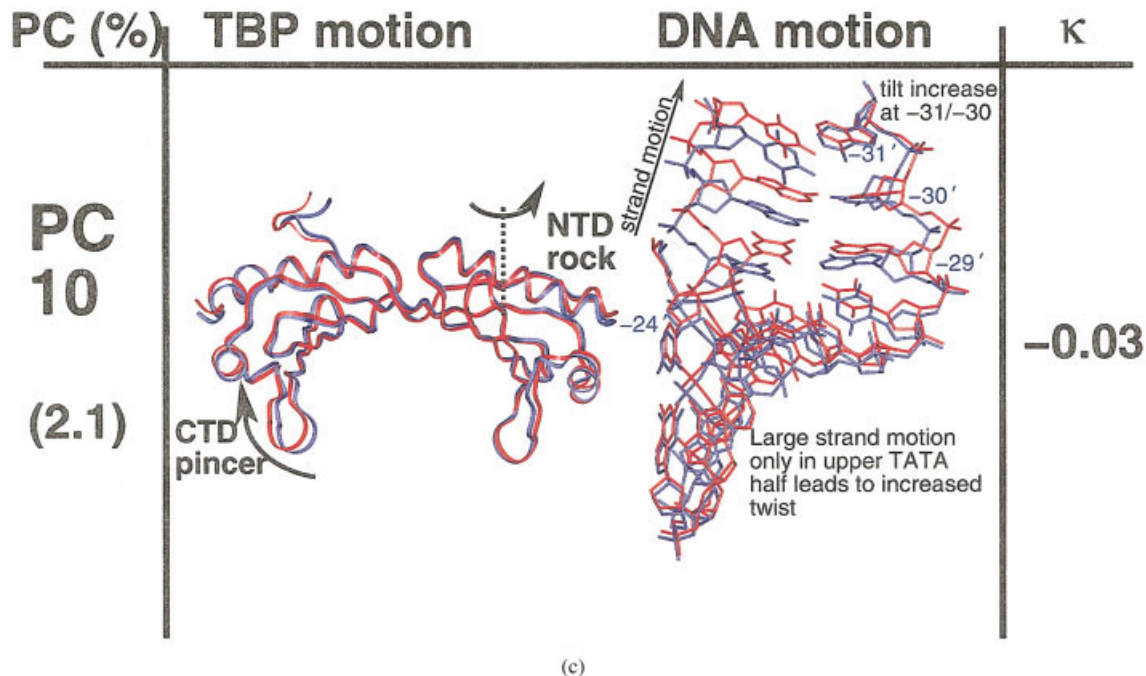
FIGURE 9   (*Continued from the prevous page.*)

proposed by Pastor *et al.*[53] Other interactions, such as between Glu 51 and phosphate $-26'$, Lys 85 and phosphates $-28'/-29'$, and Lys 169 and phosphates $-31'/-30'$, may specifically stabilize high-activity variants through exploiting solvent entropy: while these interactions form at a distance with the antisense strand backbone in all variants, these interactions in high and medium-activity variants have lower flexibility (Figure 7F) and appear to employ solvent molecules as bridging intermediates.

Global Motion. Many of the PC modes associated with high- and medium-activity variants share a common tendency: the backbone atoms of the sense strand slide back and forth along the strand direction. Although this motion occurs in both strands in PCs 5, 6, and 7, it is much stronger in the sense strand.

We show the motions of PC 5 ($\kappa = 0.53 \pm 0.09$) in Figure 10A. Part of PC 5 describes a rotation of the H2 and H2′ helices; this rotation likely affects transcription factors recognizing TBP's upper surface like TAFs and TFIIA (described above). The motions of the sense strand backbone are particularly large at the 5′ and 3′ ends of the TATA element, as evident by the isotropic fluctuations of the TATA backbone atoms; in contrast, much less motion is observed in the antisense strand (except adjacent to the end base; data not shown). We attribute this backbone motion to the different numbers of cationic residues on each side and each "leg" of TBP[7,54]; the fewer Lys/Arg residues

proximal to the sense strand TATA DNA implies an increased mobility in this half of the complex. TFIIA contacts the antisense strand backbone within the TATA element,[10] and the reduced motion of the antisense strand in PCs 5, 6, and 7 might stabilize TFIIA binding.

The PC 6 motion ($\kappa = 0.47 \pm 0.09$) separates the two domains of TBP: the C-terminal domain moves back from the DNA in a pincer motion, while the N-terminal domain "rotates" counterclockwise (Figure 10B). The TATA motions imply strand separation in bps -28 through -25, through accompanying increases in the DNA base motions stretch, shear, and stagger in this region[55] (Figure 10B). Strand separation is easier in a deformed and constrained DNA segment (like TATA/TBP); such melting motions create a void for Hoogsteen bp formation[14] and may be linked to nearby strand melting during initiation.[56]

The motion of the N- and C-terminal domains in PC 7 ($\kappa = 0.63 \pm 0.09$) is closely related to that of PC 6: "rotation" of the C-terminal domain and pincer of the N-terminal domain. This PC separates the two stirrups but does not open the central cavity. The TATA DNA accomodates the protein motion by increased twist in some regions and the sliding of the sense strand observed in other PCs (Figure 10C). The PC 7 DNA motion appears to maintain interactions between the two TBP domains and the DNA. This motion indicates that high- and medium-activity variants relieve torsional stress in the underwound TATA

DNA without significantly altering their favorable complex geometry, thus stabilizing the complexes.

Water Molecules at the Interface—All Variants

The TATA/TBP complex interface[4,5] is largely hydrophobic and dehydrated. Still, a buried water molecule was observed in the T26 cocrystal,[14] suggesting that water may either be trapped in, or penetrate into, the interface. A recent combination of experimental studies with simulations suggested that backbone interactions in the TATA/TBP complex might have low entropic penalties when solvated due to the intrinsic flexibility of certain backbone interactions.[53] The concentration of osmolytes has also been shown to be a factor affecting the *in vitro* behavior of TATA/TBP complexes, apparently increasing sequence-dependent bends in solution to a common 80° bend angle close to the crystallographic value.[18] Thus, the role of water in TATA/TBP complexes remains unclear.

We find a number of water molecules located at the interface between TBP and the TATA DNA. In our initial models, solvent molecules were placed in the available interfacial cavities and the initial placement was verified by a simulated annealing of the chemical potential in several grand canonical Monte Carlo simulations.[57] These additional Monte Carlo simulations verified occupancy of these sites *without steric clashes.*

In Figure 11, we analyze the overall lifetimes of 10 water sites (labeled I–X, Figure 11A) and the mean lifetime of water in each site; we chose these 10 sites because water molecules are favored there in at least 11 of our 13 initial structures. The lifetimes of waters in these ten sites were determined by monitoring distances (between 4.5 and 6.0 Å; 5.2 Å average) simultaneously to both nearby protein and nucleic acid atoms buried in the TATA/TBP interface; protein atoms, in particular, are buried $\approx$ 8–9 Å deep from waters in the TATA major groove.

Although some of the waters initially placed in the interfacial sites remained in, or near, the site during the simulation (as indicated by the large peak in the histogram at 5 ns of Figure 11B), many additional water molecules also migrated into the interface. The largest histogram peak indicates small binding site lifetimes and demonstrates that many water molecules exchanged into and out of the TATA/TBP interface. Although most of the interfacial sites appear to be serially or multiply occupied by different water molecules, the weak confidence limits indicate that there is no strong correlation between TE and the mean

water lifetime at any particular site (Figure 11C). Interestingly, the simulated T26 complex exchanges water rapidly near the bp 26′ site (waters VII and VIII) identified in the cocrystal,[14] suggesting that the water molecule is observed in the T26 cocrystal structure because the locations behind bps 26 and 26′ are accessible. The presence of exchangeable waters within the TATA/TBP complex is an unexpected result, and may explain observations such as the effect of osmolytes on TATA/TBP binding states,[18] and the conserved crystallographic structure of TATA/TBP complexes with varying activity.[14]

Salt Interactions—All Variants

Ions in the major groove are essential for screening the potential generated by the phosphate groups crowded together in the bent DNA and the cationic residues binding the TATA DNA backbone.[24,54] This environment explains the attraction of ions to the cavity formed between the 5′ and 3′ ends of the bent TATA DNA. In Figure 12A, we illustrate this for ten sodium ions (of a total of 37 sodium ions) in the major groove at the end of the C29 trajectory. One ion (identified in purple) is within 4 Å of the DNA, near one end of the slipped bp stack; ions in such proximity to the DNA can exert strong local forces, compelling nearby polar groups to move in concert with the ion. In such a way, these ions may promote or stabilize particular conformations typical of low-activity variants, such as Hoogsteen bps or ratcheted bp stacks.

Low-activity variants reveal a tendency ($\kappa = -0.61 \pm 0.06$) for more sodium binding events (i.e., sodiums to come within 4 Å of the base atoms lining the major groove) (Figure 12B). Overall, sodium binding events are quite common, observed in more than 75% of all trajectory snaphots, while chloride binding events are rare. The increased cation frequency in the cavity of low-activity complexes and the associated close approaches may contribute to low-activity variant instability, while altering the overall electrostatic field presented to other transcription factors by this highly charged complex.

CONCLUSION

Due to the large curvature and deformation of the TATA DNA in the complex with TBP, the motion and energy (or implied force) inherent to the bound complex play a prominent role in TATA/TBP activity. In examining systematic sequence variations of
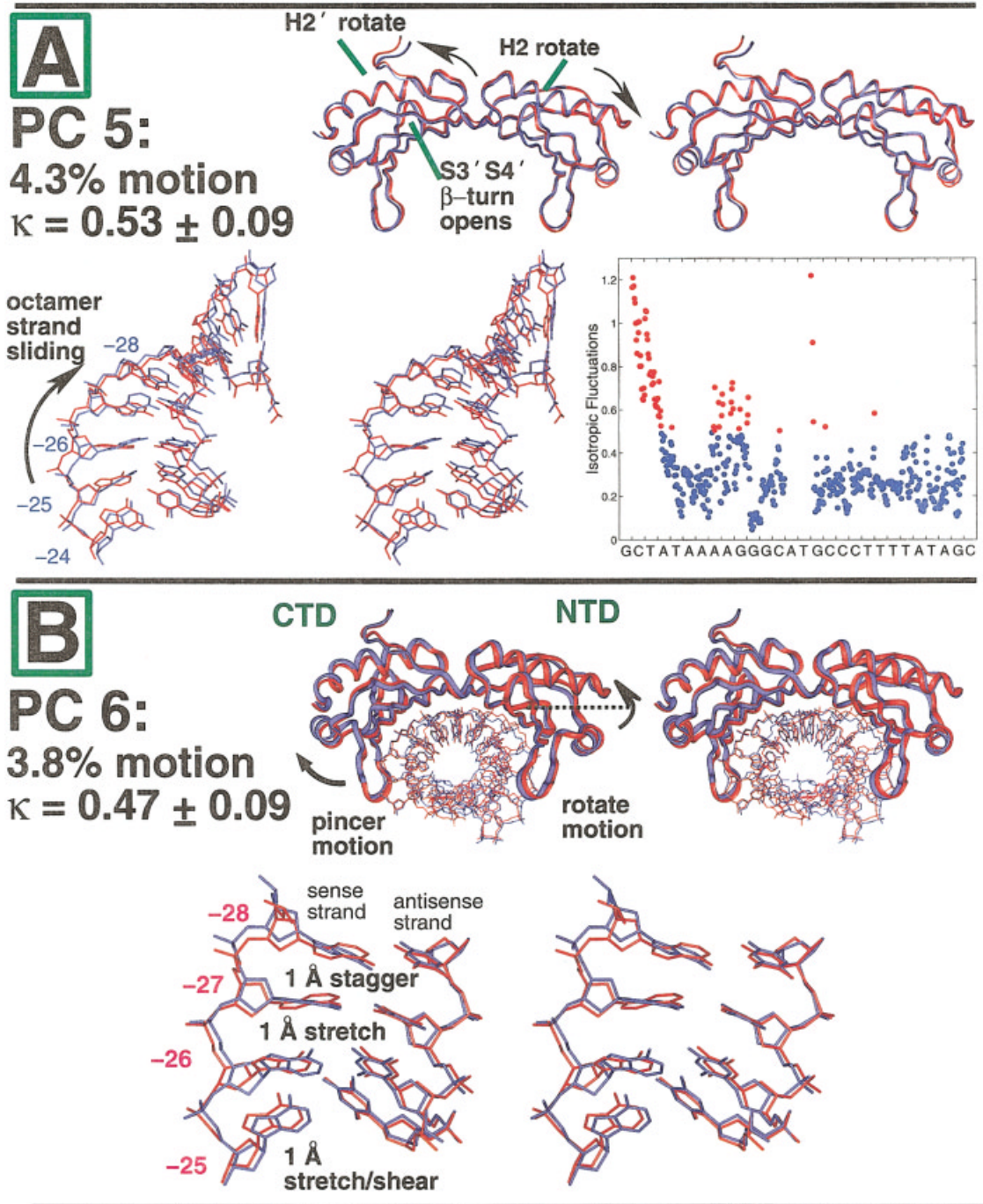
FIGURE 10 Motions in four significant PCs correlated with activity ($|\kappa| > 0.4$). (**A–C**) High-activity variant motions: PCs 5, 6, and 7. (**D**) Low-activity variant motion: PC 11. Structures were generated as described in Figure 9.

TATA/TBP complexes, various research groups have concluded that subtle motion and structural differences were closely linked,[15,17,23,24] despite global similarity in overall structures.[14] Our detailed simulations of variant TATA/TBP complexes, extending earlier investigations from several groups[15,17,18,21,22,52–54] and in good agreement with several experimental studies,[15,17,18] have revealed a variety of intrinsic motions and conformational changes in the protein and DNA components.

The conformational changes observed in our simulations might explain some aspects of observed ac-
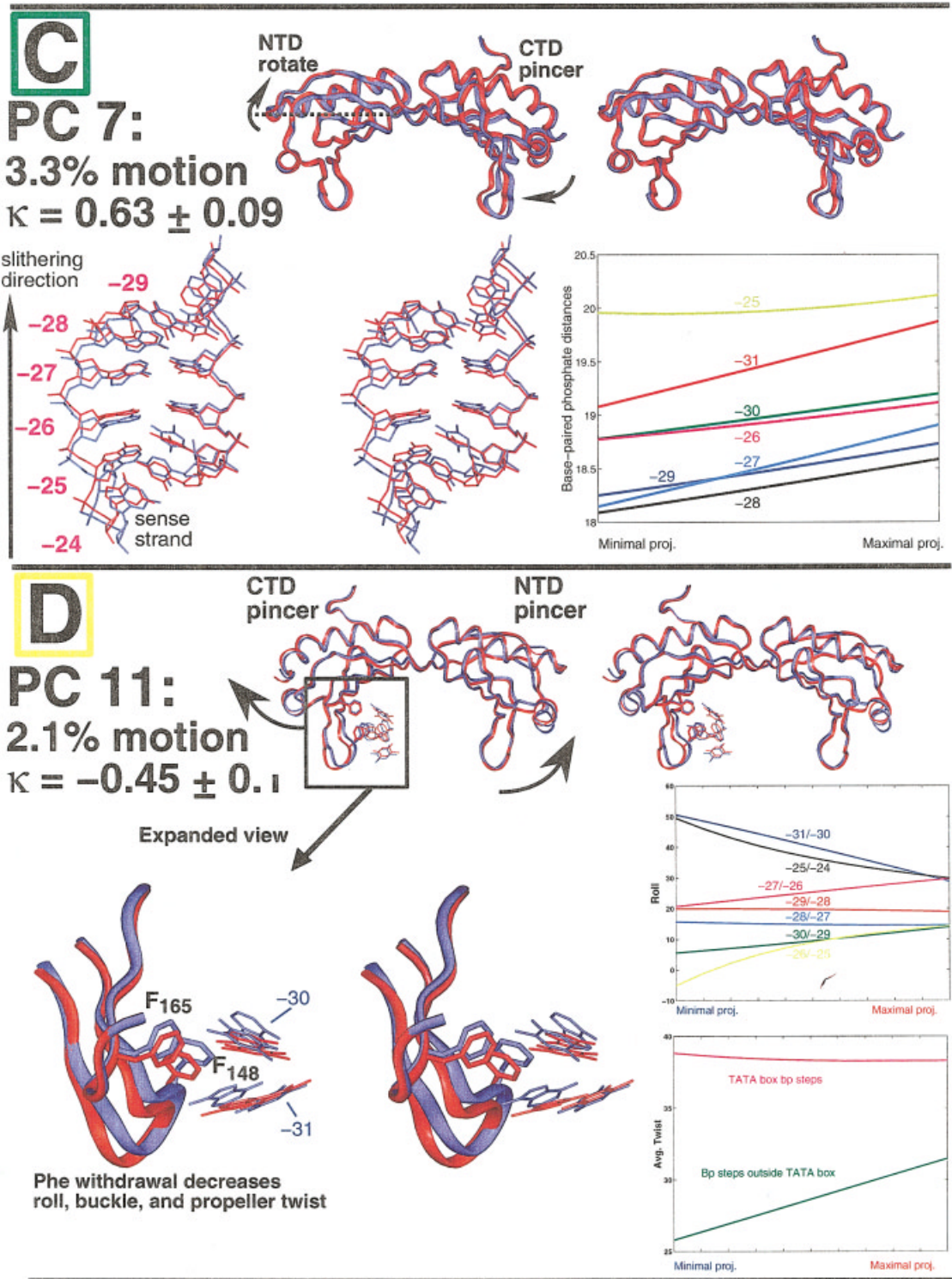
FIGURE 10 (*Continued from the prevous page.*)

tivity differences for these variant complexes.[14] Several predictions also emerge:

1. To develop the highly bent DNA observed in crystallographic[4–9,14] and other experimental[15–18] studies, the DNA passes through an intermediate with a higher bend than the crystallographic structures. This intermediate is expected to be unstable and may be spectroscopically detectable.
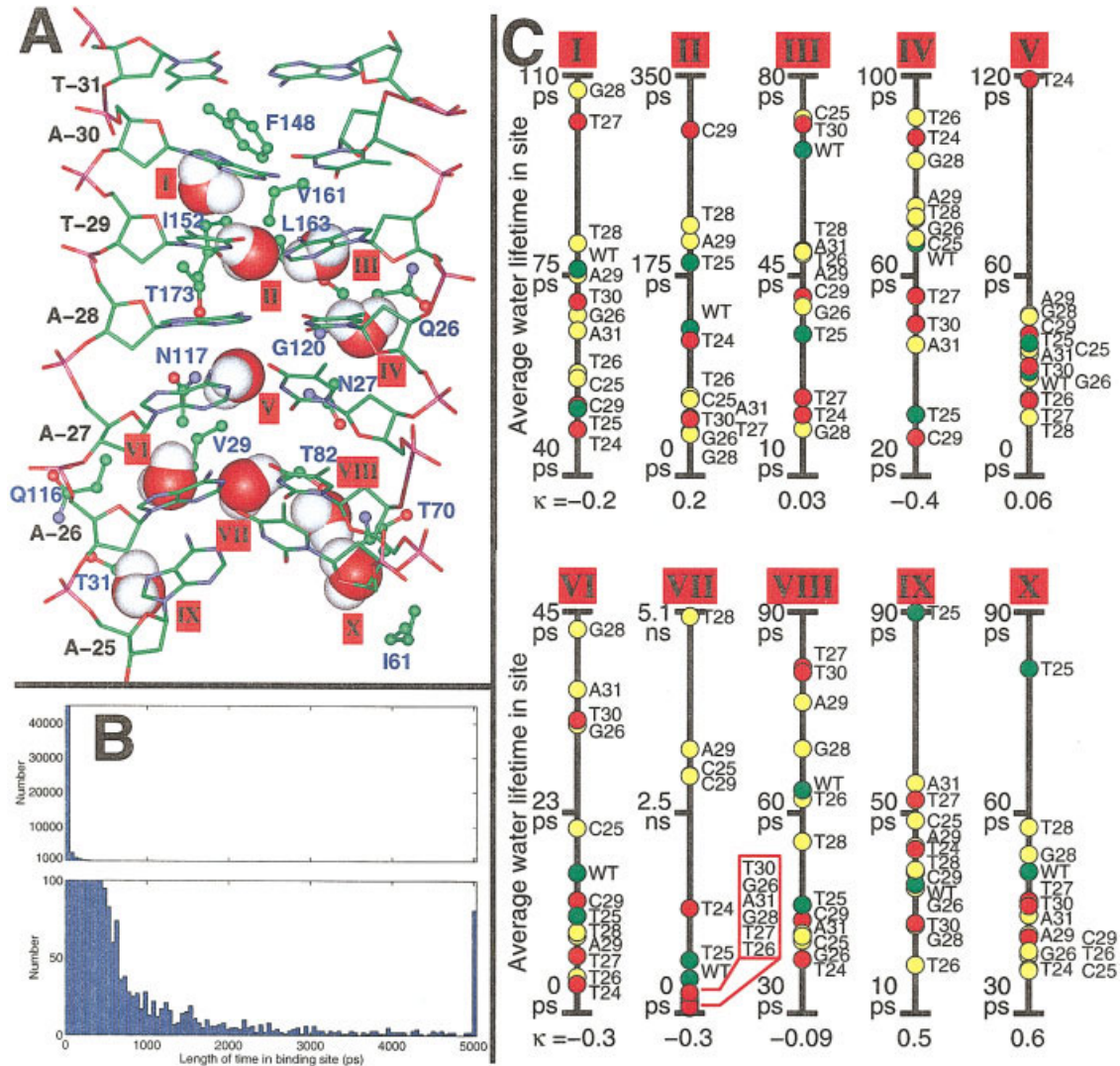
FIGURE 11 Interfacial TBP/TATA complex water sites chosen for analysis. (**A**) Ten water sites indicated by the Roman numerals next to the water molecules. Only side-chain and some backbone atoms belonging to residues used to monitor site occupancy are indicated with a ball-and-stick model; the rest of TBP is not shown. Note that the structure is mildly distorted to flatten and display the entire binding site; due to this distortion, spatial relationships between different residues and/or water molecules are illustrative only. (**B**) Histogram of lifetimes of all water molecules observed in the ten interfacial water sites. (**C**) The mean lifetimes of water molecules in a selected interfacial site of each variant TATA/TBP complex are plotted against TE. Water occupancies were monitored for 4200 snapshots spanning the entire 5.04 ns trajectory. The interfacial site corresponding to each plot is indicated by the corresponding Roman numeral. In all ten sites, the confidence limit on κ is 0.3 or larger.

2. Ratcheted bp stacks leave bases unpaired at either end of the ratcheted region. These unpaired bases expose protons normally protected by Watson–Crick hydrogen bonding. In addition, ratcheted regions create a void in the bp stack region accessible to water and available to form Hoogsteen bp.[14] Thus, TBP binding to TATA elements may expose base protons and make the protons available for exchange; these protons may exchange more readily in low-activity variants.

3. High-activity complexes tend to rotate the carboxyl end of the H2 and H2′ helices on TBP's upper surface, while displaying low rotations
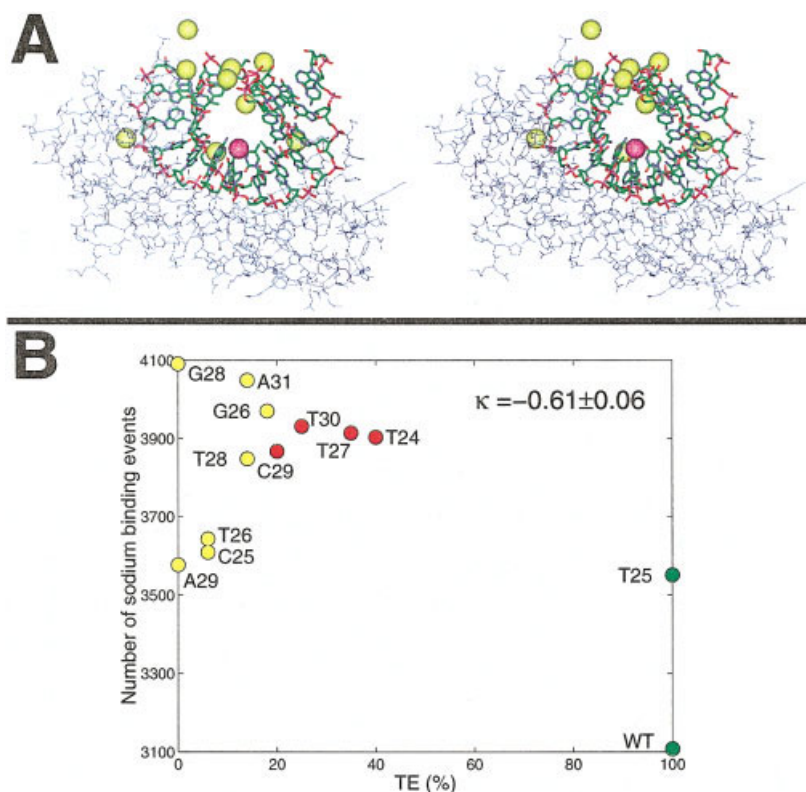
FIGURE 12 (**A**) Stereo view of sodium cations in the cavity formed between the ends of the bent TATA DNA. Only a few of the ions (yellow and purple spheres) near the C29 TATA DNA are illustrated. The structure shown is the final snapshot in a 5.04 ns trajectory. Note that a sodium ion (purple sphere) is sitting close to the "ratcheted" bp stack region. (**B**) The number of cation binding events was computed over 4200 snapshots from each variant's trajectory. Only close approaches to the DNA atoms were monitored. DNA atoms were separated into three classes: major groove atoms, minor groove atoms, and backbone atoms. Sodium binding events were measured as sodiums within 4 Å of any major groove atom (except for the end base pairs).

and translations of the solvent-exposed H1 and H1′ helices. Within our 5-ns simulations, this motion separates the C_α backbone of these helix pairs by 3 Å or more. The separation of these subdomains should be detectable with appropriate spectroscopic probes.

4. Hydroxyl radical footprints are experimentally available[17] for only 4 of our 13 variants. We have calculated footprints of the other 9 variants; these footprints predict the solvent accessibility of the DNA backbone in these complexes.

Undoubtedly, continuing experimental, structural, and computational investigations of sequence variations of the TATA promoter will shed further insights into the rich behavior and conformational ensemble relevant to eukaryotic transcription complexes. To extend our understanding of mechanistic pathways of genomic regulation, computer simulations in particular must extend their scope to temporal and spatial scales of larger macromolecular systems; ongoing work in several laboratories is addressing multiscale modeling and its application to transcriptional and genomic machinery.

APPENDIX A

Hydroxyl Radical Footprints—Simulation Validation and Predictions

The structures observed in our simulation closely resemble the cocrystal complexes,[14] despite local differences in both the protein and DNA components. To validate our results, we have compared the hydroxyl radical footprints of our simulated complexes to experimental footprints from these same variants under
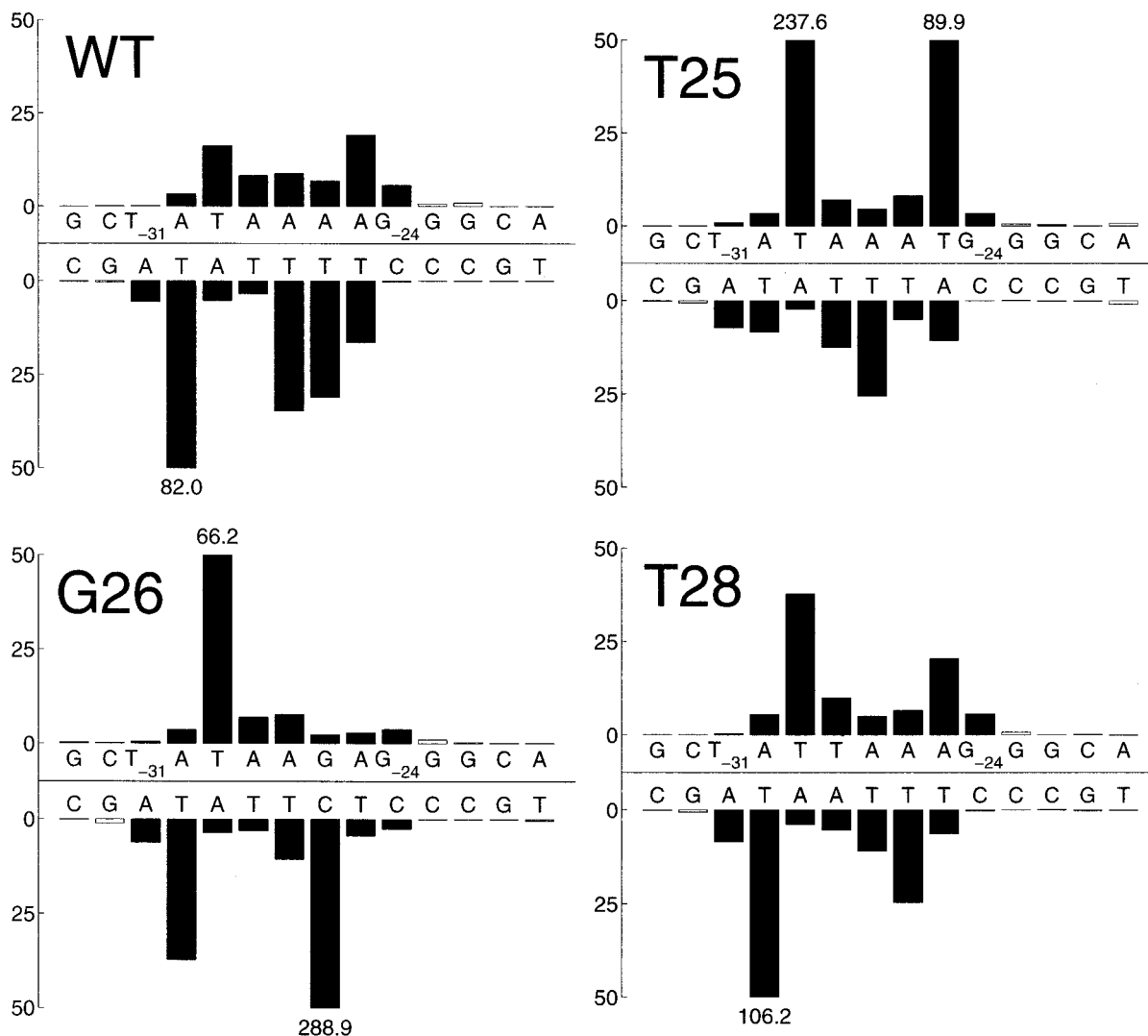
FIGURE 13 Hydroxyl radical footprints computed from the ratio of the accessible deoxyribose proton surface area of the simulated free TATA DNAs[24] to the bound complexes: $(\langle C_{i,\text{free}}\rangle / \langle C_{i,\text{bound}}\rangle) - 1$. Bars larger than a value of 50 exceed the graph boundaries and are truncated and labeled by their numerical value.

conditions of saturating TBP.[17] Recently, Pastor et al. have demonstrated[53] that MD simulations can successfully reproduce structural details (including backbone dynamics) that regulate the accessibility of TATA/TBP complex deoxyribose protons to hydroxyl radical. Of relevance to our results, experimental studies of variant TATA/TBP complexes examined in our simulations have shown that the hydroxyl radical footprints are DNA sequence specific, suggesting altered solvent accessibility and different TATA/TBP complex structures.[17]

Calculation of Hydroxyl Radical Footprint. The reagent hydroxyl radical (OH•) attacks and cleaves the DNA phosphate-deoxyribose backbone in a largely sequence-independent manner in mixed sequence DNA. The small size of the hydroxyl radical (comparable to a water molecule) allows the resulting radical-catalyzed DNA cleavage to be interpreted as a measure of the solvent accessible surface area of the DNA.[58] This measure has found utility in measuring the narrowing of the adenine tract minor groove[59] (see also http://monod.biomath.nyu.edu/index/papdir/pap_2_76.html) and the basepair-dependent details of protein-DNA complexes.[17,53,60]

Deuterium isotope effect studies of the DNA cleavage reaction conducted by Tullius and co-workers have shown all deoxyribose protons are attacked by hydroxyl radical, although the C4′ and C5′ protons are the primary attack sites.[58] The cumulative proba-
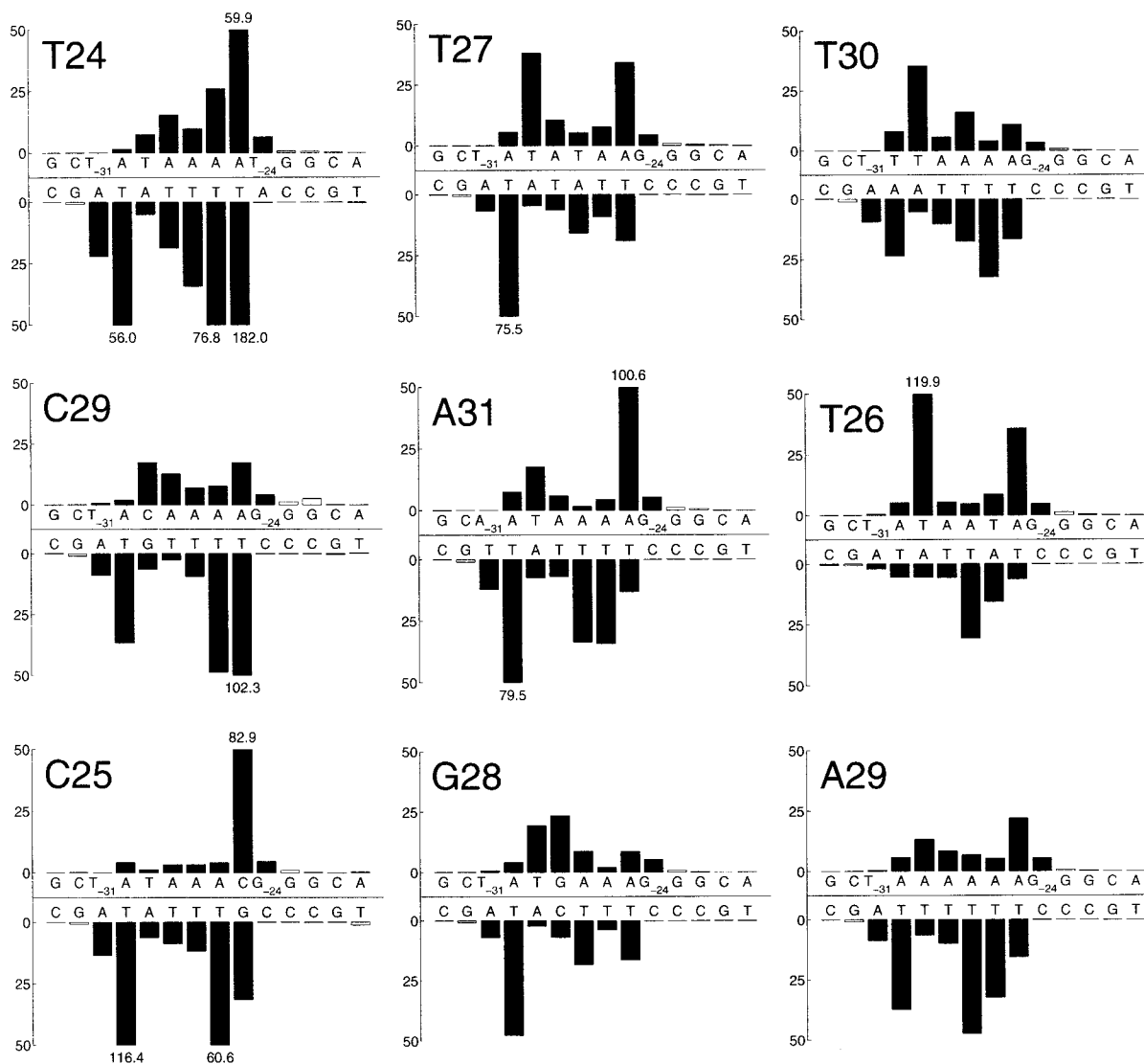
FIGURE 13 (*Continued from the previous page.*)

bility of cleaving a deoxyribose with hydroxyl radical is the sum of the individual proton probabilities P_j and is strongly correlated with each deoxyribose proton's surface area S_{ij}. Thus, we can compute from our structures the accessible surface area and use this quantity to weigh each proton's individual cleavage probability. The average cleavage probability C_i for each deoxyribose i and the corresponding standard deviation σ_{C_i} are calculated as

$$C_i = \langle \sum_{\text{protons } j} P_j S_{ij} \rangle,$$

$$\sigma_{C_i} = \sqrt{\langle \sum_{\text{protons } j} \sigma_{P_j}^2 S_{ij}^2 \rangle + \sum_{\text{protons } j} P_j^2 \sigma_{S_{ij}}^2} \quad (3)$$

where P_j, the cleavage probability for each deoxyribose proton j, is taken from deuterium isotope effect

experiments[58] and S_{ij}, the surface area of each proton j in deoxyribose i, is computed by the available procedure in CHARMM. We use the Lee and Richards algorithm[61] as implemented in CHARMM,[34] which basically rolls a water-molecule size probe around the solute surface in a plane to determine successive sections of the accessible surface area. To account for sequence-dependent differences in bp-dependent solvent accessibility in the free and TBP-bound states,[17] we plot the ratio $(\langle C_{i,\text{free}} \rangle / \langle C_{i,\text{bound}} \rangle) - 1$; this ratio is zero if there are no differences between the free DNA and the bound complex.

Comparisons to Experimental Footprints. In Figure 13, we illustrate the relative cleavage ratio of the free TATA DNA to the bound TATA DNA. The computed footprints of several variants (WT, T28, T25,

G26), which have experimentally determined footprints available for comparison,[17] reproduce the corresponding experimental pattern associated with each variant complex. We note that an earlier study from Pastor *et al.* also successfully reproduced the WT footprint pattern.[53]

There are only a few bps where differences exist in the peak height between our calculated footprints and the experimental comparisons. For example, the experimental footprint of the WT complex indicates a peak at base 28′ that is not visible in our results. Of course, perfect agreement between simulation and experiment is not expected in this particular system, since the weights employed in the footprint calculation are derived from *B*-form DNA[58] and may be different in the "northern" deoxyribose sugar pucker phases of the distorted TATA DNA. Though the experimental footprints were determined using a saturating concentration of TBP, intermediate binding states were also footprinted due to the slow isomerization times of TATA/TBP complexes. This leads to a merging of state-dependent protection patterns[17]; since the TATA DNA motion from the free form to the complexed form is very large and it is unlikely that the bending/binding processes lead to uniform monotonic changes in the protection pattern, the final experimental footprint contains an indeterminate nonuniform component from the partially associated/dissociated TATA/TBP complexes. Nevertheless, the extensive similarities between the calculated and experimental footprints imply that the overall solvent accessibility of the simulated complexes is very similar to the solution conformations.

In addition, we have computed nine additional footprints that currently have no experimental comparisons; these nine footprints (for variants T24, T27, T30, C29, A31, T26, C25, G28, and A29) predict the relative magnitudes of the various bp peaks and the sequence-dependent motions of side chains protecting each of the variant complexes deoxyribose protons.[53]

REFERENCES

1. Roeder, R. G. Trends Biochem Sci 1996, 21, 327–335.
2. Sauer, F.; Tjian, R. Curr Opin Genet Dev 1997, 7, 176–181.
3. Bucher, P. J Mol Biol 1990, 212, 563–578; http://www.epd.isb-sib.ch/promoter_elements/.
4. Kim, J. L.; Nikolov, D. B.; Burley, S. K. Nature 1993, 365, 520–527.
5. Kim, Y.; Geiger, J. H.; Hahn, S.; Sigler, P. B. Nature 1993, 365, 512–520.
6. Chasman, D. I.; Flaherty, K. M.; Sharp, P. A.; Kornberg, R. D. Proc Natl Acad Sci USA 1993, 90, 8174–8178.
7. Nikolov, D. B.; Burley, S. K. Nature Struct Biol 1994, 1, 621–637.
8. Nikolov, D. B.; Chen, H.; Halay, E. D.; Hoffman, A.; Roeder, R. G.; Burley, S. K. Proc Natl Acad Sci USA 1996, 93, 4862–4867.
9. DeDecker, B. S.; O'Brien, R.; Fleming, P. J.; Geiger, J. H.; Jackson, S. P.; Sigler, P. B. J Mol Biol 1996, 264, 1072–1084.
10. Tan, S.; Hunziker, Y.; Sargent, D. F.; Richmond, T. J. Nature 1996, 381, 127–134.
11. Geiger, J. H.; Hahn, S.; Lee, S.; Sigler, P. B. Science 1996, 272, 830–836.
12. Nikolov, D. B.; Chen, H.; Halay, E. D.; Usheva, A. A.; Lee, K. H. D. K.; Roeder, R. G.; Burley, S. K. Nature 1995, 377, 119–128.
13. Kamada, K.; Shu, F.; Chen, H.; Malik, S.; Stelzer, G.; Roeder, R. G.; Meisterernst, M.; Burley, S. K. Cell 2001, 106, 71–81.
14. Patikoglou, G. A.; Kim, J. L.; Sun, L.; Yang, S.-H.; Kodadek, T.; Burley, S. K. Genes Dev 1999, 13, 3217–3230.
15. Wu, J.; Parkhurst, K.; Powell, R.; Brenowitz, M.; Parkhurst, L. J Biol Chem 2001, 276, 14614–14622.
16. Davis, N. A.; Majee, S. S.; Kahn, J. D. J Mol Biol 1999, 291, 249–265.
17. Mollah, A. K. M. M.; Gilden, B.; Barsi, J.; Jamison, E.; Morris, S.; Pastor, N.; Parkhurst, K. M.; Parkhurst, L. J.; Burley, S. K.; Brenowitz, M. Personal communication, 2002.
18. Wu, J.; Parkhurst, K. M.; Powell, R. M.; Parkhurst, L. J. J Biol Chem 2001, 276, 14623–14627.
19. Wobbe, C. R.; Struhl, K. Mol Cell Biol 1990, 10, 3859–3867.
20. Bernués, J.; Carrera, P.; Azorín, F. Nucleic Acid Res 1996, 24, 2950–2958.
21. Pastor, N.; Pardo, L.; Weinstein, H. Biophys J 1997, 73, 640–652.
22. de Souza, O. N.; Ornstein, R. L. Biopolymers 1998, 46, 403–415.

23. Bareket-Samish, A.; Cohen, I.; Haran, T. E. J Mol Biol 2000, 299, 965–977.

24. Qian, X.; Strahs, D.; Schlick, T. J Mol Biol 2001, 308, 681–703.

25. Schlick, T. Molecular Modeling and Simulation: An Interdisciplinary Guide; Springer-Verlag: New York, 2002.

26. Schlick, T. J Comp Phys 1999, 151, 1–8 (Special Volume on Computational Biophysics).

27. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, I. T. Acc Chem Res 2000, 33, 889–897.

28. Case, D. A.; et al. AMBER 6; University of California, San Francisco, 1999.

29. Cheatham, T. E., III; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. J Am Chem Soc 1995, 117, 4193–4194.

30. 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. J Am Chem Soc 1995, 117, 5179–5197.

31. Struhl, K. Genes Func 1997, 1, 5–9.

32. Albright, S. R.; Tjian, R. Gene 2000, 242, 1–13.

33. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acid Res 2000, 28, 235–242.

34. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comp Chem 1983, 4, 187–217.

35. Qian, X.; Strahs, D.; Schlick, T. J Comp Chem 2001, 22, 1843–1850.

36. Gilson, M. K.; Sharp, K.; Honig, B. H. J Comp Chem 1987, 9, 327–335.

37. Steinbach, P. J.; Brooks, B. R. J Comp Chem 1994, 15, 667–683.

38. Barth, E.; Schlick, T. J Chem Phys 1998, 109, 1617–1632.

39. Sandu, A.; Schlick, T. J Comp Phys 1999, 151, 74–113.

40. Yang, L.; Beard, W. A.; Wilson, S. H.; Broyde, S.; Schlick, T. J Mol Biol 2002, 317, 651–671.

41. Batcho, P.; Case, D. A.; Schlick, T. J Chem Phys 2001, 115, 4019–4029.

42. Qian, X.; Schlick, T. J Chem Phys 2002, 116, 5971–5983.

43. York, D. M.; Yang, W.; Lee, H.; Darden, T.; Pedersen, L. G. J Am Chem Soc 1995, 117, 5001–5002.

44. Babcock, M. S.; Olson, W. K. J Mol Biol 1994, 237, 98–124.

45. Babcock, M. S.; Pednault, E. P. D.; Olson, W. K. J Biomol Struct Dynam 1993, 11, 597–628.

46. Babcock, M. S.; Olson, W. K. In Computation of Biomolecular Structures: Achievements, Problems, and Perspectives; Soumpasis, D. M., Jovin, T. M., Eds.; Springer-Verlag: Heidelberg, 1993.

47. Lavery, R.; Sklenar, H. Curves 5.2: Helical Analysis of Irregular Nucleic Acids; Laboratoire de Biochimie Theorique; CNRS URA 77, Institut de Biologie Physico-Chimique, Paris, France, 1997.

48. Stofer, E.; Lavery, R. Biopolymers 1994, 34, 337–346.

49. Strahs, D.; Schlick, T. J Mol Biol 2000, 301, 643–663.

50. Sherer, E. C.; Harris, S. A.; Soliva, R.; Orozco, M.; Laughton, C. A. J Am Chem Soc 1999, 121, 5981–5991.

51. Caves, L. S. D.; Evanseck, J. D.; Karplus, M. Protein Sci 1998, 7, 649–666.

52. Miaskiewicz, K.; Ornstein, R. L. J Biomol Struct Dynam 1996, 13, 593–600.

53. Pastor, N.; Weinstein, H.; Jamison, E.; Brenowitz, M. J Mol Biol 2000, 304, 55–68.

54. Pastor, N.; Weinstein, H. Protein Eng 1995, 8, 543–549.

55. Dickerson, R. E.; et al. EMBO J 1989, 8, 1–4.

56. Dvir, A.; Conaway, J. W.; Conaway, R. C. Curr Opin Genet Dev 2001, 11, 209–214.

57. Guarnieri, F.; Mezei, M. J Am Chem Soc 1996, 118, 8493–8494.

58. Balasubramanian, B.; Pogozelski, W. K.; Tullius, T. D. Proc Natl Acad Sci USA 1998, 95, 9738–9743.

59. Burkhoff, A. M.; Tullius, T. D. Cell 1987, 48, 935–943.

60. Strahs, D.; Brenowitz, M. J Mol Biol 1994, 244, 494–510.

61. Lee, B.; Richards, F. M. J Mol Biol 1971, 55, 379–400.