Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information

Enhanced Basepair Dynamics Pre-disposes Protein-Assisted Flips of Key Bases in DNA Strand Separation During Transcription Initiation

Neeladri Sekhar Roy[1†], Subrata Debnath[1†], Abhijit Chakraborty[1], Prasenjit Chakraborty[2], Indrani Bera[1], Raka Ghosh[2], Nanda Ghoshal[1], Saikat Chakrabarti[1] and Siddhartha Roy[2*]

[1]Division of Structural Biology and Bioinformatics, CSIR-Indian Institute of Chemical Biology, 4, Raja S.C.Mullick Road, Kolkata 700032, India

[2]Department of Biophysics, Bose Institute, P-1/12, C.I.T. Scheme VII M, Kolkata 700054, India.

Modeling of protein-DNA complex

The crystallographic structure of σ70 subunit of E. coli RNA polymerase (PDB id : 4IGC) was docked to the gal promoter model with expert interface of HADDOCK 2.1 [1]. HADDOCK is an information-driven flexible docking tool for modeling of biomolecular complexes. The gal promoter was modeled using 3D-DART[2]. The bases from −30 to −1 were modeled for gal promoters. For introduction of bending in the DNA models, twist, roll and slide base pair step parameters for the TATA bases (corresponding to TATG in gal) were changed within conformational space allowed for these bases[3, 4]. In another DNA model, -11A was flipped out by changing the base pair parameter of -10 , -11 and -12 bases. For docking with the gal promoter, residues 418, 414, 423, 428, 432, 433 and 454 of RNA polymerase were considered as active residues whereas bases −10, −11 and −14 were considered active residues for the gal promoter. Both the DNA molecules were considered as fully flexible. Docking was performed in solvated mode with water as the solvent. The docking protocol consists of three steps, a rigid-body energy minimization, a semi-flexible refinement in torsion angle space and a final refinement in explicit solvent. In expert interface, during the rigid body energy minimization, 1000 structures were calculated and the best 200 solutions, based on the intermolecular energy, were used for the semi-flexible simulated annealing followed by an explicit water refinement [5]. The final results were grouped in to different clusters and the final pose was selected manually by visualizing intuitively.

DNA model generation for Molecular dynamics simulation

Gal promoter ($P1^{+}P2^{-}$) DNA sequence (TTCGTTGCT**A**$_{-11}$TGGTTATTTCA and its complementary sequence) was modeled using Nucleic Acid Builder (NAB) package [6], which includes the AMBER implementation of the generalized Born model for solvation effects [7]. NAB

uses three principal techniques to build the DNA molecule. The first one is the base transformation where the DNA bases are laid out to achieve the desired helical and base-pairing configuration followed by the addition of sugar backbone and optimization using molecular mechanics energy minimization procedure. The second important part is the utilization of the distance geometry, allowing the DNA structure to satisfy sets of distance constraints. Once the initial model is constructed, the third part performs the optimization and minimization of the model using molecular dynamics simulation under AMBER force field [8]. Following the same protocol five random DNA structures were generated by maintaining the similar length and frequency distribution of nucleotides observed in the original Gal promoter DNA sequence (wild type Gal promoter and random DNA sequences/models are shown in Figure S7). The final models were used for subsequent structural studies.

Molecular dynamics simulation of DNA structures

Molecular dynamics (MD) simulations of all the six DNA models were carried out using GROMACS 4.6.1 simulation package [9]. For all of the cases, Amber FF99SB force field [10] was used for the MD simulations. At the first step DNA models were solvated in a cubic box of 9,277 TIP3P water molecules [11]. Upon solvation 52 Na^+ ions were added in the respective DNA systems to achieve charge neutrality. The final system constituting 29,544 atoms were then subjected to a two-step minimization procedure through steepest descent [9] and conjugate gradient [12] algorithms, followed up with six equilibration steps of 1 nanosecond (ns) each at 300° Kelvin (K). In each equilibration step the force constant was gradually decreased from 100,000 kJ mol^{-1} nm^{-1} to 0 kJ mol^{-1} nm^{-1}. The final simulations were carried out under NPT conditions for 1 microsecond (μs) at 300° K and pressure 1 bar. A 1 femtosecond time step was used for integrating the equations of motion using leap-frog integrator [13]. All the MD simulations were

carried out under periodic boundary condition [14]. Coulombic interactions were treated with Particle Mesh Ewald (PME) summation method [15] whereas van der Waals interactions were treated with cut-off function [10]. Temperature and pressure coupling of the whole system were handled using v-rescale [10] and Parrinello-Rahman [16] algorithms, respectively. In total, we have carried out three independent 1 μs wild type promoter DNA and five 1 μs random DNA MD simulations. For DNA base property analysis MD simulation trajectory starting from 100 nanoseconds (ns) to 1 μs were considered. The first 100 ns trajectories were considered as part of equilibration stage of the whole system.

DNA base property analysis

All the sequence dependent variation in the wild type and random DNA models were analyzed using NUPARM program [17]. Base step parameters like rise, slide, shift, tilt, roll, twist; base pair parameters including shear, stretch, stagger, opening, propeller, buckle and intra-base pair parameters C8-C6, C1-C1 distances were calculated to compare the time dependent structural variations of bases among the wild type and random DNA models (Figure S8).

To compare the intra (within the same DNA model) and inter (between wild type and random DNA models) DNA sequence dependent base properties, a normalized Z_{score} distribution was first calculated. The Z_{score} distribution of individual base properties were calculated from all the MD simulation runs (8 in total) by the following way,

$$Z_{score} = \frac{X - \mu}{\sigma}$$

Where, X is the base property value; μ is the mean of the base property values and σ is the standard deviation of the base property population. In this way we have calculated Z_{score} distribution for 14 different base properties separately (Figure S9).

Following the Z_{score} calculation of individual base properties, the Fisher's exact test [18] was performed using the following contingency tables to calculate the statistical significance of the deviation from the null hypothesis. The contingency tables and their associated null hypothesis are defined as follows,

For intra DNA sequence dependent base property p_{value} calculation:

No. of times having property Z_{score}	**DNA base position "*N*"**	**DNA base position "*not-N*"**
≥ 3 and ≤ -3	a	b
≥ 2 and ≤ -2	c	d

The associated null hypothesis is devised as there is no difference of proportion of a base property between the base position "*N"* and the rest of the base positions "*not-N"* having Z_{score} ≥ 3 and ≤ -3.

In case of inter DNA sequence dependent base property p_{value} calculation:

No. of times DNA base position "*N*" having Z_{score}	**DNA from MD simulation run X**	**DNA from rest of the MD simulation runs**
≥ 3 and ≤ -3	a	b
≥ 2 and ≤ -2	c	d

The associated null hypothesis is as there is no difference in proportion of a property between the base position "*N"* of DNA from run X compared to the same base position from the five random DNA models observed in five different MD simulation runs having $Z_{score} \geq 3$ and ≤ -3.

The p_{value} of the associated contingency tables are given by the following hypergeometric distribution,

$$p_{value} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}$$

The hypothesis is tested against 99% confidence level and in every case Odds Ratio (OR) > 1 [19] is considered for hypothesis testing. The OR is defined as,

$$OR = \frac{a/c}{b/d}$$

An OR > 1 indicates higher odds of outcome associated with the base property of base position "*N*" and MD simulation run X.

Dihedral angle based Principal Component Analysis of DNA

The base opening angle, which is defined as the angle between the lines of C1'-C8 and C1'-C6 atoms of a nucleotide sugar ring in a paired base, measures the tendency of base opening. Changes in C1'-C1' distance from equilibrium suggest an increasing or decreasing inter-chain distance between two paired bases leading to altered base pairing stability.

The Principal Component Analysis or PCA is a multivariate statistical technique that uses a linear transformation technique to diagonalize a covariance matrix of a data set to remove the linear correlations among the variables into a set of uncorrelated variables [20]. The diagonalization procedure generates a set of eigenvalues and ordering these eigenvalues decreasingly, captures

the system's fluctuation also known as principal components. The PCA in Cartesian space involves many degrees of freedom and to overcome this a more natural choice is to use internal coordinates such as the dihedral angles [20, 21], which shows more changes than bond length and angle and less degrees of freedom. In this analysis we have used the three independent 1 μs DNA MD simulation runs to generate a covariance matrix of which contains the circular movement data of DNA backbone dihedral angles, namely the α ($O3'_{[i-1]}$-P-O5'-C5'), β (P-O5'-C5'-C4'), γ (O5'-C5'-C4'-C3'), δ (C5'-C4'-C3'-O3'), ε (C4'-C3'-O3'-$P_{[i+1]}$), ζ (C3'-O3'-$P_{[i+1]}$-$O5'_{[i+1]}$). Here "i" represents the DNA base position. The dihedral angular movement distribution, generation of covariance matrix, diagonalization of the covariance matrix and further analysis was carried out using the inbuild g_angle, g_covar and g_anaeig functions GROMACS package [9], respectively.

Base properties

NUPARM program [17] uses purine and pyrimidine ring atoms to calculate the base normals. The mean base pair normal is considered as the average of the purine and pyrimidine normals, in order to minimize the differences generated due to the size of the two bases during property calculations. The Z axis is defined as the 5' to 3' direction of strand "I" while the Y axis is taken as pointing towards this strand. The X axis is considered as the direction pointing towards the major groove of the DNA. The midpoint of C6 and C8 atoms of purine and pyrimidine bases are defined as the base pair center and the Y axis is considered to be along the C6-C8 direction and passes through the base pair center. All the local helix and wedge parameters are defined in terms of the local helix axis and mean Z axis, respectively for the doublet involved.

Table S1: Basepair lifetimes of promoter region of the Gal promoter

Base pair	Lifetime (ms)
+1	30
-1	*
-2	8
-3	**
-4	48
-5	29
-6	7
-7	7
-8	*
-9	*
-10	7
-11	7
-12	12
-13	*
-14	Very rapid
-15	7
-16	38
-17	*
-18	*
-19	12
-20	8

Basepairs marked * are assigned to the most upfield group of slowly recovering peaks upon inversion. They are GC peaks with long basepair lifetimes. ** Slow relaxing, but the rate cannot be measured accurately due to overlap.

Table S2: Rate of fluorescence change of +3, 2-AP substituted templates

Oligonucleotide	Rate Constant /s
p1+p2+p3+	1.77
p1-p2-p3+	0.18
p1+p2-p3+	0.872
p1(-14)p2-p3+	0.105
p1(-11)p2-p3+	0.391

Table S3: Effect of σ^{70} amino acid substitutions on the rate of fluorescence change of +3, 2-AP substituted templates

(σ^{70}) Alanine substituted RNAP	**Rate Constants /s**
Wild type RNAP (p1+p2$^-$)	0.872 X 10^{-3}
414A	Not detectable
418A	0.120 X 10^{-3}
423A	0.099 X 10^{-3}
426A	0.701 X 10^{-3}
429A	Not detectable
430A	0.196 X 10^{-3}
432A	0.46 X 10^{-3}
433A	Not detectable
436A	Not detectable
437A	Not detectable
454A	Not detectable
455A	Not detectable
458A	Not detectable

Table S4: Sequences of Aro F oligonucleotides

Bases marked in Red are the ones that are mutated

NOMENCLATURE	SEQUENCE
Aro F wt F	**5'GAAAACTTTACTTTATGTGTTATCGTTACGTCA(+1)TCCT CGCTGAGGATCAACTATCGCAAACGA-3'**
AroF wt R	**5'TCGTTTGCGATAGTTGATCCTCAGCGAGGATGACGTAA CGATAACACATAAAGTAAAGTTTTC-3'**
AroF(T-7G) F	**5'GAAAACTTTACTTTATGTGTTATCGGTACGTCATCCTCG CTG AGG ATC AAC TAT CGC AAA CGA-3'**
AroF(T-7G) R	**5'TCGTTTGCGATAGTTGATCCTCAGCGAGGATGACGTAC CGATAACACATAAAGTAAAGTTTTC -3'**
AroF(G-8T) F	**5'GAAAACTTTACTTTATGTGTTATCTTTACGTCATCCTCG CTG AGG ATC AAC TAT CGC AAA CGA-3'**
AroF(G-8T) R	**5'TCGTTTGCGATAGTTGATCCTCAGCGAGGATGACGTAA AGATAACACATAAAGTAAAGTTTTC- 3'**
AroF(C-9A) F	**5'GAAAACTTTACTTTATGTGTTATAGTT ACG TCA TCC**

	TCG CTG AGG ATC AAC TAT CGC AAA CGA-3'
AroF(C-9A) R	**5'TCGTTTGCGATAGTTGATCCTCAGCGAGGATGACGTAA CTATAACACATAAAGTAAAGTTTTC -3'**
AroF(T-10G) F	**5'-GAAAACTTTACTTTATGTGTTAGCGTT ACG TCA TCC TCG CTG AGG ATC AAC TAT CGC AAA CGA-3'**
AroF(T-10G) R	**5'TCGTTTGCGATAGTTGATCCTCAGCGAGGATGACGTAA CGCTAACACATAAAGTAAAGTTTTC -3'**
AroF(A-11C) F	**5'GAAAACTTTACTTTATGTGTTCTCGTTACGTCATCCTCG CTG AGG ATC AAC TAT CGC AAA CGA-3'**
AroF(A-11C) R	**5'TCGTTTGCGATAGTTGATCCTCAGCGAGGATGACGTAA CGAGAACACATAAAGTAAAGTTTTC -3'**
AroF(T-12G) F	**5'GAAAACTTTACTTTATGTGTGATCGTTACGTCATCCTCG CTG AGG ATC AAC TAT CGC AAA CGA-3'**
AroF(T-12G) R	**5'TCGTTTGCGATAGTTGATCCTCAGCGAGGATGACGTAA CGATCACACATAAAGTAAAGTTTTC -3'**
AroF(T-13G) F	**5'GAAAACTTTACTTTATGTGGTATCGTT ACG TCA TCC**

	TCG CTG AGG ATC AAC TAT CGC AAA CGA-3'
AroF(T-13G) R	**5'TCGTTTGCGATAGTTGATCCTCAGCGAGGATGACGTAA CGATACCACATAAAGTAAAGTTTTC -3'**
AroF(G-14T) F	**5'GAAAACTTTACTTTATGTTTTATCGTTACGTCATCCTCG CTG AGG ATC AAC TAT CGC AAA CGA -3'**
AroF(G-14T) R	**5'TCGTTTGCGATAGTTGATCCTCAGCGAGGATGACGTAA CGATAAAACATAAAGTAAAGTTTTC -3'**
AroF(T-15G) F	**5'GAAAACTTTACTTTATGGGTTATCGTTACGTCATCCTCG CTG- AGG ATC AAC TAT CGC AAA CGA-3'**
AroF(T-15G) R	**5'TCGTTTGCGATAGTTGATCCTCAGCGAGGATGACGTAA CGATAACCCATAAAGTAAAGTTTTC -3'**

Table S5: Sequences of the PurMN oligonucleotides

Bases marked in Red are the ones that are mutated

NOMENCLATURE	SEQUENCE
PurMN WT F	**5'CAAACGTTTGCTTTCCCTGTTAGAATTGCGCC*G(+1)*AAT TTTATTTTTCTACCGCAAGTAACGCGT-3'**
PurMN WT R	**5'ACGCGTTACTTGCGGTAGAAAAATAAAATTCGGCGCAA TTCTAACAGGGAAAGCAAACGTTTG -3'**
PurMN(T-7G) F	**5'CAAACGTTTGCTTTCCCTGTTAGAAGTGCGCCGAATTTT ATTTTTCTACCGCAAGTAACGCGT-3'**
PurMN(T-7G) R	**5'ACGCGTTACTTGCGGTAGAAAAATAAAATTCGGCGCAC TTCTAACAGGGAAAGCAAACGTTTG 3'**
PurMN(A-8C) F	**5'CAAACGTTTGCTTTCCCTGTTAGACTTGCGCCGAATTTT ATTTTTCTACCGCAAGTAACGCGT-3'**
PurMN(A-8C) R	**5'ACGCGTTACTTGCGGTAGAAAAATAAAATTCGGCGCAA GTCTAACAGGGAAAGCAAACGTTTG- 3**

PurMN(A-9C) F	**5’CAAACGTTTGCTTTCCCTGTTAGCATTGCGCCGAATTTTATTTTTCTACCGCAAGTAACGCGT-3’**
PurMN(A-9C) R	**5’ACGCGTTACTTGCGGTAGAAAAATAAAATTCGGCGCAATGCTAACAGGGAAAGCAAACGTTTG -3**
PurMN(G-10T) F	**5’CAAACGTTTGCTTTCCCTGTTATAATTGCGCCGAATTTTATTTTTCTACCGCAAGTAACGCGT-3’**
PurMN(G-10T) R	**5’ACGCGTTACTTGCGGTAGAAAAATAAAATTCGGCGCAATTATAACAGGGAAAGCAAACGTTTG -3**
PurMN(A-11C) F	**5’CAAACGTTTGCTTTCCCTGTTCGAATTGCGCCGAATTTTATTTTTCTACCGCAAGTAACGCGT-3’**
PurMN(A-11C) R	**5’ACGCGTTACTTGCGGTAGAAAAATAAAATTCGGCGCAATTCGAACAGGGAAAGCAAACGTTTG- 3'**
PurMN(T-12G) F	**5’CAAACGTTTGCTTTCCCTGTGAGAATTGCGCCGAATTTTATTTTTCTACCGCAAGTAACGCGT-3’**
PurMN(T-12G) R	**5’ACGCGTTACTTGCGGTAGAAAAATAAAATTCGGCGCAATTCTCACAGGGAAAGCAAACGTTTG -3’**

PurMN(T-13G) F	**5’CAAACGTTTGCTTTCCCTGGTAGAATTGCGCCGAATTTT ATTTTTCTACCGCAAGTAACGCGT-3’**
PurMN(T-13G) R	**5’ACGCGTTACTTGCGGTAGAAAAATAAAATTCGGCGCAA TTCTACCAGGGAAAGCAAACGTTTG -3’**
PurMN(G-14T) F	**5’CAAACGTTTGCTTTCCCTTTTAGAATTGCGCCGAATTTT ATTTTTCTACCGCAAGTAACGCGT-3’**
PurMN(G-14T) R	**5’ACGCGTTACTTGCGGTAGAAAAATAAAATTCGGCGCAA TTCTAAAAGGGAAAGCAAACGTTTG -3’**
PurMN(T-15G) F	**5’CAAACGTTTGCTTTCCCGGTTAGAATTGCGCCGAATTTT ATTTTTCTACCGCAAGTAACGCGT-3’**
PurMN(T-15G) R	**5’ACGCGTTACTTGCGGTAGAAAAATAAAATTCGGCGCAA TTCTAACCGGGAAAGCAAACGTTTG -3’**

Table S6: *gal* Promoter oligos for 4ºC titration:

NOMENCLATURE	SEQUENCE
***galP1*$^{+}$*P2*$^{-}$*P3*$^{+}$ F**	5'-TTT TCG CAT CTT TTC GTT GCT ATG GTT ATT TCA TAC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'
***galP1*$^{+}$*P2*$^{-}$*P3*$^{+}$ R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AAC CAT AGC AAC GAA AAG ATG CGA AAA-3'

Table S7: 2-AP Containing oligos (2-AP at +3 Position)

Bases marked in Red are the ones that are mutated

NOMENCLATURE	SEQUENCE
***galP1*$^{+}$*P2*$^{+}$ *P3*$^{+}$ -2APF**	5'-TTT TCG CAT CTT TGT TAT GCT ATG GTT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'
***galP1*$^{+}$*P2*$^{+}$ *P3*$^{+}$-2AP R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AAC CAT AGC ATA ACA AAG ATG CGA AAA-3'
***galP1*$^{-}$*P2*$^{-}$*P3*$^{+}$-2AP F**	5'-TTT TCG CAT CTT TTC GTT ACT GCC CCT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'
***galP1*$^{-}$*P2*$^{-}$*P3*-2AP $^{+}$ R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AGG GGC AGT AAC GAA AAG ATG CGA AAA-3'
***galP1*$^{+}$*P2*$^{-}$*P3*$^{+}$ -2AP F**	5'-TTT TCG CAT CTT TTC GTT GCT ATG GTT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'

***galP1*$^{+}$*P2*$^{-}$*P3*$^{+}$ -2AP R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AAC CAT AGC AAC GAA AAG ATG CGA AAA-3'
***galP1*(T-15G)*P2*$^{-}$*P3*$^{+}$-2AP F**	5'-TTT TCG CAT CTT TTC GTG GCT ATG GTT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'
***galP1*(T-15G)*P2*$^{-}$*P3*$^{+}$ -2AP R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AAC CAT AGC CAC GAA AAG ATG CGA AAA-3'
***galP1*(G-14T)*P2*$^{-}$*P3*$^{+}$-2AP F**	5'- TTT TCG CAT CTT TTC GTT TCT ATG GTT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'
***galP1*(G-14T)*P2*$^{-}$*P3*$^{+}$ -2AP R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AAC CAT AGA AAC GAA AAG ATG CGA AAA-3'
***galP1*(C-13A)*P2*$^{-}$*P3*$^{+}$ -2AP F**	5'-TTT TCG CAT CTT TTC GTT GAT ATG GTT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'

***galP1*(C-13A)*P2*$^{-}$*P3*$^{+}$ -2AP R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AAC CAT ATC AAC GAA AAG ATG CGA AAA-3'
***galP1*(T-12G)*P2*$^{-}$*P3*$^{+}$ -2AP F**	5'-TTT TCG CAT CTT TTC GTT GCG ATG GTT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'
***galP1*(T-12G)*P2*$^{-}$*P3*$^{+}$ -2AP R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AAC CAT CGC AAC GAA AAG ATG CGA AAA-3'
***galP1*(A-11C)*P2*$^{-}$*P3*$^{+}$ -2AP F**	5'-TTT TCG CAT CTT TTC GTT GCT CTG GTT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'
***galP1*(A-11C)*P2*$^{-}$*P3*$^{+}$ -2AP R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AAC CAG AGC AAC GAA AAG ATG CGA AAA-3'
***galP1*(T-10G)*P2*$^{-}$*P3*$^{+}$-2AP F**	5'-TTT TCG CAT CTT TTC GTT GCT AGG GTT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'

***galP1*(T-10G)*P2*$^{-}$*P3*$^{+}$ -2AP R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AAC CCT AGC AAC GAA AAG ATG CGA AAA-3'
***galP1*(G-9T)*P2*$^{-}$*P3*$^{+}$ -2AP F**	5'-TTT TCG CAT CTT TTC GTT GCT ATT GTT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'
***galP1*(G-9T)*P2*$^{-}$*P3*$^{+}$ -2AP R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AAC AAT AGC AAC GAA AAG ATG CGA AAA-3'
***galP1*(G-8T)*P2*$^{-}$*P3*$^{+}$ -2AP F**	5'-TTT TCG CAT CTT TTC GTT GCT ATG TTT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'
***galP1*(G-8T)*P2*$^{-}$*P3*$^{+}$ -2AP R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT AAA CAT AGC AAC GAA AAG ATG CGA AAA-3'
***galP1*(T-7G)*P2*$^{-}$*P3*$^{+}$ -2AP F**	5'-TTT TCG CAT CTT TTC GTT GCT ATG GGT ATT TCA T2APC CAT AAG CCT AAT GGA GCG AAT TAT GAG-3'

***galP1*(T-7G)*P2*$^{-}$*P3*$^{+}$ -2AP R**	5'-CTC ATA ATT CGC TCC ATT AGG CTT ATG GTA TGA AAT ACC CAT AGC AAC GAA AAG ATG CGA AAA-3'

Figure S1

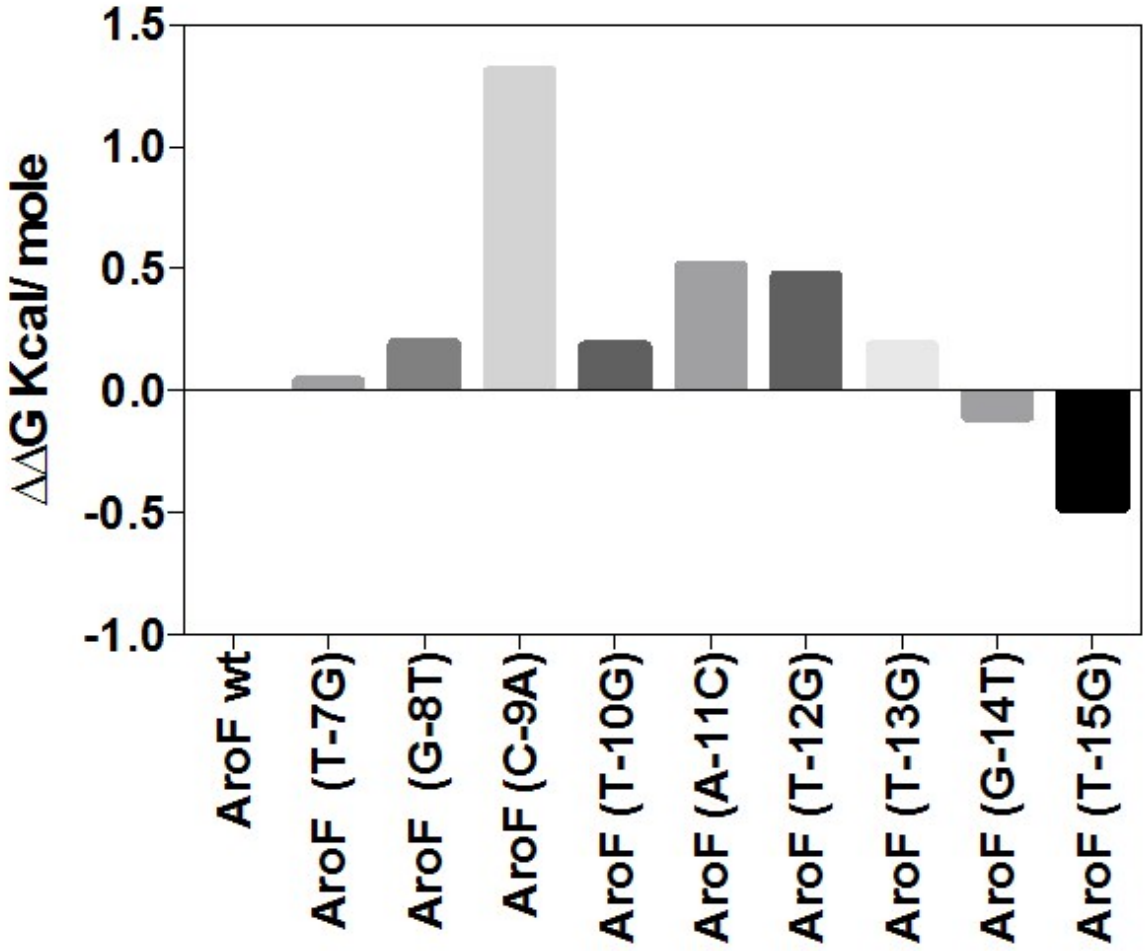


Figure S1. Effect of single base pair mutation on AroF promoter

Figure S2

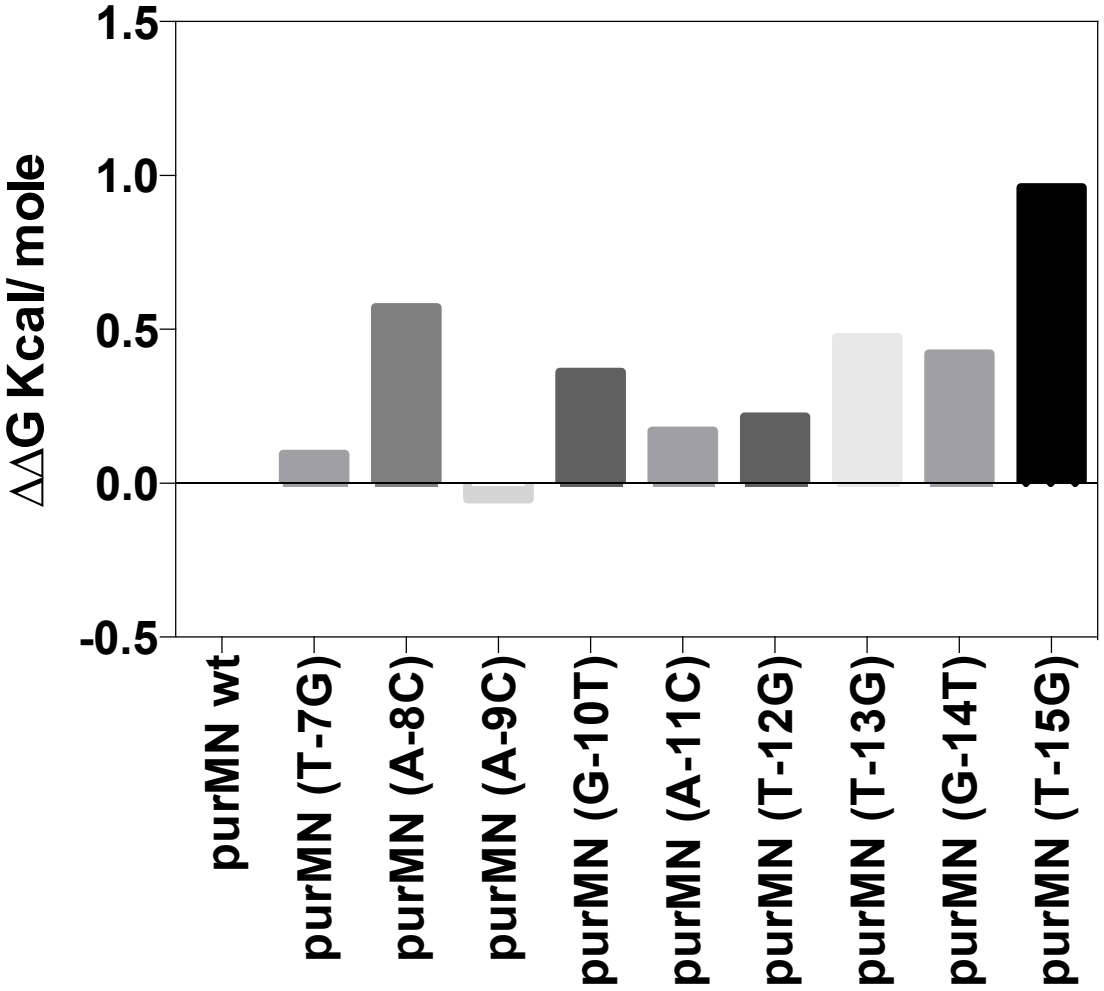


Figure S2. Effect of single base pair mutation on PurMN promoter

Figure S3

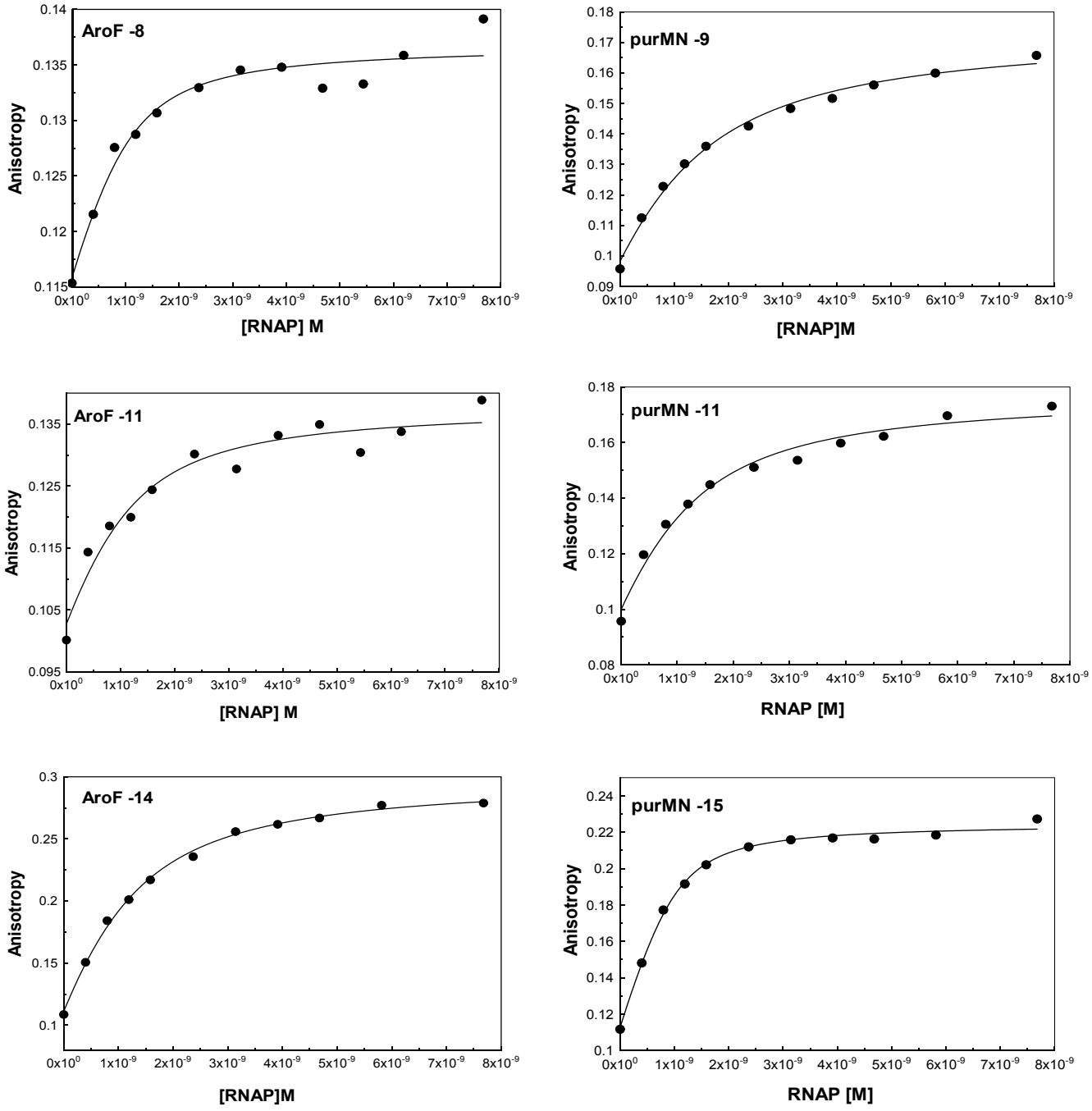


Figure S3. Some representative binding isotherms of wild type and mutant AroF and PurMN promoter sequences with RNA Polymerase.

Figure S4

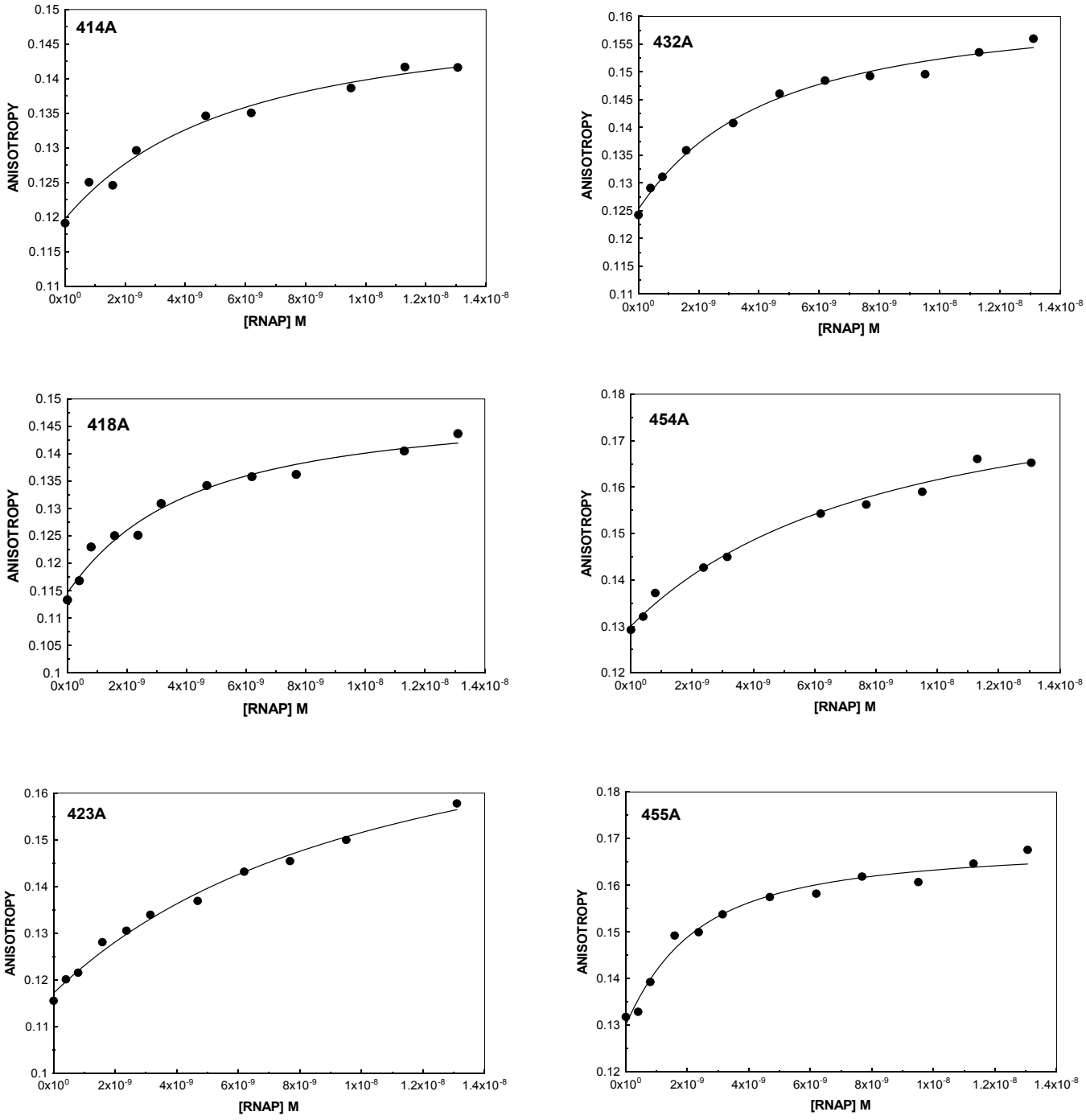


Figure S4. Some representative isotherms of σ70 substituted RNA polymerase and GalP1.

Figure S5

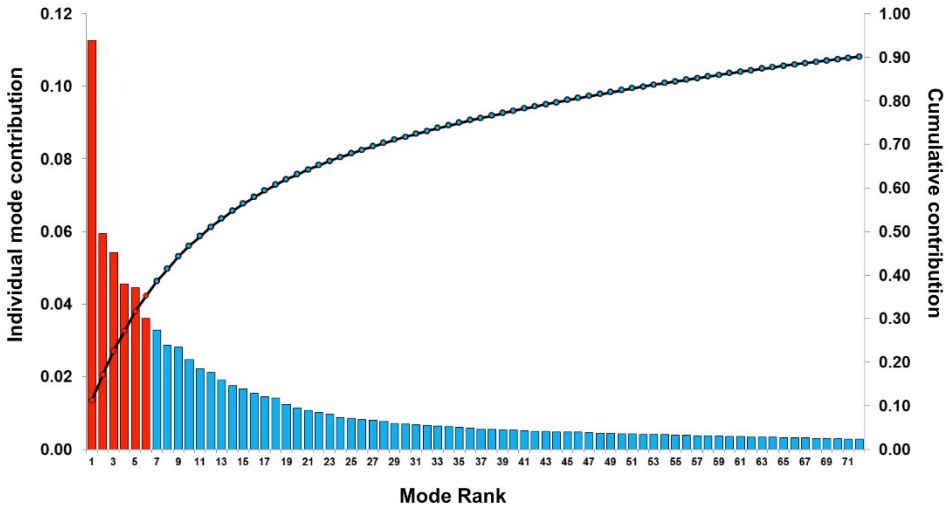


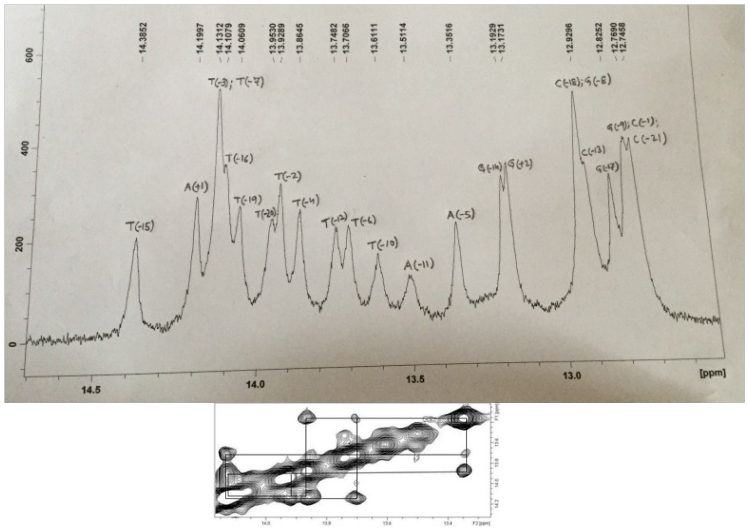
Figure S5. Principal Component Analysis of DNA in backbone dihedral angle space. The panel shows the cumulative distribution (as a line) of internal motions observed over the maximally contributing (about 90%) 72 eigenvalues. The individual eigenvectors (as shown in bar) are sorted decreasingly based on their contribution to the internal motions. First six eigenvalues (Red bars) which represents translational and rotational degrees of motion are excluded from the analysis.

Figure S6.

(A)

5’GCTT CGTTGC TATGG**TTATT** TCAGC
CGAAGCAACGATACC**AATAAA**GTCG5’
-20 -10 +1

(B)



(C)

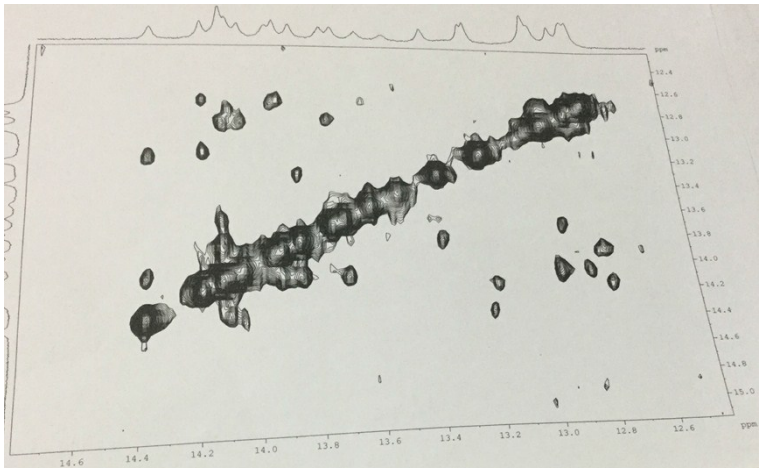


Figure S6. (A) The oligonucleotide containing the galP1 promoter. The purple colored basepair is the transcription start site, that is, the +1 basepair, the red colored basepair is the -10 basepair. (B) NMR spectra of the imino region of the oligonucleotide. The panel below shows the imino-imino connectivity by 2DNOESY between -3 to -7 basepairs. The cross-peaks are aligned with the peaks shown above. The imino spectra contains the assignment of each peak obtained through 2DNOESY. (C) The 2D NOESY spectra of the imino region.

Figure S7

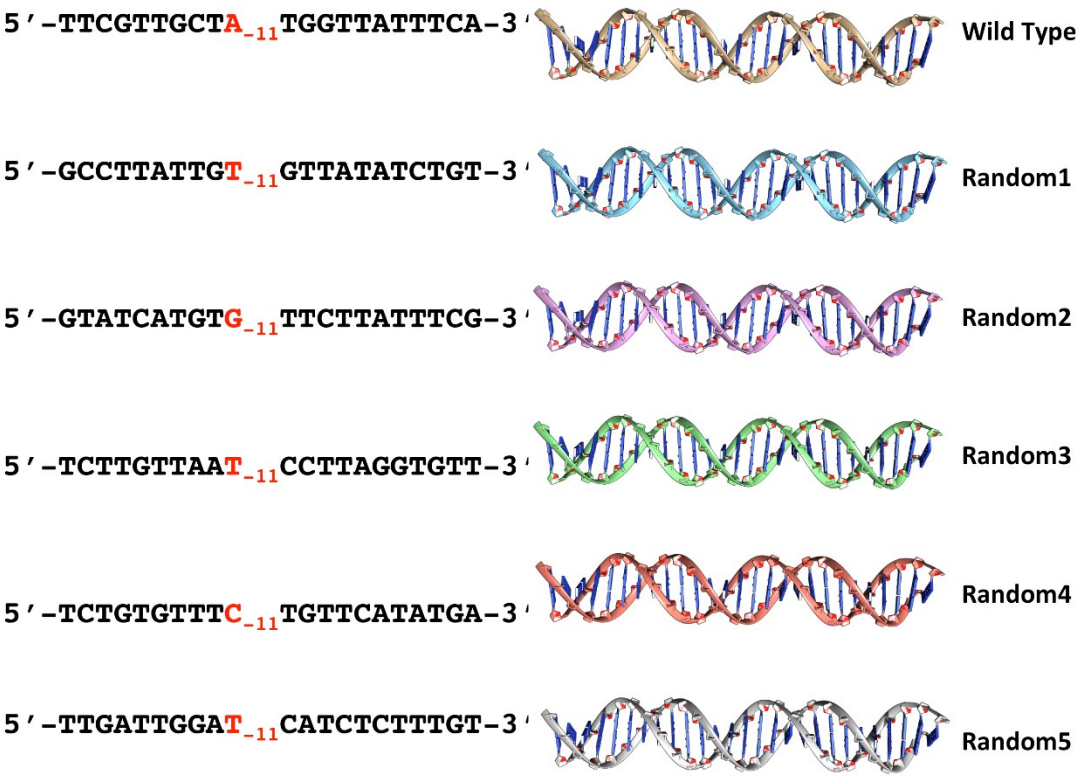
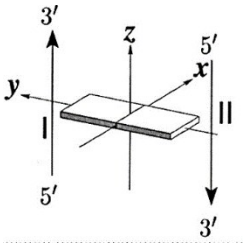


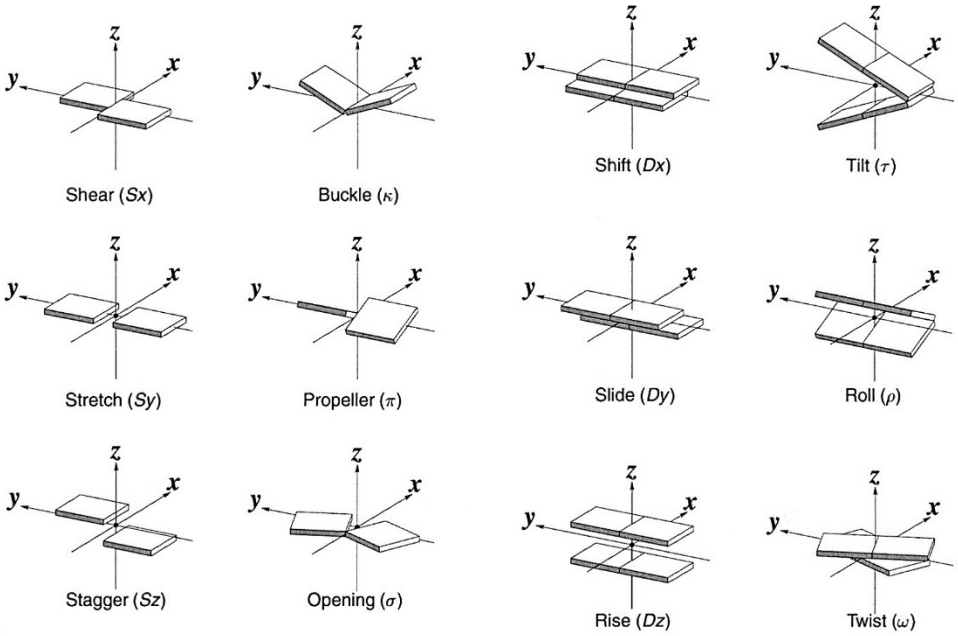
Figure S7. The panel shows the wild type Gal promoter sequence and 3D DNA structure along with the other five random DNA sequences and 3D models.

Figure S8. DNA base properties

Coordinate frame



Base properties



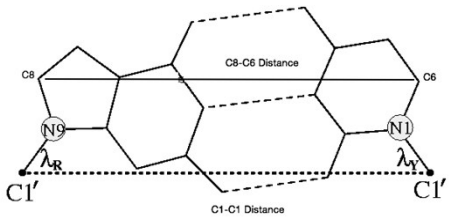


Figure S9. DNA base property Z-score distribution

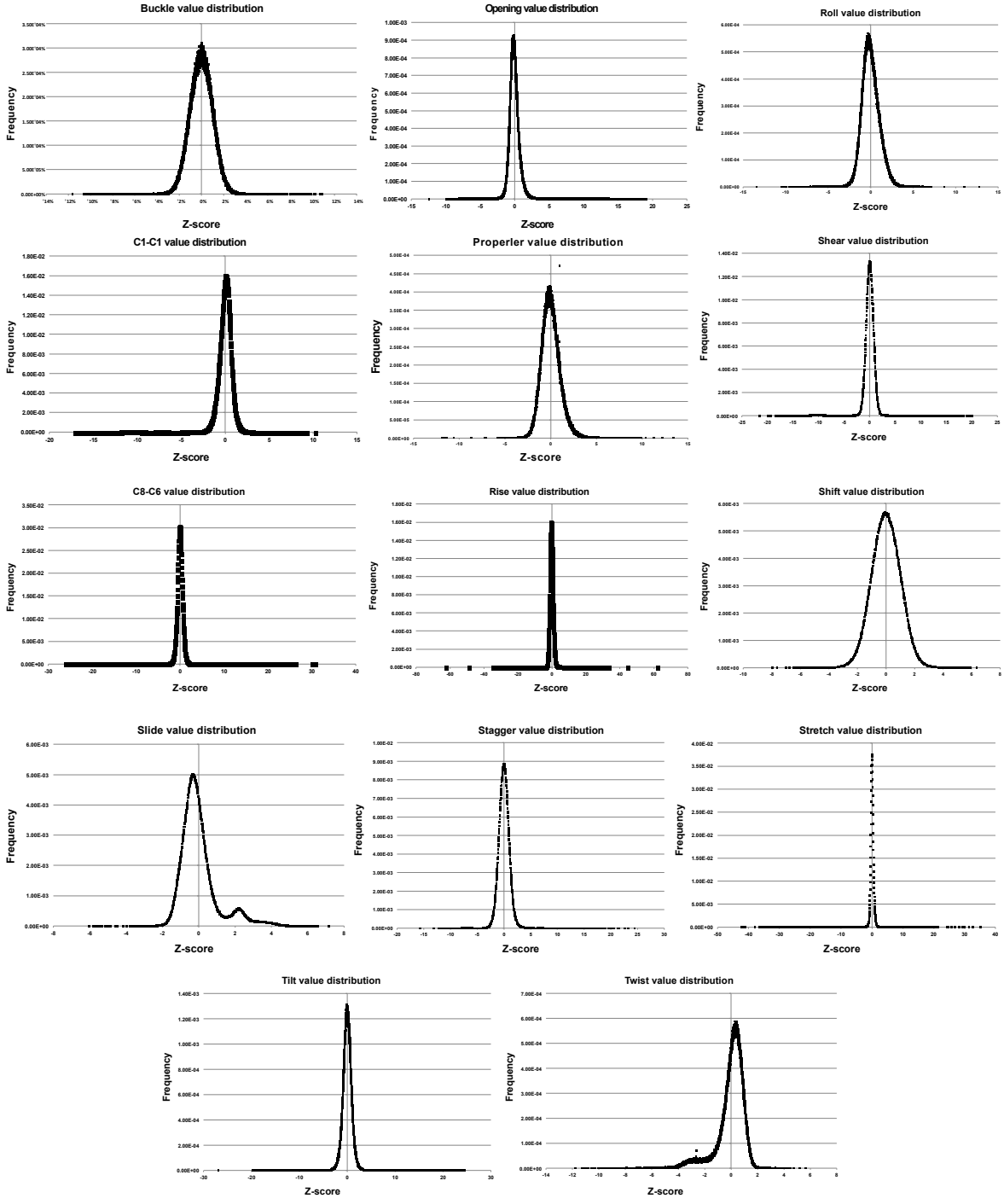


Figure S10

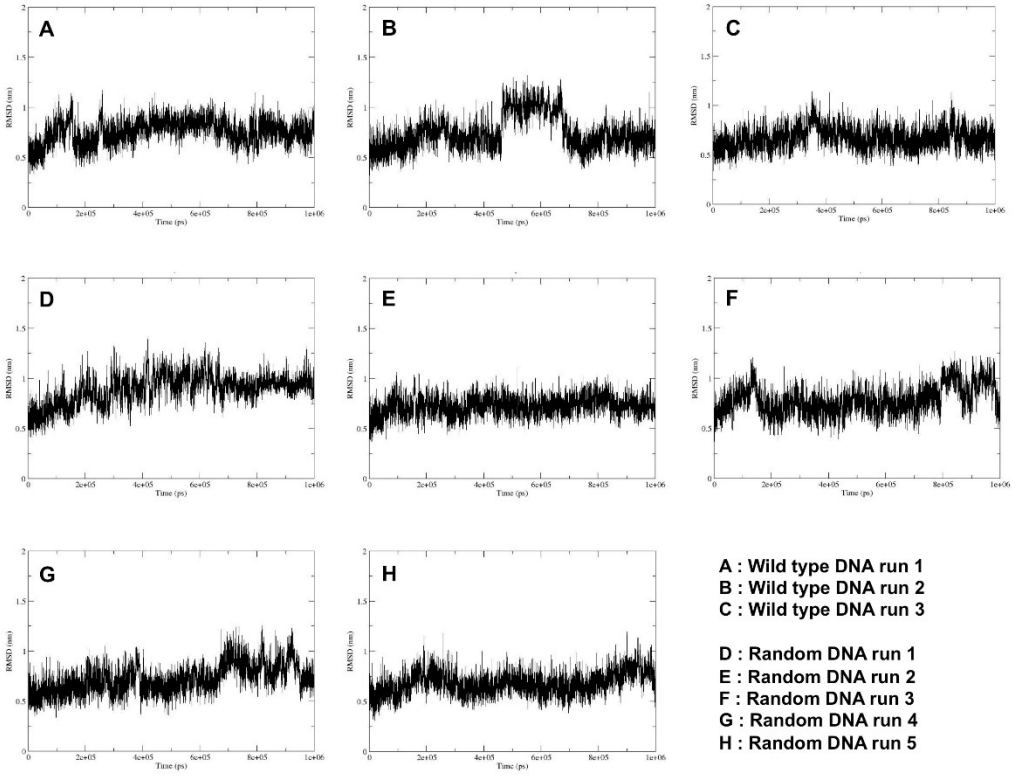


Figure S10. RMSD plots of wild type Gal promoter and random DNA models. Panel A-C shows the RMSD plots for three independent 1 μs MD simulation runs of wild type Gal promoter DNA model. Panel D-H shows the RMSD plots for 1 μs MD simulation runs for five different random DNA models.

References

1. C. Dominguez, R. Boelens and A. M. Bonvin, *J Am Chem Soc*, 2003, **125**, 1731-1737.
2. M. van Dijk and A. M. Bonvin, *Nucleic Acids Res*, 2009, **37**, W235-239.
3. A. A. Travers, *Philos Trans A Math Phys Eng Sci*, 2004, **362**, 1423-1438.
4. T. A. A, *Curr. Opin. Struc. Biol.*, 1991, **1**, 114–122.
5. S. J. de Vries, M. van Dijk and A. M. Bonvin, *Nat Protoc*, 2010, **5**, 883-897.
6. T. J. Macke and D. A. Case, 1998.
7. R. A. Brown and D. A. Case, *Journal of computational chemistry*, 2006, **27**, 1662-1675.
8. T. E. Cheatham and D. A. Case, *Biopolymers*, 2013, **99**, 969-977.
9. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. Berendsen, *Journal of computational chemistry*, 2005, **26**, 1701-1718.
10. E. Darian and P. M. Gannett, *Journal of Biomolecular Structure and Dynamics*, 2005, **22**, 579-593.
11. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *The Journal of chemical physics*, 1983, **79**, 926-935.
12. K. Zimmermann, *Journal of computational chemistry*, 1991, **12**, 310-319.
13. R. Hockney, S. Goel and J. Eastwood, *Journal of Computational Physics*, 1974, **14**, 148-158.
14. D. J. Adams, E. M. Adams and G. J. Hills, *Molecular Physics*, 1979, **38**, 387-400.
15. U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *The Journal of chemical physics*, 1995, **103**, 8577-8593.
16. M. Parrinello and A. Rahman, *Journal of Applied physics*, 1981, **52**, 7182-7190.
17. M. Bansal, D. Bhattacharyya and B. Ravi, *Computer applications in the biosciences: CABIOS*, 1995, **11**, 281-287.
18. A. Agresti, *Statistical science*, 1992, 131-153.
19. N. Scotia, *J Can Acad Child Adolesc Psychiatry*, 2010, **19**, 227.
20. A. Altis, P. H. Nguyen, R. Hegger and G. Stock, *The Journal of chemical physics*, 2007, **126**, 244111.
21. Y. Mu, P. H. Nguyen and G. Stock, *Proteins: Structure, Function, and Bioinformatics*, 2005, **58**, 45-52.