Predicting the function of eukaryotic scaffold/matrix attachment regions via DNA mechanics

2005 J. Phys.: Condens. Matter 17 S2853

(http://iopscience.iop.org/0953-8984/17/31/011)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **17** (2005) S2853–S2860 doi:10.1088/0953-8984/17/31/011

Predicting the function of eukaryotic scaffold/matrix attachment regions via DNA mechanics

Ming Li[1] **and Zhong-can Ou-Yang**

Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100080, People's Republic of China

E-mail: lmmonk@itp.ac.cn (M Li)

Received 10 June 2005
Published 22 July 2005
Online at stacks.iop.org/JPhysCM/17/S2853

Abstract

Eukaryotic chromatin undergoes a few steps of compaction to form a chromosome. Among the several levels of chromatin assembly, successive chromatin loops (5–100 kb) demarcated by the nuclear matrix are of primary importance since they behave as topologically independent domains for coordinate regulation of harboured genes. The bases of the loops are called SMARs (scaffold/matrix attachment regions). On one hand, this chromosome structure imposes stress on the DNA molecule since the double-stranded DNA is actually unwound and bent by histones and other proteins. On the other hand, the stressed DNA itself affects chromatin assembly inversely, e.g., the unwinding stress may promote SMAR binding to the matrix. The interplay between chromosome structure and unwinding stress contributes significantly to eukaryotic gene regulation. In this paper, we investigate two issues: how torsional stress may promote SMAR anchorage to the matrix; and how the formation of chromatin loops may affect basic biochemical processes. We employ the Benham model for these purposes. Our analysis gives theoretical evidence that at least some SMARs are unwound under torsional stress and at the same time could serve as topological barriers for retaining the torsional stress on the chromatin loop which may be necessary for gene transcription.

1. Introduction

How eukaryotic genes are regulated at the right place and right time has long been the central issue in biology. Since Jacob and Monod proposed the first Lac-operator model for the prokaryote *E. coli* in the light of cybernetics, numerous works in this framework have been stimulated which aimed at deducing logic relations among functionally related genes, e.g., genetic circuits or regulatory networks. It is taken for granted that the underlying biochemical

[1] Author to whom any correspondence should be addressed.

reactions are simple information-processing processes irrelevant to the structural complexity of the cell. This functionalism approach encounters difficulty in the study of prokaryotic genes due to the structured DNA organization (nucleoid), and hardly works at all in much more complex cases of eukaryotes. In fact, very few logic circuits for eukaryotes have been elucidated in this way (for a review, see [1]). On the contrary, functionally irrelevant genes are often found co-regulated apparently because they locate in the same chromosomal context, which gives rise to artefacts in regulatory network reconstruction. It is now widely accepted that there is an essential difference in spatio-temporal gene expression pattern between prokaryotes and eukaryotes due to the architectural complexity of the nucleus. In particular, eukaryotic chromatin organization plays a very important, if not leading, role in orchestrating gene regulation.

Actually, it has been known for a long time that eukaryotic DNA undergoes several steps of packaging: the double-strand DNA (dsDNA) wraps onto histone octamers to form nucleosomes, then the nucleosome string folds into a solenoid, further the solenoid is demarcated into chromatin loops by the nuclear matrix [2] and finally the whole entity assembles into chromosome. This compact structure is first of all a mechanical facility controlling gene regulation since a large amount of mechanical force is imposed on the DNA of each chromatin loop (the torsional force is the most intriguing as will be pointed out below). The bases of the loops, SMARs (scaffold/matrix attachment regions), can be identified by biochemical methods. Two classes of SMARs have been found: the constitutive (strong) ones which bind to the matrix permanently; and facultative (weak) ones which associate with the matrix temporarily. In addition, experiments demonstrated that some SMARs can function as boundary elements to delimit gene clustering, which suggests that chromatin loops may behave as both structure units and functional units [3–6]. Since SMARs may be very important, it becomes necessary to investigate two related issues, i.e., the distribution or location of SMARs on DNA and their functions. For the first goal, sequence analysis (e.g. motif identification) can be used. This approach, however, shows limitations due to the difficulty of motif definition, and it is also impossible to explain the function of SMARs by sequence analysis alone. On the contrary, SMARs show structural similarity rather than sequence similarity since they are observed as duplex unwound regions (local denatured bubbles) [7]. Hence prediction of sequence-dependent local structures could offer a new approach complementary to motif searching, and it is also natural to relate the structural properties to functions. Considering the fact that the right-handed dsDNA is significantly unwound when coiling around histone octamers in a left-handed way and thus a large amount of unwinding stress is generated on the DNA [8], the stress should be responsible for the inducing of bubble regions which may be considered as SMAR candidates. Thus mechanical analysis is a good start for our goals. In particular, it makes it possible to access the SMAR function, as will be demonstrated in this article: how SMARs may interplay with unwinding stress in gene regulation. In fact, there is accumulated evidence that unwinding stress can induce local bubbles which play important roles in gene activation [9, 10]. In the following, we will first give a brief review of the Benham model which couples the torsional state and local denaturation of dsDNA in probably the simplest way; then we will make a tentative investigation of the two issues with this model.

2. Basics of DNA mechanics

There are several parameters for describing the torsional state of dsDNA. The linking number Lk is roughly the number of helical turns. $Lk_0 = N/\gamma$ is the linking number designated for relaxed DNA (uniform B-DNA), N is the total number of base pairs in the DNA and γ is the number of base pairs per helical turn in B-DNA (often 10.5 bp/turn). Superhelical density

σ is introduced as $\sigma = (Lk - Lk_0)/Lk_0$ to measure the degree to which dsDNA is twisted. $\Delta Lk = Lk - Lk_0$ is the linking difference. It is very surprising that σ is limited to being within a very narrow range from -0.01 to -0.1 for mesophiles (most often -0.06). In particular, the superhelical density of eukaryotic DNA is about -0.06 since a DNA segment of 167 bp is unwound one helical turn per nucleosome.

Four energy terms are included in the Benham model: twisting energy; base unpairing energy; boundary energy; and inter-strand winding energy of denatured regions. Experimental measurements and parameter estimations can be found in [11, 12].

Twisting energy corresponding to ΔLk has been carefully studied for circular dsDNA in the form $G = \frac{1}{2}K(\Delta Lk)^2$, $\Delta Lk = \sigma * Lk_0$, $K = 2200RT/N$, where R is the gas constant, T is the absolute temperature [13, 14]. Considering denaturation, ΔLk is replaced by ΔLk_r, the residual linking difference, which accounts for the part of linking difference not accommodated either by local melting or by subsequent inter-strand twisting in the melting regions. The form of ΔLk_r will be given below.

Base unpairing energy is dependent on several environmental factors, such as temperature and ion concentration. In the Benham model, the unpairing energy is taken at physiological conditions, u_{AT} is 0.26 kcal mol^{-1}, and u_{GC} is 1.30 kcal mol^{-1}. The total base unpairing energy is $G = \sum_{j=1}^{N} u_j s_j$, where s_j denotes the state of the jth base pair: $s_j = 0$, intact; $s_j = 1$, unpairing.

Boundary energy B is the energy cost of the formation of a boundary between the intact B-DNA region and the adjacent unwound region. For the Benham model, $B = 10.8$ kcal mol^{-1}. Suppose that there are r unwound regions in DNA; it is easy to show that $r = \sum_{j=1}^{N} s_j(1-s_{j+1})$, and thus the corresponding energy is $G = Br = \sum_{j=1}^{N} Bs_j(1 - s_{j+1})$.

The interwinding energy of the two strands in unwound regions originates from the high torsional flexibility of the single strand. This energy term is assumed to be a harmonic oscillator potential: $G = \frac{1}{2}Cs_j{\tau_j}^2$, where C is the torsional stiffness coefficient (about 3.6 kcal mol^{-1} rad^{-2}), τ_j (rad/bp) is the twist of the denatured jth base pair and it can be treated as an independent random variable for each base pair. The relevant total energy is $G = \sum_{j=1}^{N} \frac{1}{2}Cs_j{\tau_j}^2$.

The total change in twisting can now be expressed as $\Delta Tw = \sum_{j=1}^{N}(s_j/\gamma + s_j\tau_j/2\pi)$. ΔLk_r is determined from $\Delta Lk_r = \Delta Lk - \Delta Tw$, where ΔLk is a constant. The total free energy, putting all this together, is

$$G = \tfrac{1}{2}K\left(\Delta Lk - \sum_{j=1}^{N}(s_j/\gamma + s_j\tau_j/2\pi)\right)^2 + \sum_{j=1}^{N}\tfrac{1}{2}Cs_j{\tau_j}^2 + \sum_{j=1}^{N}((B+u_j)s_j - Bs_js_{j+1}). \quad (1)$$

We are interested in the mean value of s_j, i.e., the unpairing probability p_j of the jth base pair. Simple calculation leads to

$$p_j = \sum_{\{s\}} s_j \mathrm{e}^{-\beta G(s)}/Z \quad (2)$$

where s is the state of the sequence, $\{s\}$ is the state space (2^N states). Z is the partition function:

$$Z = \sum_{\{s\}} \mathrm{e}^{-\beta G(s)} = \sum_{n=0}^{N} \Omega(n)H(n) \quad (3)$$

$$\Omega(n) = \exp\left[\frac{-2\pi^2\beta TW}{4\pi^2T + Wn}\left(\Delta Lk + \frac{n}{\gamma}\right)^2\right]\left[\left(\frac{2\pi}{\beta T}\right)^n\left(\frac{4\pi^2T}{4\pi^2T + Wn}\right)\right]^{1/2} \quad (4)$$

$$H(n) = \sum_{\{s\}} \exp\left\{-\beta \sum_{j=1}^{N} \left[\left(B + u_j\right) s_j - B s_j s_{j+1}\right]\right\} \delta\left(\sum_j s_j - n\right). \tag{5}$$

The Benham model was originally proposed for circular DNA, e.g., plasmid or bacterial DNA, since the energy form and all the parameters are determined by experiments on circular DNA. Is it still available for linear DNA? So far as we know there is no other model with the parameters examined suitable for describing stress-induced melting behaviour, and on the other hand all the results calculated using the Benham model are insensitive to most of its parameters; therefore it can be regarded as the simplest phenomenological model suited to our purpose, at least at this stage.

For a given DNA sequence, we can calculate p_j numerically, to obtain a 'transition profile'. Distinct peaks in the profiles indicate base unpairing events (local melting) which may have significant biological meaning. In the next section, we will make several case studies with the Benham model.

3. Results and discussion

We calculate the transition profiles for several DNA sequences taken from the databank NCBI (with accession number NCBI: ...) by using the Benham model exactly. The SMARs identified are from the databank S/MARt DB (with accession number SMxxxxxxx) [15]. Sequences in all the profiles are coordinated from the 5′-terminal to the 3′-terminal (left to right).

The first case is the histone gene cluster, 5′-$\overleftarrow{\text{H3}}$-$\overrightarrow{\text{H4}}$-$\overleftarrow{\text{H2a}}$-$\overrightarrow{\text{H2b}}$-$\overleftarrow{\text{H1}}$-3′. The sequence (NCBI: X14215) is from *Drosophila melanogaster* with a SMAR between gene H1 and H3 (the position and transcriptional orientation of each gene are indicated in figure 2). The five-gene cluster repeats a hundred times on the chromosome, so we construct a sequence of two adjacent units from the original sequence (NCBI: X14215) for the first round of calculation (other constructions, e.g. sequences including more units, give the same result). The profile is shown in figure 1.

In this profile, the two identical peaks indicate two unwound regions which are in accordance with the experimentally identified SMARs (SM0000037). This result offers theoretical evidence for Bode's observation that some SMARs are prone to denaturation under unwinding stress [7]. Further, we may ask the question of why SMARs are evolved just at these sites. It has been suggested that they can serve as topological barriers for maintaining the torsional state of the loop, so we should expect new unpairing events to occur elsewhere due to the retained unwinding stress. Therefore we take out the intervening region, i.e., the single unit without the flanking SMARs, for a second round of calculation. The profile is presented in figure 2. Two new peaks emerge. The left one locates just downstream of both H4 and H2A, the right one downstream of both H2B and H1. Since H4 and H2A (or H2B and H1) are convergently transcribed, the unwinding event may be necessarily involved in transcriptional termination, and further the co-occurrence of these two events may be important to the coordinate transcription of the whole gene cluster. The two unwound regions may also be SMAR candidates though there are no relevant reports. It is possible that they are facultative SMARs and may not serve as topological barriers (we did not observe any difference between the two profiles (before and after cutting them off); data not shown). Inversely, if these two unpairing events are indeed necessary in gene transcription, our assumption of SMAR function becomes quite reasonable: that the strong SMARs could serve as topological barriers for defining the chromatin loops as both structure units and function units. It is worth emphasizing that the interplay between SMARs and stressed DNA reflects the complexity of eukaryotic gene regulation: the unwinding stress induces strong bubbles (SMARs) to form chromatin loops,

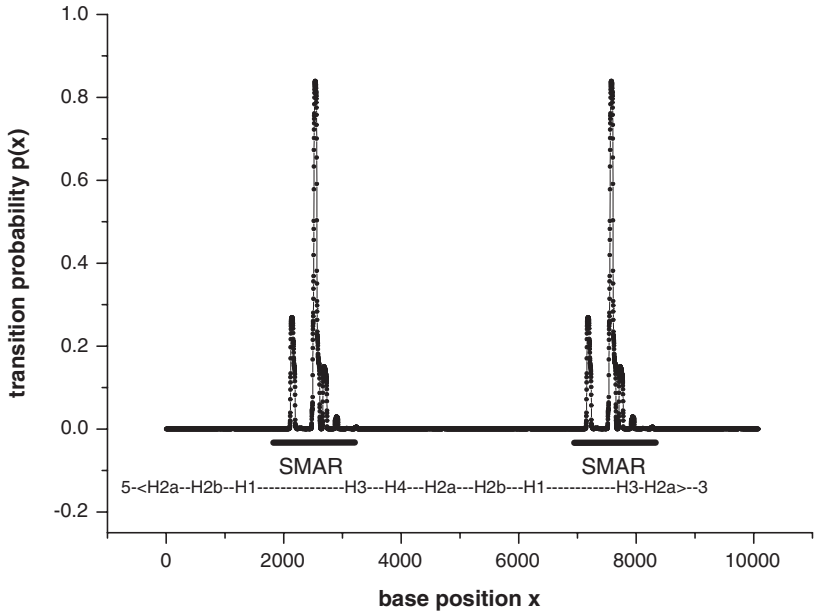

Figure 1. This sequence is two adjacent units of the histone gene cluster of *D. melanogaster* (X14215). The relative positions of these genes are indicated but not scaled. In this calculation, the superhelical density is taken as -0.06.

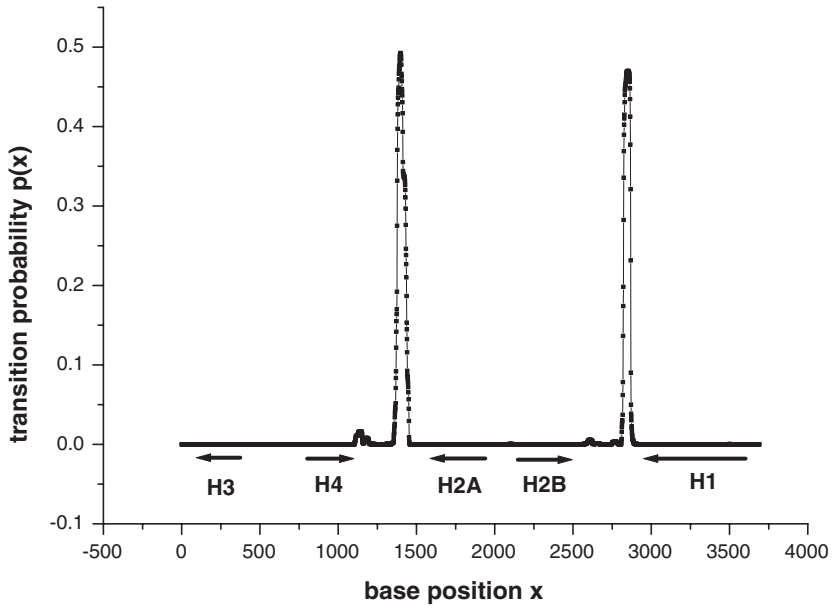

Figure 2. This sequence is a single unit without the flanking SMARs. The position and transcriptional orientation of each gene are indicated by an arrow. The superhelical density is taken as -0.06.

and inversely novel melting events induced by the conserved unwinding stress provide a new mechanism for gene regulation. The following case studies also show the complexity.

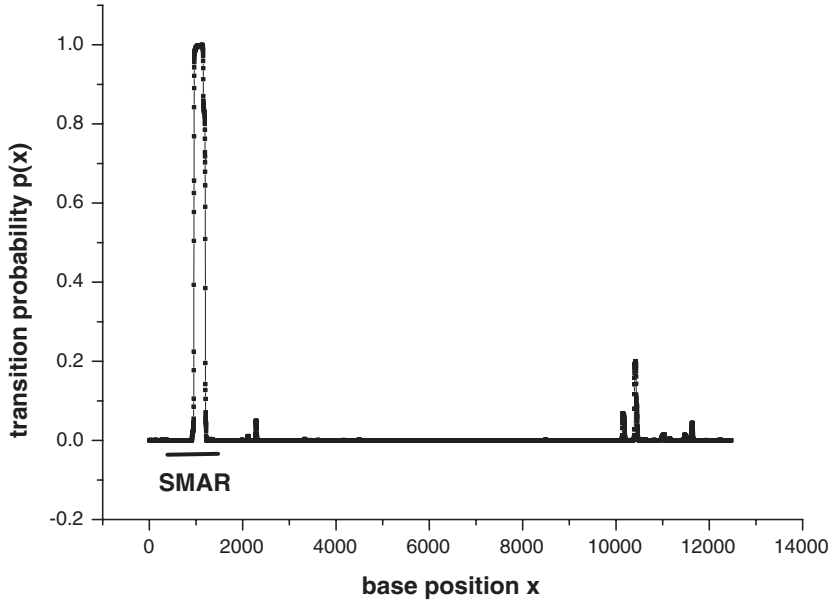

Figure 3. This sequence is a DNA region from NT086806 containing gene cspB and two flanking SMARs. The sequence coordinate is rescaled as 1:12 490. The superhelical density is taken as -0.06.

As the second example, we carry out the same analysis for another histone gene cluster from *Drosophila hydei* (NCBI:X17072). The gene order and orientation is the same as that in the above case, but the sequence differs. It is interesting that we finally obtained very similar profiles (data not shown) and again the same feature which indicates the strongest unwound region locates in the intergenic region between H1 and H3. From this we make the prediction that it is constitutive SMAR acting as domain barriers (there is no report on the SMARs for this sequence), and the same discussion of SMAR function as given above still applies in this case.

The third case is the DNA region NCBI: NT086806, REGION: 4897141..4909630, containing a single gene cspB and two flanking SMARs which are identified experimentally (5′-SMAR: SM0000074 and 3′-SMAR: SM0000075). As in the above analysis, we first take the entire sequence to calculate the transition profile (figure 3). The distinct peak at the 5′-terminal is in accordance with SM0000074 which is the most competitive SMAR. There is no signal corresponding to SM0000075. Then we cut off the 5′-SMAR region and carry out another round of calculation (figure 4). In this profile, the peak region at the 3′-terminal overlaps almost entirely with the SMAR SM0000075. We still make the same assumption as regards the functions of these two SMARs as in the first example, and take out the region just covering the gene for the third round of calculation. The profile is shown as figure 5, in which an unwound region emerges in the last intron of the gene. This region may also be a facultative SMAR candidate, which could play a regulatory role in gene transcription (there are several elucidated cases where intronic SMARs may be involved in plant gene transcription; for example, see [16]).

These results, as well as others (data not shown), indicate that at least some SMARs (strong SMARs) could serve as protectors of the unwinding stress, and it is also possible that some weaker SMARs could be very versatile in their functioning.

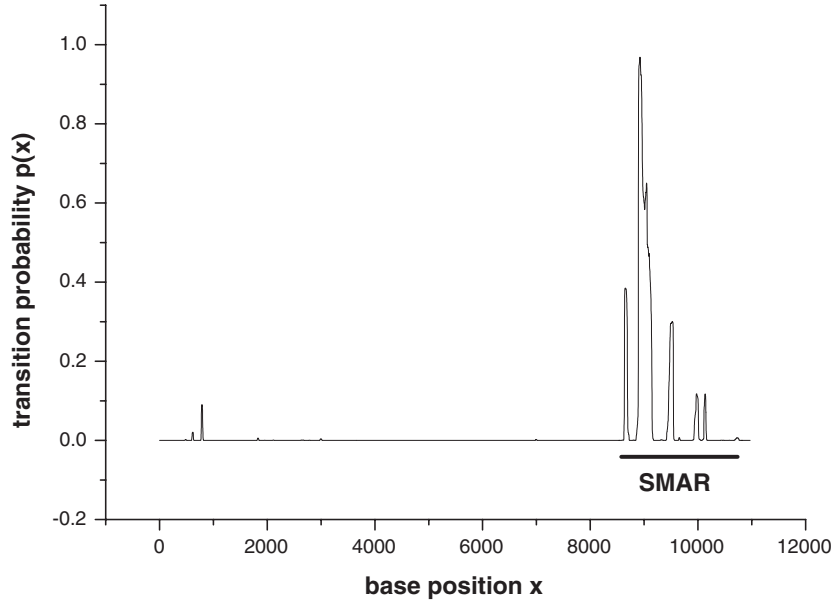

Figure 4. This sequence is the same region as in figure 3 but without the 5′ SMAR. The coordinate rescaling is 1501:12 490. The superhelical density is taken as −0.06.

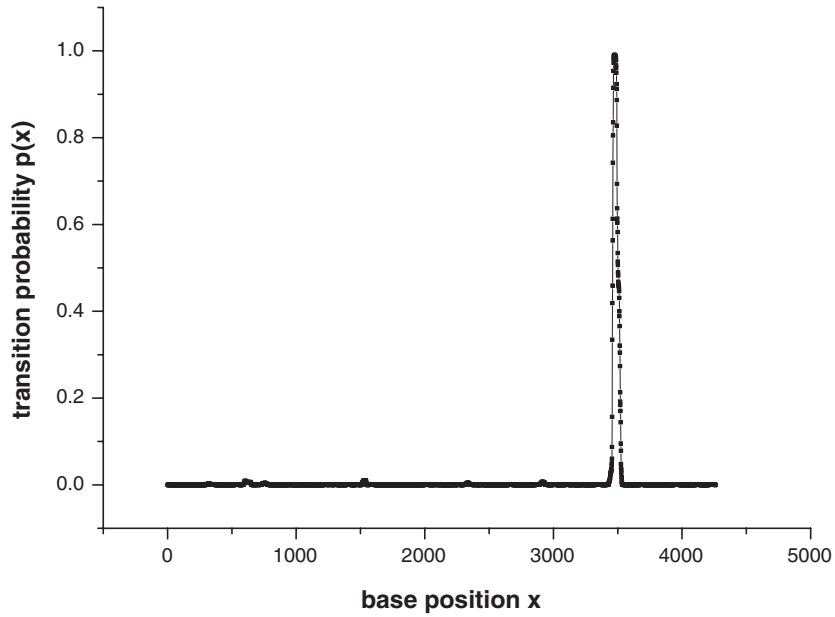

Figure 5. This sequence just covers the entire gene cspB but not the two flanking SMARs. The coordinate rescaling is 5401:9300. The superhelical density is taken as −0.06.

4. Summary

In this paper we employ the Benham model to investigate SMAR function as regards DNA mechanics. Our analysis implies that unwinding stress possibly has important functions both

in chromatin assembly (SMARs binding to the matrix) and in basic biochemical processes. It also provides preliminary support to the hypothesis that unwinding stress on a chromatin loop can be maintained by flanking SMARs functioning in gene transcription. Therefore chromatin loops are not only structure units but also function units. These qualitative conclusions still hold even when the parameters in the Benham model vary over a relatively large range, since this model is insensitive to the parameters. More case studies should be done with a more realistic model in the near future to elucidate how universal our hypothesis is.

References

[1] Freeman M 2000 *Nature* **408** 313
[2] Paulson J R and Laemmli U K 1977 *Cell* **12** 817
[3] Phi-Van L and Strätling W H 1988 *EMBO J.* **7** 655
[4] Levy-Wilson B and Fortier C 1989 *J. Biol. Chem.* **264** 21196
[5] Dillon N and Grosveld F 1993 *Trends Genet.* **9** 134
[6] Bode J, Schlake T, Rios-Ramirez M, Mielke C, Stengert M, Kay V and Klehr-Wirth D 1995 *Structural and Functional Organization of the Nuclear Matrix—International Review of Cytology* 162A ed K W Jeon and R Berezney (San Diego, CA: Academic) p 389
[7] Bode J, Kohwi Y, Dickinson L, Joh T, Klehr D, Mielke C and Kohwi-Shigematsu T 1992 *Science* **255** 195
[8] Cook P R and Brazell I A 1976 *J. Cell Sci.* **22** 287
[9] Dunaway M and Ostrander E A 1993 *Nature* **361** 746
[10] Dorman J 1997 *Mol. Gen. Genet.* **256** 93
[11] Bauer W R and Benham C J 1993 *J. Mol. Biol.* **234** 1184
[12] Bauer W R *et al* 1995 *J. Mol. Biol.* **253** 438
[13] Pulleyblank D E *et al* 1975 *Proc. Natl Acad. Sci.* **72** 4286
[14] Depew R E and Wang J C 1975 *Proc. Natl Acad. Sci.* **72** 4275
[15] Liebich I, Bode J, Frisch M and Wingender E 2002 *Nucl. Acids Res.* **30** 372
[16] Tikhonov A P, Bennetzen J L and Avramova Z V 2000 *The Plant Cell* **12** 249