

Profiling the Thermodynamic Softness of Adenoviral Promoters

Chu H. Choi,* Zoi Rapti,*[†] Vladimir Gelev,* Michele R. Hacker,[‡] Boian Alexandrov,[§] Evelyn J. Park,* Jae Suk Park,* Nobuo Horikoshi,[¶] Augusto Smerzi,[§] Kim Ø. Rasmussen,[§] Alan R. Bishop,[§] and Anny Usheva*

*Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts;

[†]Department of Mathematics, University of Illinois at Urbana-Champaign, Urbana, Illinois; [‡]Department of Obstetrics and Gynecology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts; [§]Center for Nonlinear Studies, Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico; and [¶]Radiation Oncology Department, Washington University School of Medicine, St. Louis, Missouri

ABSTRACT We showed previously that anharmonic DNA dynamical features correlate with transcriptional activity in selected viral promoters, and hypothesized that areas of DNA softness may represent loci of functional significance. The nine known promoters from human adenovirus type 5 were analyzed for inherent DNA softness using the Peyrard-Bishop-Dauxois model and a statistical mechanics approach, using a transfer integral operator. We found a loosely defined pattern of softness peaks distributed both upstream and downstream of the transcriptional start sites, and that early transcriptional regions tended to be softer than late promoter regions. When reported transcription factor binding sites were superimposed on our calculated softness profiles, we observed a close correspondence in many cases, which suggests that DNA duplex breathing dynamics may play a role in protein recognition of specific nucleotide sequences and protein-DNA binding. These results suggest that genetic information is stored not only in explicit codon sequences, but also may be encoded into local dynamic and structural features, and that it may be possible to access this obscured information using DNA dynamics calculations.

INTRODUCTION

Gene transcription has been shown to be a very selective process in every organism. On average, there is ~ 1 gene for every several thousand DNA nucleotides, and not all of these have independent start sites (1). To direct transcription at only these sites, placement of the transcribing complex must be tightly regulated. This level of specificity is extraordinary, given that RNA polymerase II, which synthesizes mRNA, is not sequence-specific. Rather, it is only specific to a single-stranded DNA template. It has been shown that protein transcription factors interact with promoter DNA and polymerase to yield specific transcription. The proper identification of these promoter regions is critical to understanding gene expression and regulation.

Researchers have long sought to define what constitutes a gene promoter, and what the minimum requirements are to get a specific transcriptional start site. Approaches using known sequence elements such as CpG islands, TATA boxes, and Initiator (Inr) elements have had some success in extracting promoters from DNA sequences, but these methods are often problematic because of the wide array of promoter types (2–6). They tend to perform well on training sets of promoters, but underperform on other DNA sequences. Even with the advent of position-weighted matrices, hidden Markov models, Bayesian networks, and support vector machines, the results are still not straightforward to apply.

Other methods have attempted to identify promoter regions based on clustering of transcription factor binding sites, which are presumably needed for transcriptional regulation. A recent study showed that the TATA box, the Inr element, the downstream promoter element, and the TFIIB-recognition element were each found in the minority of promoters from a rather extensive data set (7). Surprisingly, the TATA box was found only in 21.8% of 1871 human promoters taken from the eukaryotic promoter database (8,9), and in just 10.4% of 8793 human promoters taken from the database of transcriptional start sites (10,11).

DNA structural dynamics are intricately connected with function. The double helix must be temporarily disrupted during transcription, replication and repair, and to understand the relationship between DNA sequence and genomic function one needs to consider the most probable structural states as well as the available extremes of motion. Both experiments and theory have shown that the strands constantly move in relation to one another, sometimes forming temporary openings, or soft spots, within the archetypal helix (12–15). This phenomenon displays a complex sequence dependence, as different nucleotide sequences can yield similar dynamic properties, and it has been proposed that such soft spots may play an important role in transcription (16–18).

Previous work with the adeno-associated virus P5 promoter showed that although DNA supercoiling and transcription factor (TF) binding are necessary for transcription in the wild-type promoter, a five nucleotide mismatch opening at the transcriptional start site results in bidirectional RNA polymerase II transcription in the absence of those “requisite” factors (19). This evidence strongly suggests an important role for local strand separation in transcription

Submitted October 10, 2007, and accepted for publication March 18, 2008.

Chu H. Choi and Zoi Rapti contributed equally to this work.

Address reprint requests to Chu H. Choi, Tel.: 617-667-1377; E-mail: cchoi@bidmc.harvard.edu.

Editor: Ruth Nussinov.

© 2008 by the Biophysical Society
0006-3495/08/07/597/12 \$2.00

doi: 10.1529/biophysj.107.123471

initiation, and that promoter melting may be sufficient to recruit RNA polymerase in some cases.

Instead of using biological function to extract promoter regions, we wanted to implement a simple physical model to study the physical properties of promoter DNA, to ascertain the role of duplex “softness” in determination of a promoter region. We chose the Peyrard-Bishop-Dauxois (PBD) model as an intuitive physical model to investigate DNA softness (16,17), and chose a transfer integral approach to analyze the statistical mechanics involved for each of these promoters (20,21).

A large component of the PBD Hamiltonian is represented by the hydrogen bonding energy term and therefore the opening probability of DNA is highly sensitive to the overall A-T basepair content. Although A-T (or T-A) basepair density is the chief component in our dynamical profiles, we have also showed that the exact propensity profile displays a complex dependence on the fine-scale distribution of A-T and G-C basepairs. This lends support to a report that some GC-rich DNA sequences can exhibit high rates of opening (22).

The adenoviral genome is ~36,000 basepairs long, but only nine promoters have been identified that regulate expression of 33 known gene products. We report calculations of the DNA dynamic profile these promoters, along with a comprehensive survey of genome regulatory elements derived from published articles and databases. The human adenovirus serotype 5 was chosen for its well-annotated genome and the abundance of functional studies on its viral transcription. We have previously published studies on one of the adenovirus promoters (16,17), and now extend our analysis to all the known promoters in adenovirus 5.

MATERIALS AND METHODS

Human adenovirus serotype 5 genome sequence and annotated data

The complete 35,938 bp sequence of DNA basepairs (Refseq AC_000008) is available from the NCBI Entrez Genome Database. Transcriptional start site and regulatory site locations for the adenoviral genes examined in this work are available in the literature (9,23–25).

DNA probability calculations

The nine known promoters from human adenovirus type 5, with their reported transcriptional start sites (Table 1), were analyzed using a PBD model and the transfer integral operator (TIO) approach to evaluate for average separation displacement of each basepair from equilibrium.

For each promoter, we selected a region for calculations from 230 bp in the 5' direction of transcribed strand to 120 bp downstream of the transcriptional start site (TSS) in the 3' direction. The full sequence was calculated using open boundary conditions, and to eliminate sequence terminal effects, we have thrown out the data for 80 bp on each end of the sequence, leaving us with a 200 bp region stretching from –150 bp to +50 bp relative to the TSS. We also chose two 200 bp nonpromoter control sequences from the genome starting at positions 10,600 and 19,000. The first sequence represents a rare intergenic sequence in adenovirus (over 90% of the genome is coding), whereas the second sequence is from the coding region hexon gene. Both of these controls have no other reported function or binding sites associated with them, making them good models for comparison.

TABLE 1 Human adenovirus 5 promoters with TSS and location on the plus or minus strand

Promoters	TSS	Strand (+/–)
Early		
E1A	499	+
E1B	1702	+
E2E	27051	–
E3	27567	+
E4	35609	–
Late		
E2L	25910	–
IX	3582	+
MLP	6039	+
IVa2	5828	–

TSS, transcriptional start sites.

Double-stranded DNA was modeled using the PBD model as described previously (16). This phenomenological model represents basepair bonding as fitted Morse potentials for one-dimensional A-T and G-C basepairs. Stacking energy between consecutive basepairs is represented by a single parameterized coupling term. The model also incorporates a nonlinear term to account for cooperative effects, such as sugar-phosphate backbone interactions. This nonlinear element effectively models the change in DNA stiffness as the double stranded duplex is opened due to entropic effects.

The potential energy function for basepair n is given by

$$E = \sum_{n=1}^N [V(y_n) + W(y_n, y_{n-1})],$$

where the term for hydrogen bonding and base stacking between basepairs is

$$V(y_n) = D_n(e^{-a_n y_n} - 1)^2$$

and the nonlinear nearest-neighbor coupling is represented as

$$W(y_n, y_{n-1}) = \frac{k}{2}(1 + \rho e^{-b(y_n + y_{n-1})})(y_n - y_{n-1})^2.$$

Morse potential parameter values used were $D_{GC} = 0.075$ eV, $a_{GC} = 6.9$ Å^{–1} for a G-C basepair, and $D_{AT} = 0.05$ eV, $a_{AT} = 4.2$ Å^{–1} for A-T basepairs. For the nonlinear coupling term, $k = 0.025$ eV/Å², $\rho = 2$, and $b = 0.35$ Å^{–1}. All calculations assumed a temperature of 300 K.

This relatively simple model has proven to be effective for describing the opening dynamics of DNA (26). In its current form the PBD model lacks parameters to account for purine- and pyrimidine-specific base stacking (A-T versus T-A) and noncanonical DNA structures. We are currently developing, but have not yet used a modification that would differentiate A-T basepairings from T-A and G-C basepairings from C-G in the DNA sequence. Such a change would not be difficult to implement, and would not significantly impact calculation times, but accurate determination of parameters is underway.

As described previously (20), a thermodynamic partition function was constructed to represent the probability p of an opening of length $n = 10$ basepairs (the given basepair and the 9 subsequent basepairs) all being separated by more than a threshold distance $x_i = 1.5$ Å. Preliminary studies suggested that strand separation distance threshold $x_i = 1.5$ Å and probability of 10 basepairs ($n = 10$) to be displaced simultaneously from equilibrium are good indicators of transcriptionally relevant openings. Using the TIO method to solve this function mathematically (21), we derived the propensity p of an “open” state (according to the parameters n and x_i) for each basepair along the sequence. The resulting probabilities p are not normalized along a given sequence because they are instead normalized to the possible states at each given basepair; to avoid confusion, we will henceforth refer to p as being an opening propensity and not a probability. In any case, relative values of p at each basepair will be comparable, and not cause an evaluative problem.

Because it is a thermodynamics-based calculation, the TIO method yields no information about the timescale of the open/closed state equilibrium, and by design it can not distinguish between different basepair separation distances beyond the specified threshold of 1.5 Å, i.e., it can not measure the relative magnitude of open states. Please note that TIO-derived information is strictly equilibrium-dependent and thermodynamic in nature, but because the properties we focus on derive from the dynamics of double-stranded DNA, we use “dynamic” and “thermodynamic” somewhat interchangeably. Calculations were carried out in parallel using MATLAB 7.2 (The Mathworks, Natick, MA) on nodes of the Harvard Medical School High Performance Computing Shared Cluster, Orchestra.

Statistical analysis

The mean p was calculated for each binding site, for the binding and non-binding regions within each promoter sequence, and for each 200 basepair sequence. Standard deviations (SD) are reported for each mean. Within each promoter sequence the mean p for the binding regions was compared with the mean p for the two control sequences combined. Similarly, the mean p for the nonbinding regions was compared with the mean p for the two control sequences combined. Each promoter sequence in its entirety was also compared with the control sequences. The data were not normally distributed; thus, the Mann-Whitney U test was used for the statistical comparisons. This statistical test yielded a p -value for each comparison; p -values ≤ 0.05 were considered to be statistically significant.

RESULTS AND DISCUSSION

Because our softness profiles are probabilities and not static structural properties, we focus on local maximum peaks rather than simply on areas of high opening propensity. Our reasoning is that should an opening initiate at a basepair that is not a maximum, cooperativity effects should take advantage of the local energetics, rapidly migrating the opening toward the peak locations. This soft spot in the double helix will remain at these preferred spots until the structure reforms, or else it may be able to nucleate into another, larger opening. Therefore, these bases are the most available “soft” positions that proteins are likely to encounter.

Our hypothesis was that promoter regions, which need to separate for transcription to occur, could show patterns of DNA softness in adenovirus. Calculation of two negative control sequences shows muted propensity profiles, with average propensity values (p) not far above the baseline (Fig. 1). Comparison with promoter propensity profiles shows curves with more and higher peaks (Figs. 2 and 3). This observation suggests that promoter DNA may exhibit unique opening propensity patterns. When the overall area under the curve, or the sum of all data points p , is compared for the promoter and the control sequences, we find that the promoters do show more softness in general, as well as lower %GC content than the control regions (Table 2). As expected, the more G-C basepairs there are in the sequence, the more stable the duplex becomes. However, that is clearly not the only factor, because sequences with similar %GC values show opening propensities that differ, such as the IX and early region 1A (E1A) promoters. Both control sequences show a tendency toward lower propensity values. The intergenic sequence at 10,600 bp (Control 1) displays the

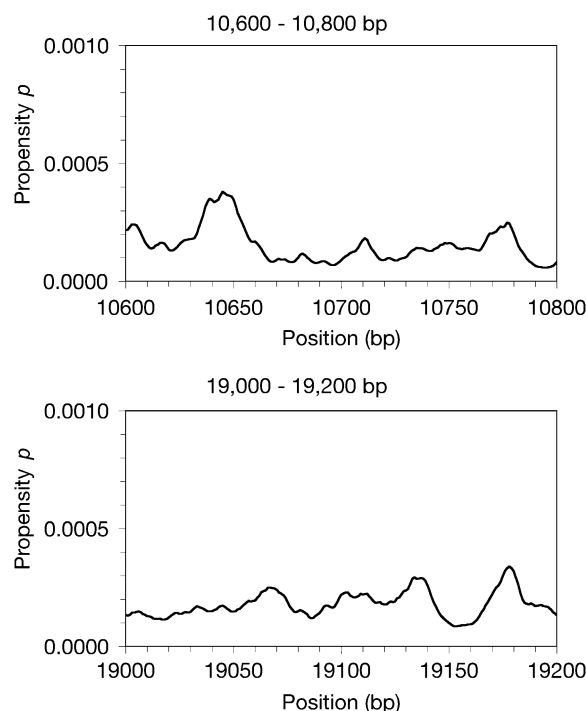


FIGURE 1 Calculated propensity profiles showing soft areas for adenovirus 5 control sequences. A total of 200 bp were analyzed using a PBD-TIO method to calculate propensity of 10 bp openings starting at each basepair to analyze for DNA softness. The x axis reflects the global bp position based on the published genomic sequence.

lowest overall area under the curve, representing the least opening propensity. The coding sequence at 19,000 bp (Control 2) shows more opening propensity than both the major late promoter (MLP) and E3 promoters, but also features a higher overall baseline value with smaller peaks. This higher baseline is likely the result of the sequence at 19,000 bp containing fewer overall G-C basepairs (60% compared to 63% for the sequence at 10,600 bp).

Beyond this general characterization of the promoter sequences, we also found that the promoters typically exhibited 5–6 softness propensity maxima distributed in similar patterns over the 200 bp sequence (Table 3). Using bp

TABLE 2 Quantitative characterization of studied sequences, sorted by greatest area under the propensity curve

Sequence	Area $\times 10^{-4}$ (sum p)	%GC
E4	735.57	44
E1B	657.84	44
IX	610.69	53
E1A	558.30	53
E2E	476.11	55
E2L	409.19	58
IVa2	369.10	60
Control 2	358.72	60
MLP	349.53	60
E3	336.26	61
Control 1	309.53	63

coordinates relative to the transcriptional start site (+1 bp), we found that the peak maxima often appear in similar locations for this collection of promoters. Most significantly, one peak appears in the -55 to -70 region for all nine promoters and another in the region from -15 to -30 bp upstream of the TSS for seven of nine promoters. This softness pattern may play a role in the function of these DNA sequences as promoters.

Although peak profiles for these promoters varied significantly, we found some common features. In addition to transcriptional start sites, we also examined the locations of known transcription factor binding sites in the same promoters (Table 4), using the TRANSFAC database, not to be confused with the TRANSFAC transcription factor binding site prediction engine (25). We have included all major binding sites, as well as sites protected from generalized DNase I digestion, which indicate the binding of some undetermined transcription factor or factors.

It is interesting that the average p of the control DNA sequences differ significantly from that of the MLP (p -value = 0.19) and E3 (p -value = 0.49) promoters (Table 4). The Iva2 promoter exhibits a similar, though not statistically significant, pattern (p -value = 0.03). We observe that the opening propensities alone are not sufficient to predict gene transcription, because there are many other factors and processes involved. This possibly suggests that these less pliable promoters need to be transactivated to become transcriptionally active. Indeed, some of the most active early promoters exhibit the highest overall opening propensities and standard deviations in our study (E1A, E1B, E4), and their expression likely aids in the activation of these other genes.

We have attempted to quantify the matching of reported binding sites with peaks in our propensity profiles in Table 4. If it is true that binding sites often match with these peaks, then the average p of the bases included in those sites should be significantly greater than the average p when compared to the control sequences. We expect that this measure will not work for G-C basepair rich sites, such as Sp1, because the GC-box will likely not open spontaneously relative to the rest of the sequence. Because this data is derived from various experiments with limitations on DNA primer length, some of the recorded binding regions in TRANSFAC are considerably longer than the actual binding site, which complicates the analysis. Rather than assume consensus binding sites are sufficient for binding of factors, we simply use the experimentally corroborated data, and we still observed a correlation between propensity maxima in the DNA profile and transcription factor binding sites and DNase I protection sites reported in the literature.

A statistical comparison of the binding regions within each promoter and the control sequences showed that the average p for the reported binding regions is indeed statistically significantly higher than the average p for the control series in eight of the nine promoters (p -values < 0.0001). Although the differences were less pronounced, the average p for the

nonbinding regions in eight of the nine promoter sequences also differed significantly from the average p for the control sequences (p -values < 0.01). These differences were statistically significant due to the large number of basepairs included in the analyses despite the observation that in all nine promoters the average p for the nonbinding regions was smaller, and thus more similar to that of the control sequences, than the average p for the binding regions. These results suggest that the peak profiles found in the promoter regions are distinct from those found in the control, coding DNA.

E1A promoter

The E1A promoter initiates transcription at bp 499 in the type 5 adenovirus. It has been shown that once viral replication begins, there are multiple alternative start sites 200–300 bp upstream of this site, but this is the primary site for early transcription. The E1A gene product is an important transactivator of viral transcription in adenovirus, and involved with almost all the other promoters.

Our calculated profile for this promoter shows several soft regions within the 200 bp fragment (Fig. 2 *a*). Five major propensity peaks are observed and their intensities listed in Table 3. We note that there are four peaks upstream from the start site, and one peak downstream, with the dominant peak located at -60 bp. Interestingly, some of these peaks correlate with known protein binding sites. Direct gel shift analysis has shown that the cAMP response element-binding (CREB) transcription factor binds the segment from -21 to -52 bp (27), and DNase I experiments in HeLa cell extract show protection of the region from -21 to -38 bp, -45 to -72 bp, and -91 to -99 bp (28), in agreement with three observed peaks in our propensity profile. A nuclear factor 1 (NF-1) binding site ~ 20 bp downstream of the TSS has also been reported for adenovirus 12 (29). This would not be inconsistent with our observed peak at bp 518.

We see that in the E1A promoter, all of the binding sites show relatively higher propensity values when compared with the surrounding sequence, as well as larger SD than both the nonbinding DNA and the controls (Table 4). In a well-studied promoter like E1A, we would expect this kind of good agreement with our model.

Early region 1B promoter

The early region 1B (E1B) promoter initiates specific transcription on the plus strand from bp 1702 in this serotype. We again observe four upstream peaks and one undefined cluster of peaks downstream of the TSS, but this time the most dominant peak is located -136 bp relative to the start site (Fig. 2 *b*). There is a large peak centered at -68 bp, somewhat similar to the E1A promoter.

Again, previous work with the closely related adenovirus type 2 has shown some correlation of our peaks with

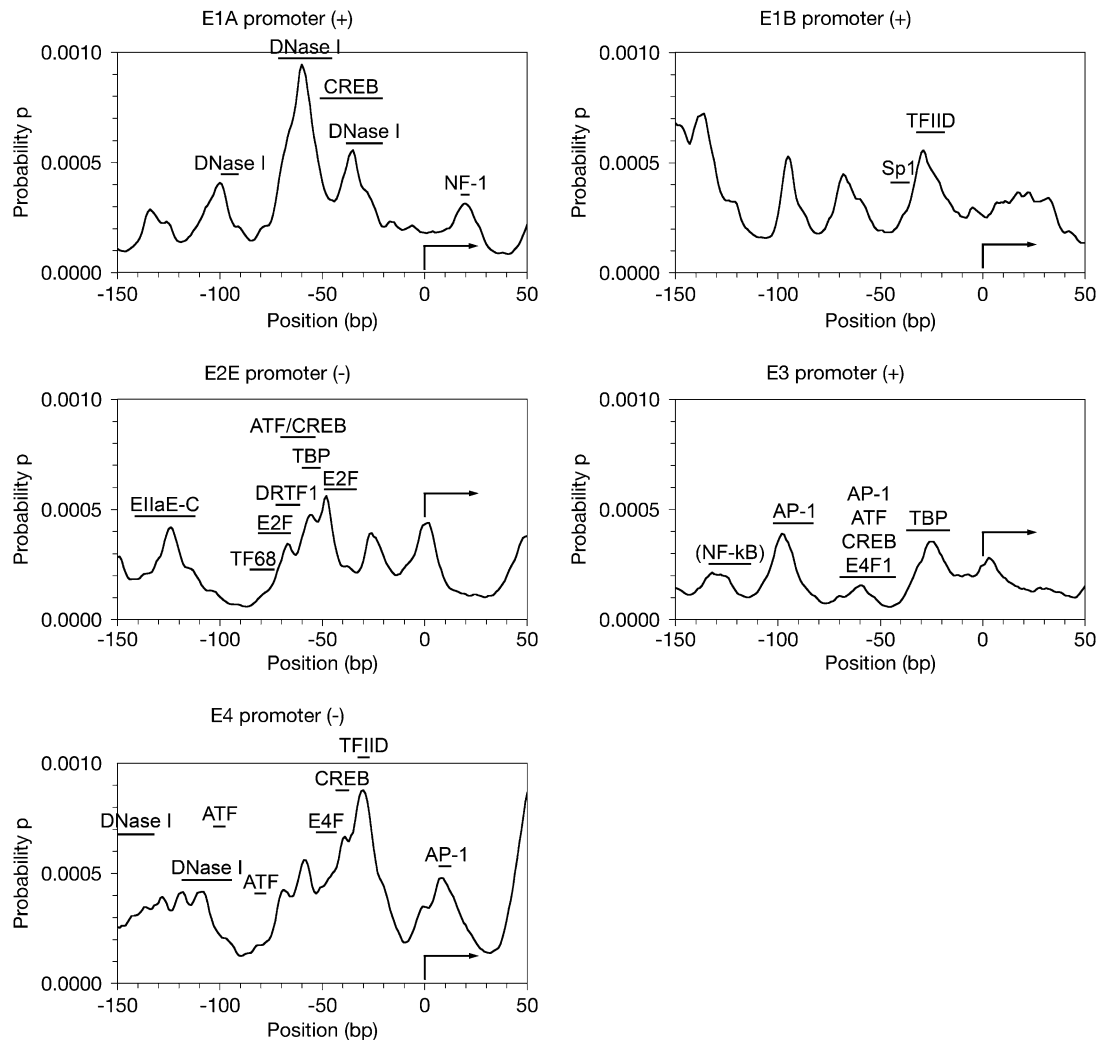


FIGURE 2 Calculated propensity profiles showing soft areas for adenovirus 5 early promoters. A total of 200 bp of the promoter region around the transcriptional start site was analyzed using a PBD-TIO method to calculate propensity of 10 bp openings starting at each basepair to analyze for DNA softness. The x axis reflects the bp position relative to the TSS (+1), which is indicated by the right-facing arrow in the graph. Reported transcription factor binding sites or sites of protection from DNase I are indicated with bars above the softness profile and labeled with the name of the transcription factor. References for the binding sites are included in the Discussion.

reported experiments. Direct gel shift, DNase I footprinting and antibody supershift all suggest that TATA box-binding protein (TBP) and transcription factor IID (TFIID) bind in the locus from -18 to -33 bp (30–32). This area shows a significantly higher average p relative to the rest of the sequence (Table 4). An Sp1 binding site that does not match a propensity peak has been shown at positions -45 to -35 bp using affinity chromatography, DNase I, gel shift competition, and primer extension footprinting (31,33–35). As expected, the average propensity in that GC-box is low, actually lower than that of the rest of the DNA sequence.

Comparison of the binding DNA properties and the non-binding properties show that they are very similar in average p , and the nonbinding DNA actually has a larger SD. As E1B is not well-studied as a promoter, this result could be indicative of additional binding sites that have not yet been

reported. It is unlikely that this promoter contains only two binding sites.

Early region 2 early promoter

The early region 2 early (E2E) promoter directs transcription on the minus strand from bp 27051. Our softness profile maps a major peak at the nucleotide adjacent the TSS, and at least five significant peaks upstream of the start site, with three of them overlapping in the range from ~ -40 to -70 bp (Fig. 2 *c*).

There has been much work done on this promoter in adenovirus 2. Overall, the softness of this promoter region supports our hypothesis, as there is a wealth of binding sites reported for this promoter in the upstream bases. ATF/CREB was shown to bind the promoter at positions -82 to -66 bp and TBP to the sequence from -60 to -51 bp (36), and E2F

TABLE 3 Major propensity profile peaks from human adenovirus type 5 promoters

	Relative bp		
	Max p ($\times 10^{-4}$)	Peak range	Peak max
E1A	2.860	−137 to −124	−134
	4.078	−108 to −96	−100
	9.438	−69 to −53	−60
	5.556	−43 to −26	−35
	3.139	+13 to +26	+20
E1B	7.225	−144 to −130	−136
	5.285	−99 to −91	−95
	4.480	−73 to −57	−68
	5.568	−35 to −17	−29
	3.631	+2 to +36	+23
E2E	4.193	−129 to −119	−124
	3.442	−71 to −63	−67
	4.747	−62 to −52	−56
	5.610	−51 to −44	−48
	3.930	−30 to −19	−26
	4.400	−6 to +10	+2
E3	2.141	−135 to −122	−131
	3.900	−102 to −91	−97
	1.557	−62 to −56	−58
	3.537	−31 to −16	−24
	2.809	−4 to +11	+4
E4	3.927	−134 to −124	−128
	4.117	−123 to −114	−119
	4.161	−113 to −103	−108
	4.251	−73 to −65	−69
	5.604	−64 to −53	−59
	6.667	−44 to −37	−39
	8.770	−36 to −23	−30
	4.784	−1 to +18	+8
E2L	3.266	−147 to −134	−139
	3.185	−133 to −123	−127
	1.972	−89 to −85	−87
	3.743	−64 to −49	−56
	4.028	−29 to −19	−21
	4.157	−18 to −12	−17
IX	2.258	+18 to +27	+22
	4.910	−155 to −128	−145
	6.486	−102 to −89	−95
	7.097	−76 to −59	−70
	4.638	−11 to +10	−1
MLP	2.485	+41 to +48	+46
	1.787	−150 to −138	−144
	2.597	−115 to −102	−109
	3.718	−83 to −62	−70
	2.228	−32 to −18	−28
	1.974	+2 to +20	+8
IVa2	1.958	+44 to +50	+48
	3.718	−149 to −128	−141
	2.597	−112 to −96	−102
	1.769	−73 to −60	−68
	1.915	−6 to −36	−40
	2.542	−11 to +1	−5
	3.189	+2 to +21	+15

Peak range was determined using peak widths at 50% of maximum peak values above an assumed baseline of $p = 0.0001$.

was shown to bind to the sequences from −71 to −53 bp and −49 to −33 bp upstream of the TSS (37). A subsequent study placed TF68 and DRTF1 sites at positions −86 to −73 bp and −73 to −61 bp (38). The TF68 site does not match well with our propensity profile. The ATF/CREB site, TBP site and upstream E2F site correspond well with the three overlapping peaks observed in our profile, but the other two sites are located where there are more rigid segments of DNA. Because the E2F sites have been shown to operate synergistically (39), it is conceivable that initial binding at one site could activate another site. Another study suggests that EIIaE-C binds from −142 to −112 bp upstream of the TSS (40). This could align with the peak centered at −124 bp. Five of seven identified binding sites display a higher average opening propensity p than the rest of the sequence (Table 4).

Early region 3 promoter

The early region 3 (E3) promoter resides on the plus strand, and a specific transcript begins at bp 27,567. In this promoter, our analysis shows one peak just downstream of the TSS, and four significant peaks upstream (Fig. 2 *d*). This early promoter does not show the propensity magnitudes typical of the other early promoter regions.

Again referencing research on the extremely similar adenovirus 2, the E3 promoter is known to bind TFIID from bases −37 to −16 upstream from the TSS (41). The softness peak at −26 bp seems to fit that description. Several binding partners have been identified in the region from −70 bp to −42 bp, including AP-1, ATF, CREB, and E4F1 (27,41), but the softness profile reflects only a small, but significant, widened peak in that region. AP-1 has also been implicated at positions −103 to −83 bp (41), where there is a large propensity peak. Finally, a NF- κ B-like factor has been shown to bind in the area from −134 to −113 bp upstream of the TSS (42), and our profile does show a plateau-like feature with a maximum at −131 bp. Most of the binding sites do show a higher average p than nonbinding DNA, except for the multiple protein site at −70 to −42 bp (Table 4). In our model, that site could require an enhancer or DNA conformation shift to activate protein binding.

E3 is the promoter with the highest overall G-C content and the lowest overall softness, with an average p of just 1.68×10^{-4} . The promoter peak pattern is still discernible, but the DNA is relatively rigid. Because of this, the propensity profile is similar to that of the control DNA (p -value = 0.11) (Table 4).

Early region 4 promoter

The early region 4 (E4) promoter is also located on the minus strand, at bp 35069. Our propensity profile shows one peak downstream of the TSS, and at least seven peaks upstream, though they seem to be overlapped into three main groups (Fig. 2 *e*). The softness profile is striking in its overall high propensity values, and a large region of very soft DNA from

TABLE 4 Quantification and characterization of TRANSFAC-reported binding sites in adenovirus type 5 promoters

Gene	Binding site	Relative location (bp)	Average p ($\times 10^{-4}$)	SD p (10^{-4})	p -value
Control 1	—	—	1.54	0.75	
Control 2	—	—	1.79	0.54	
E1A	DNase I	−99 to −91	2.79	0.75	
	DNase I	−72 to −45	6.00	2.00	
	CREB	−52 to −21	3.87	0.90	
	DNase I	−38 to −21	3.90	1.11	
	NF-1	+20 to +24	3.14	0.24	
	<i>Binding</i>	—	4.55	1.95	<0.0001
	<i>Nonbinding</i>	—	1.92	0.67	<0.0001
	<i>All</i>	—	2.79	1.76	<0.0001
E1B	Sp1	−45 to −35	2.45	0.51	
	TBP/TFIID	−33 to −18	4.55	0.66	
	<i>Binding</i>	—	3.69	1.21	<0.0001
	<i>Nonbinding</i>	—	3.23	1.43	<0.0001
	<i>All</i>	—	3.29	1.41	<0.0001
E2E	E1aE-C	−142 to −112	2.67	0.75	
	TF68	−86 to −73	1.12	0.39	
	ATF/CREB	−82 to −66	1.97	0.91	
	DRTF1	−73 to −61	2.91	0.48	
	E2F	−71 to −53	3.63	0.75	
	ATF/CREB	−60 to −51	4.43	0.33	
	E2F	−49 to −33	3.24	0.75	
	<i>Binding</i>	—	2.78	1.19	<0.0001
	<i>Nonbinding</i>	—	2.09	1.07	0.0028
	<i>All</i>	—	2.38	1.17	<0.0001
E3	NF- κ B	−134 to −113	1.64	0.43	
	AP-1	−103 to −83	2.62	0.92	
	AP-1/ATF/CREB/E4F1	−70 to −42	1.04	0.32	
	TFIID	−37 to −16	2.61	0.75	
	<i>Binding</i>	—	1.90	0.93	0.11
	<i>Nonbinding</i>	—	1.49	0.50	0.01
	<i>All</i>	—	1.68	0.76	0.49
E4	DNase I	−152 to −132	1.74	0.35	
	DNase I	−118 to −94	5.04	0.89	
	ATF	−100 to −94	4.98	0.13	
	ATF	−80 to −74	5.64	0.25	
	E4F	−53 to −43	14.25	0.39	
	CREB	−45 to −37	9.49	0.68	
	TBP/TFIID	−30 to −23	2.61	1.43	
	AP-1	+7 to +13	3.66	0.28	
	<i>Binding</i>	—	3.95	1.64	<0.0001
	<i>Nonbinding</i>	—	3.49	1.79	<0.0001
	<i>All</i>	—	3.68	1.74	<0.0001
E2L	CRF	−155 to −128	2.76	0.41	
	USF-2	−127 to −97	1.88	0.42	
	CRF	−90 to −65	1.42	0.36	
	Sp1	−56 to −36	2.34	0.88	
	<i>Binding</i>	—	2.06	0.72	<0.0001
	<i>Nonbinding</i>	—	2.03	1.00	0.01
	<i>All</i>	—	2.05	0.87	<0.0001
IX	DNase I	−70 to −45	3.66	2.49	
	TBP	−30 to −23	2.70	0.36	
	<i>Binding</i>	—	3.43	2.21	<0.0001
	<i>Nonbinding</i>	—	2.98	1.56	<0.0001
	<i>All</i>	—	3.05	1.69	<0.0001
MLP	USF-1	−120 to −90	1.89	0.36	

(Continued)

TABLE 4 (Continued)

Gene	Binding site	Relative location (bp)	Average p ($\times 10^{-4}$)	SD p (10^{-4})	p -value
	CCAAT-box	−82 to −71	3.36	0.29	
	USF-1	−63 to −52	1.82	0.48	
	TBP	−39 to −17	1.84	0.26	
	Inr	+8 to +26	1.64	0.24	
	<i>Binding</i>	—	2.00	0.61	<0.0001
	<i>Nonbinding</i>	—	1.51	0.65	0.0032
	<i>All</i>	—	1.75	0.68	0.19
IVa2	USF-1	−113 to −89	1.94	0.39	
	Sp1	−50 to −36	1.65	0.17	
	Inr	−9 to +6	2.28	0.16	
	TBP	+10 to +22	2.82	0.44	
	<i>Binding</i>	—	2.12	0.51	<0.0001
	<i>Nonbinding</i>	—	1.71	0.80	0.24
	<i>All</i>	—	1.85	0.74	0.03

~−10 bp to −75 bp. If softness is related to protein binding, we would expect a great number of binding partners in this area.

We have multiple positive alignments of our propensity peaks with protein binding sites for this promoter. There are a number of factors that bind from −152 to −132 bp and −118 to −94 bp (43–45), which is reflective of the multiple peaks we see in our softness profile. ATF binds to the promoter at −100 bp and −80 bp (46), which bookend a softness valley in our profile. E4F has been reported to bind from positions −53 to −43 bp (47), which is very soft. CREB has been shown to bind at −45 bp (27), which is within the very soft, active region of the E4 promoter. DNase I footprinting confirms that TBP/TFIID binds around −30 bp upstream of the start site (46), and the largest profile peak is exactly at −30 bp in our calculated profile. An AP-1 site has been identified at positions +7 to +13 bp downstream of the TSS (48), and that matches our peak at +7 bp. Our opening propensity model suggests that this promoter is highly active, and could consequently be problematic to characterize (Table 4). We do see, however, that there are a host of factors that are known to bind to this DNA, and that is consistent with our hypothesis.

Early region 2 late promoter

The early region 2 late (E2L) is perhaps the least well-characterized promoter region in adenovirus. It initiates transcription on the minus strand from the 25,910 bp. As the name implies, transcription does not occur until the late early stages of the viral life cycle. Again, we might expect the softness profile to differ from our previous examples. It is not as pronounced for another transcript that is not expressed right away, and that appears to be a trend (Fig. 3 *a*). The largest peak is at −17 bp relative to the TSS, though there is a slightly smaller peak at −56 bp.

There is CRF binding reported in the −155 to −128 bp region, a USF-2 site from −127 to −97 bp, another CRF site from −90 to −65 bp and an Sp1 site from −56 to −36 bp in adenovirus 2 (49). The multiple peaks from −127 to −139

might fit the first CRF site, but there is no match for the USF-2 or second CRF sites. The peak with a maximum at −56 bp could possibly match the Sp1 binding site, but it is not a strong match. This promoter does not show good alignment of binding sites with propensity peaks (Table 4), but this is likely due to two of the four sites being GC-rich and the fact that some binding sites are probably still uncharacterized on this promoter.

Protein IX promoter

This promoter, also on the plus strand of the genome, is known to direct transcription from position 3582 bp. This entire gene is contained within the E1B transcription unit, which affects its regulation (50). The E1B and IX promoters share some striking sequence similarities (51). There is a TATA box present in this serotype, and in the related serotype 2, but not in adenovirus 3 (52).

As the sequences are somewhat similar, the IX promoter profile is very similar to that of E1B (Figs. 2 *b* and 3 *b*). Once again, the peak with the highest propensity value is at −70 bp relative to the TSS. A TATA box has been reported at position −30 in this promoter, to which TBP binds and directs transcription (53). There is also some other, unidentified activating upstream element from −45 to −70 bp relative to the TSS. Although these are not exact matches for the observed peaks, the correlation is not inconsistent (Table 4). There may also be unreported binding sites for this promoter; it would be unique if the IX promoter only bound TBP and perhaps one other protein.

MLP

The MLP is the most well-studied of the adenoviral promoters and is known to drive transcription of a major mRNA product that encodes for multiple genes. It initiates transcription at bp 6039 on the plus strand. We see from our calculated profile that the overall softness is less than the early promoters, and also that the largest peak is at centered at

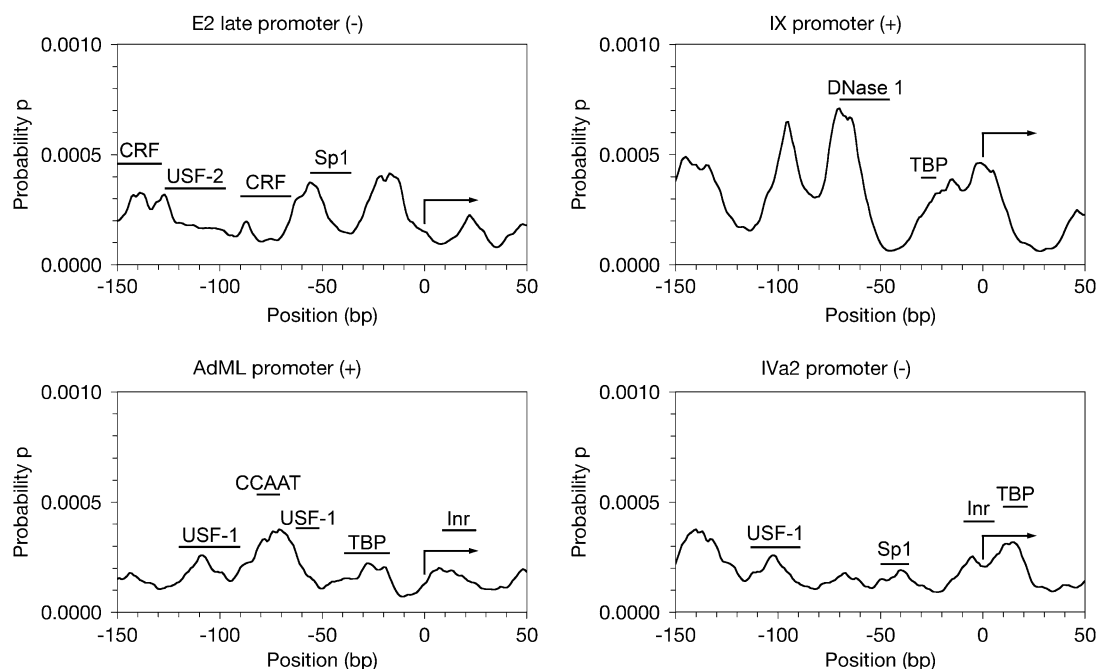


FIGURE 3 Calculated propensity profiles showing soft areas for adenovirus 5 late promoters. A total of 200 bp of the promoter region around the transcriptional start site was analyzed using a PBD-TIO method to calculate propensity of 10 bp openings starting at each basepair to analyze for DNA softness. The x axis reflects the bp position relative to the TSS (+1), which is indicated by the right-facing arrow in the graph. Reported transcription factor binding sites or sites of protection from DNase I are indicated with bars above the softness profile and labeled with the name of the transcription factor. References for the binding sites are included in the Discussion.

5969 bp, or -70 bp relative to the TSS (Fig. 2 *c*). There are four peaks upstream of the promoter and two peaks downstream in this case.

TBP has been identified to bind between positions -17 and -39 bp by DNase I footprinting and gel retardation experiments (54). Many subsequent experiments have reconfirmed this in a myriad of other ways. A large propensity peak is not present in our profile, but there is a soft spot centered on the -28 position. An Inr element has also been identified from positions $+8$ to $+26$ bp on this promoter by DNase I, exonuclease III, and gel retardation (55–57), which would seem to correspond to an asymmetric peak with its maximum at $+7$ bp. An upstream factor 1 (USF-1) site has also been reported from positions -52 to -63 bp (54), which does not match our calculated propensity profile. Another USF-1 binding site is at -90 to -120 bp (58), which could match a peak centered at 5930 bp (position -109 relative to the TSS). Finally, an inverted CCAAT-box has been identified by direct gel shift and methylation interference assays from positions -71 to -82 bp (59–61), where there is a prominent peak in our propensity profile. All of these binding sites exhibit higher average opening propensities than the other DNA, though the profile as a whole is extremely quiet (Table 4).

Despite the relative rigidity of MLP, the peaks are clearly distinguished and their locations appear to be significant. As the most studied promoter in adenovirus, we do not believe that there are binding sites that have not been established. Rather, we speculate that as a late and inducible promoter,

other mechanisms may play a role in activating transcription that do not leverage DNA softness in the same way.

IVa2 promoter

The IVa2 promoter initiates transcription at position 5828 bp on the minus strand. The promoter for the protein IVa2 gene is unique in at least two aspects. First, it is the only promoter that does not have an identifiable TATA box-like element, and second, it is a promoter that is completely inactive until the end of the early stage, when viral replication begins. If DNA softness is indeed related to the behavior of promoter regions, we would expect that this promoter would show different properties than all of the others studied here. The promoter is on the minus strand and located within close proximity to the major late promoter, such that the promoters are arranged in a back-to-back fashion.

The first noticeable characteristic is that the overall softness profile is less than the early promoters we have analyzed (Fig. 3 *d*). There is a small peak present just upstream from the TSS, and again four peaks upstream of the promoter, and there is one peak downstream. Despite the lack of a TATA-like sequence, the IVa2 promoter has a few reported binding sites: USF-1 binds at -113 to -89 bp (62), there is an Sp1 GC-box at -50 to -36 bp (63), an Inr element from -9 to $+6$ bp (55), and TBP actually binds downstream of the TSS from $+10$ to $+22$ bp (64). These binding sites could fit nicely with the peaks centered at -102 , -40 , -5 and $+14$ bp. Three

of the four binding sites show elevated opening propensity over the rest of the sequence, save Sp1, which binds to a GC box (Table 4).

The calculated profiles do share some similarities. The early promoters are, in general, softer pieces of DNA than are the late promoter regions. This can be partially explained by the different nucleotide constitution of the two types of promoters. The five early promoter regions we studied are, on average, 51%GC, compared to 58% for the four late promoters. However, this does not offer a complete explanation, as three of the early promoters were over 50%GC and the E3 promoter actually had the highest GC content of all promoters, at 61%. Four of five early promoters show a clear soft spot in the region from -35 to -25 bp upstream of the TSS, representative of the TATA box. However, only one of the late promoters shows that profile peak, despite three of the four promoters containing TATA boxes. All nine promoters exhibited a significant peak in the region -70 to -55 bp upstream of the promoter, though the propensity values varied somewhat. The promoters with the weakest peaks in this region were the E3 and the IVa2 promoters. From these data, it seems that the sequence-mediated local thermodynamic softness of promoter DNA does seem to have an effect on promoter activity and/or promoter strength.

The alignment of opening propensity peaks with transcription factor binding sites in many cases is an interesting and unexpected result. It is well-established that specific DNA-protein binding in most cases involves significant conformational changes in the binding partners, as in indirect recognition (65). Transcription factor binding sites have been correlated with numerous DNA properties, such as stiffness, A-philicity, deformability, and other structural parameters (66–70). Our results suggest that DNA-protein recognition may correlate with DNA strand separation dynamics. It is possible that “soft spots” with enhanced dynamics also show the conformational plasticity required for induced fit binding. Another possibility is that the enhanced exposure of unpaired DNA strands at those sites establishes initial contacts between the protein and the DNA, resulting in higher association rates. If confirmed in more detailed studies, our findings may suggest a facile method to extract protein binding site data using mathematically simple DNA calculations.

Recently, bulk core promoter sequences have been characterized energetically using known promoter elements and multiple position weight matrices (66). It is interesting that the reported duplex disrupt energy profiles of mammalian promoters, calculated using the Breslauer model (71), mirror the propensity profiles described here. The primary difference is that our method measures softness, whereas their results are plotted on a converse scale of energetic stability, with the TSS positions located within local energy minima. Interestingly, the thousands of promoters clustered into categories exhibited properties similar to the average profile of the promoters we analyze here, suggesting that our descriptors of DNA structure may share a common physical origin.

In contrast to such related methods, however, our PBD approach can be applied to any unique sequence of DNA to determine whether it fits such a profile.

CONCLUSION

Analysis of the known promoter regions of adenovirus type 5 leads us to conclude that there may be common softness patterns in these promoters, though the pattern may differ between early and late-expressed genes. Perhaps trans-activating factors may be necessary to establish the right softness conditions for transcription on late genes. Comparison of the results to functional data for the adenoviral genome strongly supports the idea that genetic information is stored not only in the form of nucleotide primary sequence, but is also encoded into the dynamical structure of DNA. Our calculations suggest that at least some of this secondary structure information can be extracted by simple consideration of the softness of the double helix. In its current version the PBD model is unable to describe DNA dynamics arising from effects other than the distribution of hydrogen bonding potentials along the double helix. A refined model that also accounts for DNA structural inhomogeneity due to patterns of pyrimidine and pyridine rings should add to the accuracy of simulations at the genomic scale.

We thank Priscilla Schaffer's laboratory for helpful discussions. The authors also thank the High Performance Computing Group at Harvard Medical School for their assistance with the Orchestra Shared Research Cluster, and Hendrata Dharmawan for assistance with UNIX scripting and data file manipulation.

The authors acknowledge financial support from the National Institutes of Health-National Institute of General Medical Sciences (grant No. R01 GM073911 to A.U.). Research at Los Alamos is carried out under the auspices of the National Nuclear Security Administration of the U.S. Department of Energy at Los Alamos National Laboratory under Contract No. DE-AC52-06NA25396 to K.R.

REFERENCES

1. International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature*. 431:931–945.
2. Bajic, V. B., M. R. Brent, R. H. Brown, A. Frankish, J. Harrow, U. Ohler, V. V. Solovyev, and S. L. Tan. 2006 Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol.* 7(Suppl 1):S3.1–13.
3. Werner, T. 2003. The state of the art of mammalian promoter recognition. *Brief. Bioinform.* 4:22–30.
4. Ohler, U., and H. Niemann. 2001. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.* 17:56–60.
5. Pedersen, A. G., P. Baldi, Y. Chauvin, and S. Brunak. 1999. The biology of eukaryotic promoter prediction—a review. *Comput. Chem.* 23:191–207.
6. Fickett, J. W., and A. G. Hatzigeorgiou. 1997. Eukaryotic promoter recognition. *Genome Res.* 7:861–878.
7. Gershenson, N. I., and I. P. Ioshikhes. 2005. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics.* 21:1295–1300.

8. Cavin Périer, R., T. Junier, and P. Bucher. 1998. The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.* 26:353–357.
9. Schmid, C. D., R. Perier, V. Praz, and P. Bucher. 2006. EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res.* 34:D82–D85.
10. Suzuki, Y., R. Yamashita, K. Nakai, and S. Sugano. 2002. DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.* 30:328–331.
11. Yamashita, R., Y. Suzuki, H. Wakaguri, K. Tsuritani, K. Nakai, and S. Sugano. 2006. DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. *Nucleic Acids Res.* 34:D86–D89.
12. Levitt, M. 1983. Computer simulation of DNA double-helix dynamics. Cold Spring Harbor Symposium on Quantum Biology. 47:251–62.
13. Patel, D. J., S. Ikuta, S. Kozlowski, and K. Itakura. 1983. Sequence dependence of hydrogen exchange kinetics in DNA duplexes at the individual base pair level in solution. *Proc. Natl. Acad. Sci. USA.* 80:2184–2188.
14. Kearns, D. R. 1984. NMR studies of conformational states and dynamics of DNA. *CRC Crit. Rev. Biochem.* 15:237–290.
15. Dauxois, T., and M. Peyrard. 2006. Physics of Solitons. Cambridge University Press, Cambridge.
16. Choi, C. H., G. Kalosakas, K. Ø. Rasmussen, M. Hiromura, A. R. Bishop, and A. Usheva. 2004. DNA dynamically directs its own transcription initiation. *Nucleic Acids Res.* 32:1584–1590.
17. Kalosakas, G., K. Ø. Rasmussen, A. R. Bishop, C. H. Choi, and A. Usheva. 2004. Sequence-specific thermal fluctuations identify start sites for DNA transcription. *Europhys. Lett.* 68:127–133.
18. Prohofsky, E. W. 1988. Solitons hiding in DNA and their possible significance in RNA transcription. *Phys. Rev. A.* 38:1538–1541.
19. Usheva, A., and T. Shenk. 1996. YY1 transcriptional initiator: protein interactions and association with a DNA site containing unpaired strands. *Proc. Natl. Acad. Sci. USA.* 93:13571–13576.
20. Rapti, Z., A. Smerzi, K. Ø. Rasmussen, A. R. Bishop, C. H. Choi, and A. Usheva. 2006. Healing length and bubble formation in DNA. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 73:051902.
21. Rapti, Z., A. Smerzi, K. Ø. Rasmussen, A. R. Bishop, C. H. Choi, and A. Usheva. 2006. Lengthscales and cooperativity in DNA bubble formation. *Europhys. Lett.* 74:540–546.
22. Domberger, U., M. Leijon, and H. Fritzsche. 1999. High base pair opening rates in tracts of GC base pairs. *J. Biol. Chem.* 274:6957–6962.
23. Berk, A. J. 1986. Adenovirus promoters and E1A transactivation. *Annu. Rev. Genet.* 20:45–79.
24. Parks, C. L., S. Banerjee, and D. J. Spector. 1988. Organization of the transcriptional control region of the E1b gene of adenovirus type 5. *J. Virol.* 62:54–67.
25. Matys, V., E. Fricke, R. Geffers, E. Gössling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, and S. Land. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31:374–378.
26. Campa, A., and A. Giansanti. 1998. Experimental tests of the Peyrard-Bishop model applied to the melting of very short DNA chains. *Phys. Rev. E.* 58:3585.
27. Hardy, S., and T. Shenk. 1988. Adenoviral control regions activated by E1A and the cAMP response element bind to the same factor. *Proc. Natl. Acad. Sci. USA.* 85:4171–4175.
28. Yoshida, K., M. Narita, and K. Fujinaga. 1989. Binding sites of HeLa cell nuclear proteins on the upstream region of adenovirus type 5 E1A gene. *Nucleic Acids Res.* 17:10015–10034.
29. Koikeda, S., R. Ibuki, Y. Sawada, K. Nagata, H. Shibata, Y. Masamune, and Y. Nakanishi. 1990. Nuclear factor I stimulates transcription of the adenovirus 12 E1A gene in a cell-free system. *Biochim. Biophys. Acta.* 1048:85–92.
30. Schmidt, M. C., C. C. Kao, R. Pei, and A. J. Berk. 1989. Yeast TATA-box transcription factor gene. *Proc. Natl. Acad. Sci. USA.* 86:7785–7789.
31. Schmidt, M. C., Q. Zhou, and A. J. Berk. 1989. Sp1 activates transcription without enhancing DNA-binding activity of the TATA box factor. *Mol. Cell. Biol.* 9:3299–3307.
32. Zhou, Q., P. M. Lieberman, T. G. Boyer, and A. J. Berk. 1992. Holo-TFIID supports transcriptional stimulation by diverse activators and from a TATA-less promoter. *Genes Dev.* 6:1964–1974.
33. Wu, L., D. S. Rosser, M. C. Schmidt, and A. Berk. 1987. A TATA box implicated in E1A transcriptional activation of a simple adenovirus 2 promoter. *Nature.* 326:512–515.
34. Sogawa, K., Y. Kikuchi, H. Imataka, and Y. Fujii-Kuriyama. 1993. Comparison of DNA-binding properties between BTEB and Sp1. *J. Biochem.* 114:605–609.
35. Parks, C. L., and T. Shenk. 1996. The serotonin 1a receptor gene contains a TATA-less promoter that responds to MAZ and Sp1. *J. Biol. Chem.* 271:4417–4430.
36. SivaRaman, L., and B. Thimmappaya. 1987. Two promoter-specific host factors interact with adjacent sequences in an E1A-inducible adenovirus promoter. *Proc. Natl. Acad. Sci. USA.* 84:6112–6116.
37. Yee, A. S., R. Reichel, I. Kovcsdi, and J. R. Nevins. 1987. Promoter interaction of the E1A-inducible factor E2F and its potential role in the formation of a multi-component complex. *EMBO J.* 6:2061–2068.
38. La Thangue, N. B., B. Thimmappaya, and P. W. Rigby. 1990. The embryonal carcinoma stem cell Ela-like activity involves a differentiation-regulated transcription factor. *Nucleic Acids Res.* 18:2929–2938.
39. Jansen-Durr, P., H. Boeuf, and C. Kédinger. 1989. Cooperative binding of two E2F molecules to an Ela-responsive promoter is triggered by the adenovirus Ela, but not by a cellular Ela-like activity. *EMBO J.* 8:3365–3370.
40. Jalinot, P., B. Devaux, and C. Kédinger. 1987. The abundance and in vitro DNA binding of three cellular proteins interacting with the adenovirus E1a early promoter are not modified by the Ela gene products. *Mol. Cell. Biol.* 7:3806–3817.
41. Garcia, J., F. Wu, and R. Gaynor. 1987. Upstream regulatory regions required to stabilize binding to the TATA sequence in an adenovirus early promoter. *Nucleic Acids Res.* 15:8367–8385.
42. Williams, J. L., J. Garcia, D. Harrich, L. Pearson, F. Wu, and R. Gaynor. 1990. Lymphoid specific gene expression of the adenovirus early region 3 promoter is mediated by NF-kappa B binding motifs. *EMBO J.* 9:4435–4442.
43. Watanabe, H., T. Imai, P. A. Sharp, and H. Handa. 1988. Identification of two transcription factors that bind to specific elements in the promoter of the adenovirus early-region 4. *Mol. Cell. Biol.* 8:1290–1300.
44. Rooney, R. J., P. Raychaudhuri, and J. R. Nevins. 1990. E4F and ATF, two transcription factors that recognize the same site, can be distinguished both physically and functionally: a role for E4F in E1A trans activation. *Mol. Cell. Biol.* 10:5138–5149.
45. Watanabe, H., J. Sawada, K. Yano, K. Yamaguchi, M. Goto, and H. Handa. 1993. cDNA cloning of transcription factor E4TF1 subunits with Ets and notch motifs. *Mol. Cell. Biol.* 13:1385–1391.
46. Horikoshi, M., T. Hai, Y. S. Lin, M. R. Green, and R. G. Roeder. 1988. Transcription factor ATF interacts with the TATA factor to facilitate establishment of a preinitiation complex. *Cell.* 54:1033–1042.
47. Raychaudhuri, P., R. Rooney, and J. R. Nevins. 1987. Identification of an E1A-inducible cellular factor that interacts with regulatory sequences within the adenovirus E4 promoter. *EMBO J.* 6:4073–4081.
48. Merino, A., L. Buckbinder, F. H. Mermelstein, and D. Reinberg. 1989. Phosphorylation of cellular proteins regulates their binding to the cAMP response element. *J. Biol. Chem.* 264:21266–21276.
49. Goding, C. R., S. M. Temperley, and F. Fisher. 1987. Multiple transcription factors interact with the adenovirus-2 E1A-late promoter: evidence for a novel CCAAT recognition factor. *Nucleic Acids Res.* 15:7761–7780.
50. Vales, L. D., and J. E. Darnell. 1989. Promoter occlusion prevents transcription of adenovirus polypeptide IX mRNA until after DNA replication. *Genes Dev.* 3:49–59.

51. Babiss, L. E., and L. D. Vales. 1991. Promoter of the adenovirus polypeptide IX gene: similarity to E1B and inactivation by substitution of the simian virus 40 TATA element. *J. Virol.* 65:598–605.
52. Engler, J. A. 1981. The nucleotide sequence of the polypeptide IX gene of human adenovirus type 3. *Gene.* 13:387–394.
53. Matsui, T. 1989. Novel regulation of transcription initiation of the peptide IX gene of adenovirus 2. *Mol. Cell. Biol.* 9:4265–4271.
54. Sawadogo, M., and R. G. Roeder. 1985. Interaction of a gene-specific transcription factor with the adenovirus major late promoter upstream of the TATA box region. *Cell.* 43:165–175.
55. Carcamo, J., L. Buckbinder, and D. Reinberg. 1991. The initiator directs the assembly of a transcription factor IID-dependent transcription complex. *Proc. Natl. Acad. Sci. USA.* 88:8052–8056.
56. Nakajima, N., M. Horikoshi, and R. G. Roeder. 1988. Factors involved in specific transcription by mammalian RNA polymerase II: purification, genetic specificity, and TATA box-promoter interactions of TFIID. *Mol. Cell. Biol.* 8:4028–4040.
57. Garfinkel, S., J. A. Thompson, W. F. Jacob, R. Cohen, and B. Safer. 1990. Identification and characterization of an adenovirus 2 major late promoter CAP sequence DNA-binding protein. *J. Biol. Chem.* 265:10309–10319.
58. Pognonec, P., and R. G. Roeder. 1991. Recombinant 43-kDa USF binds to DNA and activates transcription in a manner indistinguishable from that of natural 43/44-kDa USF. *Mol. Cell. Biol.* 11:5125–5136.
59. Chodosh, L. A., A. S. Baldwin, R. W. Carthew, and P. A. Sharp. 1988. Human CCAAT-binding proteins have heterologous subunits. *Cell.* 53:11–24.
60. Mahajan, P. B., and E. A. Thompson. 1990. Glucocorticoid inhibition of transcription from adenovirus major late promoter. *Mol. Endocrinol.* 4:1515–1521.
61. Tronche, F., A. Rollier, D. Sourdive, S. Cereghini, and M. Yaniv. 1991. NFY or a related CCAAT binding factor can be replaced by other transcriptional activators for co-operation with HNF1 in driving the rat albumin promoter in vivo. *J. Mol. Biol.* 222:31–43.
62. Moncollin, V., R. Stalder, J. M. Verdier, A. Sentenac, and J. M. Egly. 1990. A yeast homolog of the human UEF stimulates transcription from the adenovirus 2 major late promoter in yeast and in mammalian cell-free systems. *Nucleic Acids Res.* 18:4817–4823.
63. Albrecht, G., B. Devaux, and C. Kedinger. 1988. Genomic footprinting detects factors bound to major late and IVa2 promoters in adenovirus-infected HeLa cells. *Mol. Cell. Biol.* 8:1534–1539.
64. Carcamo, J., E. Maldonado, P. Cortes, M. H. Ahn, I. Ha, Y. Kasai, J. Flint, and D. Reinberg. 1990. A TATA-like sequence located downstream of the transcription initiation site is required for expression of an RNA polymerase II transcribed gene. *Genes Dev.* 4:1611–1622.
65. Paillard, G., and R. Lavery. 2004. Analyzing protein-DNA recognition mechanisms. *Structure.* 12:113–122.
66. Florquin, K., Y. Saeys, S. Degroove, P. Rouzé, and Y. Van de Peer. 2005. Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Res.* 33:4255–4264.
67. Gromiha, M. M. 2005. Influence of DNA stiffness in protein-DNA recognition. *J. Biotechnol.* 117:137–145.
68. Olson, V. A., J. A. Wetter, and P. D. Friesen. 2003. The highly conserved basic domain I of baculovirus IE1 is required for hr enhancer DNA binding and hr-dependent transactivation. *J. Virol.* 77:5668–5677.
69. Olson, W. K., A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA.* 95:11163–11168.
70. Kim, J. L., and S. K. Burley. 1994. 1.9 Å resolution refined structure of TBP recognizing the minor groove of TATAAAAG. *Nat. Struct. Biol.* 1:638–653.
71. Breslauer, K. J., R. Frank, H. Blöcker, and L. A. Marky. 1986. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA.* 83:3746–3750.