Sequence-specific thermal fluctuations identify start sites for DNA transcription

EUROPHYSICS LETTERS 1 October 2004

Europhys. Lett., **68** (1), pp. 127–133 (2004)
DOI: 10.1209/epl/i2004-10167-8

Sequence-specific thermal fluctuations identify start sites for DNA transcription

G. Kalosakas[1](*), K. Ø. Rasmussen[1], A. R. Bishop[1], C. H. Choi[2] and A. Usheva[2]

[1] *Theoretical Division and Center for Nonlinear Studies Los Alamos National Laboratory - Los Alamos, NM 87545, USA*
[2] *Endocrinology, Beth Israel Deaconess Medical Center Department of Medicine, Harvard Medical School 99 Brooklin Ave., Boston, MA 02215, USA*

received 12 May 2004; accepted in final form 28 July 2004

PACS. 87.15.Aa – Theory and modeling; computer simulation.
PACS. 87.15.Ya – Fluctuations.
PACS. 87.15.He – Dynamics and conformational changes.

Abstract. – We report successful comparisons between model predictions for intrinsic thermal openings and experimental transcription data, showing that large and slow thermally induced openings (bubbles) of double-stranded DNA coincide with the location of start sites for transcription. Investigating viral and bacteriophage DNA gene promoter segments, we find that the largest opening occurs at the transcription start site in all cases studied. Other probable large openings predicted in our model appear to be related to other regulatory sites. The coherent dynamics is determined by a combination of sequence specificity (disorder), nonlinearity, and entropy, controlled by the long-range consequences of local base-pair stacking constraints.

One of the most challenging subjects in biophysics is the relation between biomolecular motions and function [1]. Recently, nuclear magnetic-resonance relaxation measurements were able to observe at atomic resolution dynamics of proteins in action [2, 3]. These experiments revealed the presence of dynamical hot spots and collective conformational fluctuations at the active domains of the studied proteins, that are explicitly linked to function.

Here, corroborating this picture, we present examples supporting functionality arising from structurally coherent dynamics, controlled by an essential combination of: *sequence specificity, nonlinearity and entropy*; the nonlinearity from local base-pair stacking constraints is crucial. Successful comparisons of molecular-dynamics simulations of a minimal model (below) of transverse dynamics for gene promoter DNA segments with *in vitro* transcriptional experiments, show that the combination of all the above-mentioned components together control coherent "bubble" fluctuational openings of base-pairs around specific sites of promoter DNA. In all cases studied, the largest bubble openings occur at the transcription start site, while other large openings coincide with regulatory sites at which transcription factors and other assisting proteins are bound. These results demonstrate the importance of the sequence structure, not simply as a static and passive element, but rather to provide the template for specific coherent fluctuations determining function features. This is an example of the importance of a (dynamic) landscape of substates [1].

(*) Present address: Max Planck Institute for the Physics of Complex Systems - Nöthnitzer Str. 38, Dresden 01187, Germany.

We have used a microscopic model proposed by Peyrard, Bishop, and Dauxois [4, 5] to describe the dynamics of the openings of double-stranded DNA. This model focuses only on the most obvious degrees of freedom for openings, namely the transverse stretching of the hydrogen bonds connecting complementary bases in the opposite strands of the double helix. Its reduced character makes it suitable for simulations over relatively long times and appropriate for gathering sufficient statistics. The potential energy of this model reads [5]

$$V = \sum_n \left[D_n \left(e^{-a_n y_n} - 1 \right)^2 + \frac{k}{2} \left(1 + \rho \, e^{-\beta (y_n + y_{n-1})} \right) \left(y_n - y_{n-1} \right)^2 \right]. \tag{1}$$

Here the sum is over all the base-pairs of the DNA and y_n denotes the displacement from the equilibrium position of the relative distance between the bases within the n-th base-pair, divided by $\sqrt{2}$. The Morse potential (other similar potentials can also be used) in the first term provides the effective interactions between complementary bases; it represents both the attraction due to the hydrogen bonds forming the base-pairs and the repulsion of the negatively charged phosphates in the backbone of the two strands screened by the surrounding solvent. The parameters D_n and a_n of the on-site potential distinguish between the two possible combinations of bases, *i.e.* adenine-thymine (A-T) or guanine-cytosine (G-C), at site n, depending on the particular sequence. The second term in the total potential energy (1) represents the stacking interaction between adjacent base-pairs. Here the nonlinear inter-site coupling, given by the exponential term that effectively modifies a harmonic spring constant, is essential for representing *local constraints* in nucleotide motions, which result in long-range cooperative effects [5]. As in elastic materials [6, 7], such constraints control lattice vibrations, yielding essential entropic terms [8]: the stiffening of the coupling in the compact state compared to that in the open state leads to an abrupt entropy-driven transition [5]. Physically, the constraint describes the change of the next-neighbor stacking interaction due to the distortion of the hydrogen bonds connecting a base-pair, mediated by the redistribution of the electrons on the corresponding bases.

The apparent simplicity of this model is deceptive. As noted earlier, it combines essential ingredients: The stacking constraint induces nonlinearity and entropic effects resulting in coherent bubble openings in a precursor thermal window preceding melting; the disorder from the particular sequence selects specific locations for large bubbles. This model has successfully reproduced not only the sharp melting (denaturation) transition occurring when the two strands of a long DNA molecule separate, but also the details of the precursor (nucleation) fluctuational openings and the dynamics upon approaching the denaturation transition. The coexistence of two essential features is necessary for obtaining the first-order transition [9]: i) the nonlinear coupling constant that decreases in the denaturated phase providing an increase in entropy and ii) a plateau in the on-site potential which allows the exploration of a large domain in phase space, with little energy cost. Regarding the precursor fluctuations, intrinsic localized modes nucleate as nonlinear bubble opening events that subsequently interact and grow, providing the experimentally observed denaturation bubbles [4, 10, 11]. This nucleation regime, precursor to the melting, extends over temperatures several tens of kelvins below the melting transition, *i.e.* the biologically relevant regime.

The model has also been used to reproduce the melting curves of very short heterogeneous DNA segments, in excellent quantitative agreement with experimental data [12]. Furthermore, it provides the characteristic multi-step melting observed in single heterogeneous DNA molecules [13]. Recently, the model has been used to describe charge transport properties in a flexible DNA chain, where the charge is coupled to the lattice degrees of freedom [14, 15]. The bubbles, as long-lived intrinsic inhomogeneities ("hot spots") [16], represent a colored noise

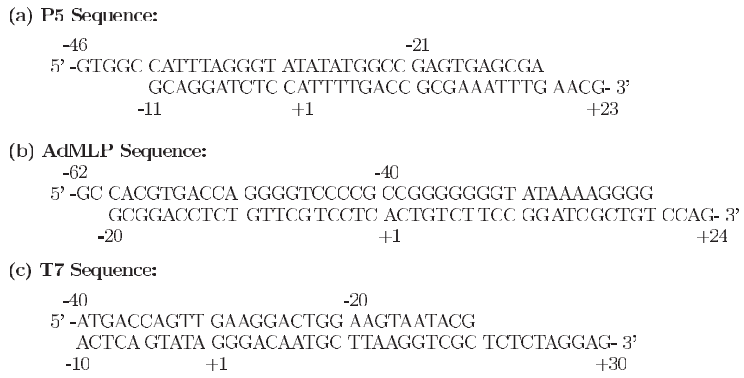

Fig. 1 – Base-pair sequences of the studied DNA gene promoter fragments. (a) 69 base-pair long viral P5 promoter, (b) 86 base-pair long viral AdMLP promoter, and (c) 70 base-pair long bacteriophage T7 promoter.

environment, which determines charge dynamics [14, 17].

Motivated by the successful descriptions of the nonlinear thermal fluctuations, we have applied this model to explore the possible role of the intrinsic bubble openings for the transcriptional initiation and regulatory sites of specific promoter DNA sequences for which we have detailed experimental data. In particular, we have studied fragments of the adenoassociated viral P5 promoter (P5) [18], the adenovirus major late promoter (AdMLP) and the bacteriophage T7 core promoter (T7). The base-pair sequences of these promoters are presented in fig. 1. *In vitro* transcription experiments, identifying the specific initiation of RNA polymerase II transcription from DNA templates containing the corresponding promoter fragments, are shown in fig. 2. See ref. [19] for more details regarding the transcription experiments.

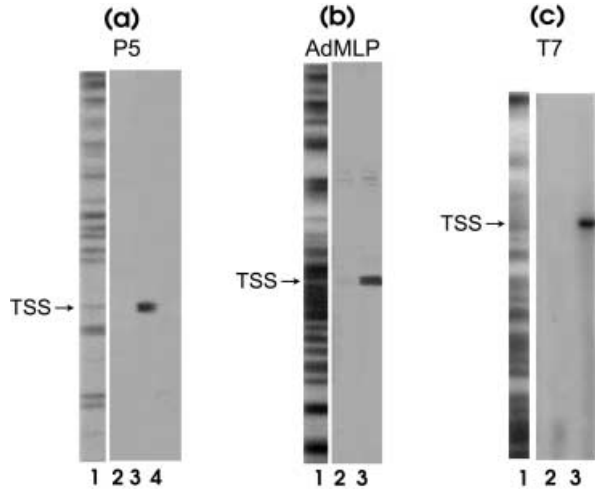

Fig. 2 – Auto-radiography of [^{32}P]-labeled reverse transcripts after separation by gel electrophoresis (lanes 3). Specific transcription started from the site +1, denoted as TSS (transcription start site), in all cases. Lanes 1 indicate base position markers obtained by chemical sequencing. (a) P5 promoter, (b) AdMLP promoter, and (c) T7 promoter. Lane 4 in (a) shows elimination of the transcription for the mutated P5 promoter, where the nucleotides at +1 and +2 have been changed from AT to GC.

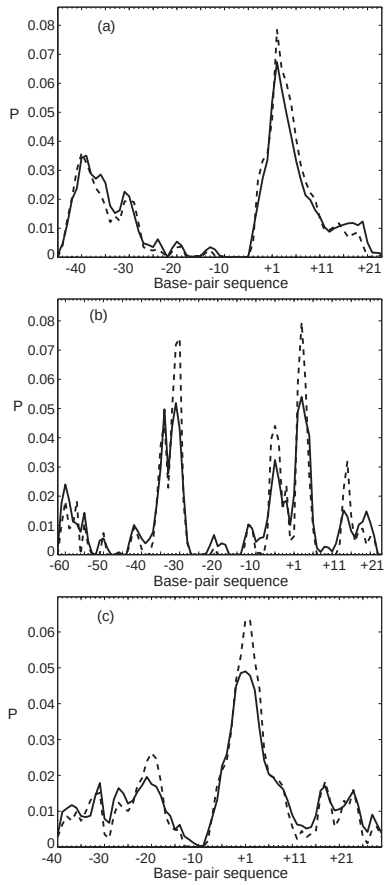

Fig. 3 – Probability P for the occurrence of an opening of 10 base-pair width and amplitude of more than 2.1 Å (thick solid line) or 1.4 Å (dotted line) starting at a particular site n of the DNA fragment, as a function of n, for the (a) P5 promoter, (b) AdMLP promoter, and (c) T7 promoter. All the probabilities have been normalized to unity along the DNA segment.

We performed Langevin molecular dynamics (thereby modeling thermal fluctuations and dissipation effects) for nucleotides of mass m evolving in the potential V of eq. (1). We used the parameter values given in ref. [12]: $k = 0.025\,\mathrm{eV/\AA^2}$, $\rho = 2$, $\beta = 0.35\,\mathrm{\AA^{-1}}$ for the inter-site coupling, while for the Morse potential $D_{\mathrm{GC}} = 0.075\,\mathrm{eV}$, $a_{\mathrm{GC}} = 6.9\,\mathrm{\AA^{-1}}$ for a GC base-pair, and $D_{\mathrm{AT}} = 0.05\,\mathrm{eV}$, $a_{\mathrm{AT}} = 4.2\,\mathrm{\AA^{-1}}$ for an AT pair. These parameters accurately reproduce measured melting curves in oligonucleotides [12]. The simulated temperature was 300 K (below the melting temperature but in the precursor regime of bubble formation). The other parameters are $m = 300$ amu and the dissipation coefficient $\gamma = 0.005\,\mathrm{ps^{-1}}$ (the friction term is $-m\gamma\frac{\mathrm{d}y}{\mathrm{d}t}$). The preheating time allowing the system to reach thermal equilibrium at each realization was 92 ps and during this period the dissipation was $0.05\,\mathrm{ps^{-1}}$. At the end of the preheating period the mean kinetic energy of the system is fluctuating around the corresponding thermal equilibrium value. We have used periodic boundary conditions in the numerical calculations in order to eliminate artificial opening propensity at the boundaries.

The statistics of the thermally induced openings was obtained using 1000 different realizations for each DNA sequence studied. We ran each realization for 1 ns, after reaching thermal

equilibrium, and monitored the state of the system every 1 fs. Thus we have 10^6 events for each of the 1000 realizations. At every event we checked the displacements of the base-pairs at each site n and the following $l-1$ (l varying from 1 to 20) base-pairs. If the openings at *all* these l subsequent sites are greater than a threshold value $y_{\rm th}$ (varying from one tenth to a few Å) we assign a contribution to the opening event at the n-th base-pair of the sequence. The obtained opening probabilities along the studied DNA segments for large bubble sizes of $l = 10$ base-pairs and thresholds $y_{\rm th} = 1.0$ and 1.5 Å for accepting an opening are presented in fig. 3. (Recall that the real openings are equal to $y\sqrt{2}$.)

Remarkably, in all these cases the most probable large openings are located at the experimental transcription start site TSS (at +1). Furthermore, in the viral cases the other distinct openings seem to be related to known regulatory and transcription factor binding sites: in P5, the opening at the A/T-rich region between −40 and −35 corresponds to the binding site of the transcription factor Yin Yang 1 [20], while in AdMLP the second largest opening is close to the binding site of the TATA-box binding protein [21] that is necessary for transcription. Openings of such large widths and amplitudes are rare events in our microscopic simulations, therefore requiring sufficient statistics. Note also that the transcription start site is located at the largest bubble which is also the slowest: initiation is thus favored by both the coherent deformation and its dynamics.

The locations of the large openings predicted by our model agree very well with experimental measurements, where S1 nuclease selectively cleaves single-stranded DNA [19]. These results verify the biological relevance of the numerically calculated openings.

We stress that similar local sequences do *not* exhibit the same opening probabilities; equal size segments of relatively weakly bound A/T pairs in different parts of the promoter show very different statistics (compare, for example, the region between −30 and −25 with that around +1 in P5). Furthermore, the larger openings do not necessarily occur in regions with longer A/T stretches, as might be intuitively anticipated because of the weaker bonding. Effective long-range cooperativity (resulting from the nonlinear inter-site potential in eq. (1)) and competing localization lengths due to the disorder (base-pair sequence) and the nonlinearity are responsible for this striking specificity: in general, length scale competition in nonlinear systems is known [22] to lead to complex spatio-temporal (dynamic landscape) behavior. The sensitivity of the cooperative/competing phenomena enhances the predictive power of our model; for instance, a small mutation of the sequence (at a specific location) is sufficient to completely eliminate bubble formation at the transcription initiation site, as found experimentally. For instance, in fig. 4 we show numerical calculations of the opening probabilities for a mutated P5 promoter, where nucleotides +1 and +2 have been changed from AT to GC. This mutation completely eliminates the opening at the previous transcription start site, in agreement with the absence of transcriptional events in the corresponding experiment (see fig. 2a, lane 4). We mention that the enhanced opening probability in the region from −40 to −30 in this case, compared to the wild type P5 (fig. 3a), is only partly due to the normalization used along the DNA chain. Nonlocal effects are also present due to anharmonic stacking interactions, and they have been confirmed experimentally [19].

We emphasize that, as in previous applications of our model, the nonlinear inter-site coupling is crucial for its success [5, 13]. For example, as can be seen in fig. 5, linearizing the stacking interaction term ($\rho = 0$) results in totally modified statistics for the openings of the P5 promoter, changes the position of the peaks along the sequence, and totally eliminates the successful comparison with the experimentally observed transcription. The nonlinear inter-site coupling constitutes a minimal representation of the local stacking constraint between neighboring base-pairs. As in more general situations of displacive structural phase transitions [6, 7], such local constraints can result in long-range "elastic" interactions and

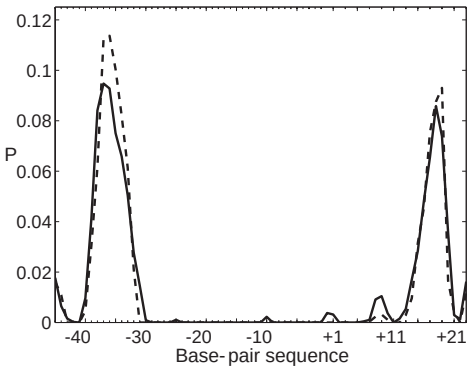

Fig. 4

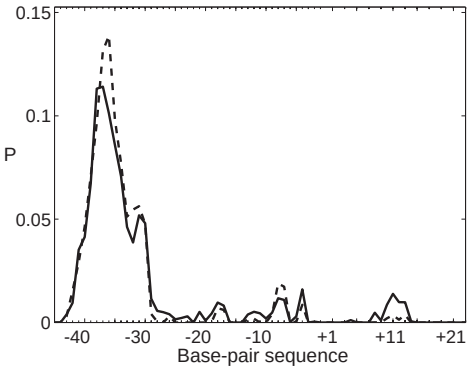

Fig. 5

Fig. 4 – Probability P for the occurrence of an opening of 10 base-pair width and amplitude of more than 2.1 Å (thick solid line) or 1.4 Å (dotted line) starting at a particular site n of the DNA fragment, as a function of n for the mutated P5 promoter (see text). The probabilities have been normalized to unity along the DNA segment.

Fig. 5 – Probability P for the occurrence of an opening of 10 base-pair width and amplitude of more than 2.1 Å (thick solid line) or 1.4 Å (dotted line) starting at a particular site n of the DNA fragment, as a function of n for the P5 promoter, by linearizing the stacking interaction term of the potential (*i.e.* setting $\rho = 0$ in the potential energy (1)). The probabilities have been normalized to unity along the DNA segment.

macroscopic cooperativity. This leads to entropy being critical in our approach —specifically in the DNA dynamics itself.

Thermal stability of a DNA double helix has also been studied by Yeramian, seeking a correlation between opening probabilities and genetic maps of a coarser length scale level [23]. He finds a distinct relation between genetic maps and intrinsic structural stability properties in some cases (probably relics of an archaic functional organization), where noncoding regions of DNA display larger propensity for thermal disruption of the double helix compared to coding regions. The helix-coil transition model used in this coarse-grained analysis (at the level of complete genomes or chromosomes) is not a dynamical, but rather a statistical model. The model used in the present report provides information on a finer scale, more relevant for dynamical procedures like transcription, and where the important role of nonlinearity for biological function is amplified.

In summary, our model and simulations, and the successful comparisons with transcriptional assays, indicate that structurally specific, large, slow, coherent thermal fluctuations occur at biologically relevant temperatures and at locations in the DNA sequences where the RNA polymerase initiates transcription. Further, we find indications that large thermal fluctuations may also help in recruiting other protein complexes participating in the transcriptional process, by separating the DNA double strand at specific locations. These bubbles are intrinsic properties of the double helix that precede protein binding and, therefore, a possible biological role is limited to the very initial steps of the transcription. Furthermore, studies of more DNA sequences and geometries are clearly called for —we can expect, for example, that local torsions will readily accompany the softest regions at the largest bubbles. However, our results already suggest that DNA, through structurally specific intrinsic dynamics, may participate in directing its own transcription. They also raise the exciting prospect for reverse-engineering of sequences for desired functions.

* * *

We would like to thank H. FRAUENFELDER, P. FENIMORE and J. A. KRUMHANSL for valuable discussions. This research was supported by the US DoE, under contract *W*-7405-*ENG*-36 and the NIH.

REFERENCES

[1] FRAUENFELDER H., SLIGAR S. G. and WOLYNES P. G., *Science*, **254** (1991) 1598.
[2] VOLKMAN B. F., LIPSON D., WEMMER D. E. and KERN D., *Science*, **291** (2001) 2429.
[3] EISENMESSER E. Z., BOSCO D. A., AKKE M. and KERN D., *Science*, **295** (2002) 1520.
[4] PEYRARD M. and BISHOP A. R., *Phys. Rev. Lett.*, **62** (1989) 2755.
[5] DAUXOIS T., PEYRARD M. and BISHOP A. R., *Phys. Rev. E*, **47** (1993) R44.
[6] MORRIS J. R. and GOODING R. J., *Phys. Rev. Lett.*, **65** (1990) 1769; *Phys. Rev. B*, **43** (1991) 6057.
[7] KERR W. C., HAWTHORNE A. M., GOODING R. J., BISHOP A. R. and KRUMHANSL J. A., *Phys. Rev. B*, **45** (1992) 7036.
[8] The importance of entropic terms for enhancing cooperativity and providing distinct transitions, similarly to the case of the model we used, has also been stressed recently in a different context, *viz.* β-hairpin folding models. See GUO C., LEVINE H. and KESSLER D. A., *Phys. Rev. Lett.*, **84** (2000) 3490.
[9] DAUXOIS T. and PEYRARD M., *Phys. Rev. E*, **51** (1995) 4027.
[10] ZENG Y., MONTRICHOK A. and ZOCCHI G., *Phys. Rev. Lett.*, **91** (2003) 148101.
[11] We note that these large-amplitude bubbles (breathers) are not traveling, in contrast to mobile low-amplitude nonlinear excitations proposed to account for hydrogen exchange experiments in, *e.g.*, ENGLANDER S. W., KALLENBACH N. R., HEEGER A. J., KRUMHANSL J. A. and LITWIN S., *Proc. Natl. Acad. Sci. USA*, **77** (1980) 7222.
[12] CAMPA A. and GIANSANTI A., *Phys. Rev. E*, **58** (1998) 3585.
[13] CULE D. and HWA T., *Phys. Rev. Lett.*, **79** (1997) 2375.
[14] KOMINEAS S., KALOSAKAS G. and BISHOP A. R., *Phys. Rev. E*, **65** (2002) 061905.
[15] MANIADIS P., KALOSAKAS G., RASMUSSEN K. Ø. and BISHOP A. R., *Phys. Rev. B*, **68** (2003) 174304.
[16] PEYRARD M. and FARAGO J., *Physica A*, **288** (2000) 199.
[17] KALOSAKAS G., RASMUSSEN K. Ø. and BISHOP A. R., *J. Chem. Phys.*, **118** (2003) 3731; *Synth. Met.*, **141** (2004) 93.
[18] The adeno-associated viral P5 has a single-stranded DNA. However, once the virus infects a host, the protective coating is shed and this DNA is replicated, resulting in transcriptionally active double-stranded DNA. The promoter is in its active form once it has become double-stranded. See, for example, STRYER L., *Biochemistry*, 3rd edition (W. H. Freeman) 1988; FIELDS B. N., KNIPE D. M. and HOWLEY P. M. (Editors), *Fields Virology*, Vol. **3** (Lippincott-Raven) 1990.
[19] CHOI C. H., KALOSAKAS G., RASMUSSEN K. Ø., HIROMURA M., BISHOP A. R. and USHEVA A., *Nucleic Acids Res.*, **32** (2004) 1584.
[20] USHEVA A. and SHENK T., *Cell*, **76** (1994) 1115.
[21] CONCINO M. F., LEE R. F., MERRYWEATHER J. P. and WEINMANN R., *Nucleic Acids Res.*, **12** (1984) 7423.
[22] SANCHEZ A. and BISHOP A. R., *SIAM Rev.*, **40** (1998) 579.
[23] YERAMIAN E., *Gene*, **255** (2000) 139; YERAMIAN E. and JONES L., *Nucleic Acids Res.*, **31** (2003) 3843.