3904–3911 *Nucleic Acids Research, 1997, Vol. 25, No. 19*

Analysis of the chromatin domain organisation around the plastocyanin gene reveals an MAR-specific sequence element in *Arabidopsis thaliana*

Cornelis M. van Drunen, Rob W. Oosterling[1], Gerbiënne M. Keultjes[1], Peter J. Weisbeek[1], Roel van Driel* and Sjef C. M. Smeekens[1]

E. C. Slater Institute, University of Amsterdam, Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands and [1]Department of Molecular Cell Biology, University of Utrecht, Padualaan 8, 3584 CH Utrecht, The Netherlands

Received June 12, 1997; Revised and Accepted August 19, 1997 DDBJ/EMBL/GenBank accession nos Z83320, Z83321, Z82043

ABSTRACT

The *Arabidopsis thaliana* genome is currently being sequenced, eventually leading towards the unravelling of all potential genes. We wanted to gain more insight into the way this genome might be organized at the ultrastructural level. To this extent we identified matrix attachment regions demarking potential chromatin domains, in a 16 kb region around the plastocyanin gene. The region was cloned and sequenced revealing six genes in addition to the plastocyanin gene. Using an heterologous *in vitro* nuclear matrix binding assay, to search for evolutionary conserved matrix attachment regions (MARs), we identified three such MARs. These three MARs divide the region into two small chromatin domains of 5 kb, each containing two genes. Comparison of the sequence of the three MARs revealed a degenerated 21 bp sequence that is shared between these MARs and that is not found elsewhere in the region. A similar sequence element is also present in four other MARs of *Arabidopsis.* Therefore, this sequence may constitute a landmark for the position of MARs in the genome of this plant. In a genomic sequence database of *Arabidopsis* the 21 bp element is found approxiamally once every 10 kb. The compactness of the *Arabidopsis* genome could account for the high incidence of MARs and MRSs we observed.

INTRODUCTION

Higher order chromatin structure of the eukaryote genome is thought to play an important role in regulation of transcription. According to current ideas differences in accessibility of chromatin for transcription-related proteins result in regions of the genome that are either poised for transcriptional activity or are transcriptionally silent (1,2). How the regions of distinct chromatin structure are defined is unclear. An attractive hypothesis is that these regions are correlated with the postulated organisation of the eukaryote chromosome in chromatin domains (3). In this model chromatin is bound at regular intervals to the nuclear matrix via specific genomic sequences (scaffold/matrix attachment regions, S/MARs), thereby creating domains of variable size. The average size of such a domain in mammalian cells has been shown to be ~90 kb, with actual sizes ranging from a few up to several hundred kilobases (4). The chromatin structure is assumed to be either in an 'open' or in a 'closed' conformation (5,6). In the 'closed' conformation transcription factors cannot bind to *cis*-acting regulatory sequences, such as enhancers and promoters, whereas in the 'open' conformation promoters and enhancers are accessible. In this manner the genes within a domain are thought to be coordinately regulated via changes in chromatin structure (7,8). The boundaries of these domains are defined by evolutionary conserved S/MARs. By using a heterologous *in vitro* nuclear matrix binding assay such S/MARs have been found in all eukaryotic organisms analysed so far, including yeast, plants and vertebrates (9–15). It is striking that S/MAR identified in one organism can bind to the nuclear matrix from another species; evidently these matrix–S/MAR interactions are evolutionary conserved (16). Therefore, it is likely that these sequence elements are part of a general chromatin organizing principle in the eukaryotic genome.

An important clue towards the functional role of S/MARs in defining independently controlled chromatin domains comes from studies on transgenic plants and vertebrates. Generally, the expression level of a transgene that is stably integrated into a genome is highly variable. This is believed to be due to differences in chromatin structure at the integration site and to *cis*-acting elements, like enhancers and silencers located near the integration site (17). If the transgene is flanked by S/MARs, expression is enhanced and the variability is reduced (18–22). The effect of S/MARs has been attributed to their putative ability to define an independently controlled chromatin domain and their

*To whom correspondence should be addressed. Tel: +31 20 5255150; Fax: +31 20 5255124; Email: van.driel@chem.uva.nl

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

postulated role as enhancer-blocker (23). However, it is not clear whether the physical interaction of S/MARs with the nuclear matrix has any direct bearing on these functions. The *in vivo* association of S/MARs with the interphase nuclear matrix is still a matter of debate, although the role of S/MARs in the higher order organisation of the eukaryote genome is generally accepted.

S/MARs are not the only DNA elements that are thought to be involved in the functional compartmentalisation of the eukaryotic genome. Sequences have been identified that are able to interfere with long range interactions between *cis*-acting regulatory elements. When these sequences (called boundary elements) are positioned between a promoter and an enhancer, the enhancer is no longer able to activate transcription from that promoter. This is not due to general transcriptional silencing, since the effect is not observed if the boundary element is positioned elsewhere than between enhancer and promoter elements (24–26). Examples of boundary elements are the SCS and SCS′ sequence elements flanking the *Drosophila* heat shock locus (27), a sequence element in the gypsy transposon of *Drosophila* (28) and a sequence element defining the upstream border of the chicken β-globin locus (29). Interestingly, the BEAF32 protein, which through binding to SCS′ is responsible for its boundary function, has recently been shown to be associated with many sites on polytene chromosomes (30). This points to a general role of boundary elements in the organisation of eukaryotic genomes.

Although many S/MARs have been identified, only a limited number of studies have addressed the domain organisation of specific genomic regions (31–33). The best documented analyses are those of the 300 kb *rosy-Ace* region (34), an 800 kb region of chromosome 1 of *Drosophila melanogaster* (35–37) and the 280 kb region around the maize *adh*1 gene (38). These studies indicate that individual domains are diverse in size (5–100 kb) and can contain a variable number of genes. Due to the large size of these regions, their unknown nucleotide sequence and the use of non-overlapping restriction fragments in the S/MAR screen, an unambiguous identification of all S/MARs and genes in these regions was not possible.

In *Arabidopsis thaliana* a large sequence database is compiled containing the primary sequence of its genome. However, no data is available on how this primary sequence could be organized into higher order structures such as chromatin domains. In order to resolve this problem we started a detailed analysis of the organisation of a specific genomic region. In a 16 kb region around the light-regulated plastocyanin gene (PC) of *A.thaliana* (39,40) we have mapped all open reading frames and S/MARs that mark the boundaries of putative chromatin domains. Our analysis reveals a high coding potential with seven genes and we find three S/MARs using a matrix binding assay. All S/MARs are located in the intergenic areas of the PC region, defining the borders between two chromatin domains, each containing two genes.

The three *Arabidopsis* S/MARs are A+T-rich and all have characteristics of S/MARs from other plants, yeast and animals. Interestingly, the three *Arabidopsis* S/MARs share a unique, degenerated 21 bp sequence that is only present in the S/MARs and nowhere else in the 16 kb PC region. From two non-related genomic regions we identified four additional S/MARs. All these S/MARs contain sequences that strongly resemble the 21 bp DNA element found in S/MARs of the plastocyanin region. Based on the alignment of seven *Arabidopsis* S/MAR we propose a 21 bp MAR recognition sequence (MRS) that is unique for *Arabidopsis* S/MARs.

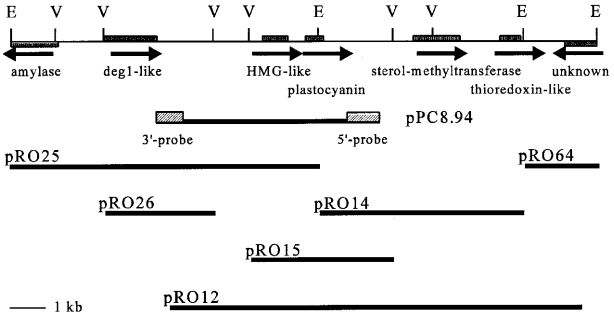

Figure 1. Cloning and characterization of the plastocyanin region. The alignment of the lambda clones (pROx) obtained from the genomic screen is shown in respect to a partial restriction map of the plastocyanin (PC) region with its identified open reading frames. Some of these have domains with significant similartity with the indicated proteins but the identification is not sufficiently established. The arrows indicate the direction of transcription for these genes. Below the map are the positions of the 5′ *Hin*dIII–*Cla*I and 3′ *Bam*HI–*Hin*dIII probes (striped boxes) derived from pPC8.94 that were used in the screen. The restriction sites given are *Bam*HI (B), *Cla*I (C), *Eco*RI (E), *Eco*RV (V) and *Hin*dIII (H).

MATERIALS AND METHODS

Strains and plasmids

The genomic lambdaGEM-11 library of *A.thaliana* ecotype Colombia (constructed by Dr J. Mulligan and Dr R. Davis of Stanford University) was obtained via the Köln DNA Center of the European Union BRIDGE *Arabidopsis* project. Probes for the screen for flanking genomic sequences of the PC gene were derived from pPC8.94 (41), a plasmid containing the *Arabidopsis* plastocyanin gene on a 5.2 kb genomic *Hin*dIII fragment.

Cloning and sequencing of the PC region

Positive lambdaGEM phages were shot-gun cloned using either *Eco*RI, *Eco*RV or *Bam*HI (partial digest) in pBluescript. From a subsequent screen with a 5′ probe of pPC8.94 (*Hin*dIII–*Cla*I) we obtained pRO25 (*Eco*RI insert) and pRO26 (*Eco*RV insert). From the screen with the 3′ probe (*Bam*HI–*Hin*dIII) we obtained pRO14 (*Eco*RI insert) and pRO15 (*Eco*RV insert). The plasmid pRO12 (partial *Bam*HI insert) hybridized with both 5′ and the 3′ probes. A screen with a 3′ probe of pRO12 (*Xba*I–*Bam*HI) resulted in pRO64 (*Eco*RI insert). An overview of these clones is provided in Figure 1. Overlapping subclones from the pRO-series were prepared in order to sequence the entire region. Sequence analysis was supported by the European Scientists Sequencing *Arabidopsis* (ESSA) project of the European Union. Both DNA strands were sequenced, each fragment at least four times. Identification of potential open reading frames was done with the BLAST sequence comparison program (42) and our results were independently confirmed by the Martinsried Institute for Protein Sequencing (MIPS). The sequences of 16 kb PC region (accession number z83321), the 11.2 kb ATH1 region (accession number z83320), and the 4.3 kb ATB2 region (accession number z82043) have been submitted to the EMBL DNA library.

Isolation of nuclei

Nuclei from rat liver cells were isolated as described before (43) and were kept at –80°C in storage buffer [7.5 mM Tris–HCl, pH 7.4, 40 mM KCl, 1 mM EDTA, 0.25 mM spermidine, 0.1 mM spermine, 1% (v/v) thioglycol, 0.2 M sucrose, 50% (v/v) glycerol] at a density of 10^7 nuclei/ml.

Nuclear matrix preparation

Procedures were essentially as described before (44). To obtain matrices for the binding assay rat liver nuclei were subjected to a lithium 3,5-diiodosalicylate (LIS)-extraction protocol (45). Nuclei of 10^7 cells were washed once in 10 ml washing buffer [3.75 mM Tris–HCl, pH 7.4, 20 mM KCl, 0.5 mM EDTA, 0.125 mM spermidine, 0.05 mM spermine, 1% (v/v) thioglycol, 0.1% (w/v) digitonin and 20 µg/ml aprotinin]. After pelleting (300 *g* for 10 min at 4°C) nuclei were gently resuspended in 0.5 ml washing buffer and stabilised by incubation for 20 min at 42°C. Non-matrix proteins were extracted by adding 10 ml of 10 mM LIS in extraction buffer [20 mM HEPES–KOH, pH 7.4, 100 mM lithium acetate, 1 mM EDTA, 0.1 mM PMSF, 0.1% (w/v) digitonin and 20 µg/ml aprotinin] and followed by incubation for 15 min at 25°C. The resulting nuclear halos were collected by centrifugation (15 000 *g* for 5 min at 4°C) and washed four times with 10 ml of digestion buffer (20 mM Tris–HCl, pH 7.4, 70 mM NaCl, 20 mM KCl, 10 mM $MgCl_2$, 0.125 mM spermidine, 0.05 mM spermine and 10 µg/ml aprotinin). For the *in vitro* S/MAR binding assay rat nuclear matrices were obtained by restriction of the genomic DNA of the halos in 1 ml digestion buffer containing 1000 U each of *Eco*RI, *Hin*dIII and *Xho*I for 2 h at 37°C.

S/MAR binding experiments

Rat liver matrix preparation was adjusted to a final concentration of 15 mM EDTA and 120 µg/ml *Escherichia coli* competitor DNA. To identify S/MARs in the 16 kb PC region, nuclear matrices from 10^6 cells were incubated overnight at 37°C with 15 ng of the appropriate [α-^{32}P]dATP end-labelled restriction fragments. After separation into pellet and supernatant fractions by centrifugation (15 000 *g* for 30 min at 4°C) DNA was purified by incubation at 37°C for 60 min with 0.1% SDS and 50 µg/ml proteinase K, followed by phenol–chloroform extraction. DNA was precipitated, dissolved in 50 µl TE and subsequently half of pellet, supernatant or input fractions were loaded on a 1.2% agarose gel. After electrophoresis the gel was dried on Whatman 3MM paper, followed by overnight autoradiography on Kodak X-Omat S film. The quality of our preparations was checked with a positive control: a *Hin*dIII–*Pst*I–*Eco*RI–*Ava*I digest to release the 1000 bp intergenic H1–H3 histone MAR of *Drosophila melanogaster* (45).

RESULTS

Cloning of the plastocyanin region

The analysis of the organisation of the *Arabidopsis* genome is facilitated by its compactness. Earlier studies and the progressing genome project have revealed that major parts of the genome contain coding regions with hardly any interspersed repetitive sequences (46,47). Starting from clone pPC8.94 (41), which spans the plastocyanin (PC) gene, we screened a genomic lambdaGEM library (Materials and Methods) walking in upstream and in downstream directions. Figure 1 depicts the relative position of the six overlapping clones (pROx) that cover ~16 kb around the PC gene. These clones were used in our search for S/MARs. The correct alignment and integrity of the clones was confirmed by Southern blot analysis of the *Arabidopsis* genomic PC region (data not shown). The PC gene has been mapped on chromosome 1 near RFLP marker g6838, using the CEPH/INRA/CNRS YAC library by Drs D. Bouchez and C. Camilleri.

Genomic organisation of the plastocyanin region

Sequence analysis revealed the presence of six open reading frames in addition to the PC gene. To all genes except one we could assign a tentative function on the basis of sequence similarities with known genes in the sequence databases. In Figure 1 we also provide an overview of the position of the newly identified *Arabidopsis* genes and their direction of transcription.

The 5′-edge of the cloned region codes for the N-terminal part of a putative α-amylase gene. The first five exons and part of the sixth exon of the *Arabidopsis* α-amylase gene have a high degree of sequence similarity (86%) with the amylase gene from *Solanum tuberosum* (patent no. WO 9012876-A, DANSICO A/S). The second gene in the PC region contains a central block of 150 amino acids that is also present in a protein of similar molecular weight in *Saccharomyces cerevisiae* (DEG-1) (48) and a rRNA-methyl transferase from *Caenorhabditis elegans* (49). Although the yeast DEG-1 protein is required for normal growth, its function is unknown. An *Arabidopsis* EST (T76494) with 94% identity over 420 bp maps in this area, establishing the functionality of this predicted gene. Immediately upstream of the PC gene we identified a functional gene (EST AT2862; 100% identical over 181 bp) of the HMG-1 superfamily. It codes for a protein that contains a single HMG box, preceded by a basic N-terminal region. This structure closely resembles that of the yeast proteins NHP6A and NHP6B (50), the HMP1 protein from *Plasmodium falciparum* (51), and the NHP1 protein from *Babesia bovis* (52). These small HMG-like proteins are thought to bring distant protein binding sites on the DNA into close proximity by inducing bending of the DNA strand (53).

Downstream of the PC gene we find a gene that probably encodes an enzyme from the sterol metabolism pathway. This putative delta(24)-sterol methyl transferase has a similarity of 61% with the yeast equivalent LIS1/ERG6 (54). There are two non-overlapping *Arabidopsis* ESTs that originate from this gene, ATTS3237 which is 94% identical over 420 bp and AT2536 which is 93% identical over 480 bp. The next gene encodes a thioredoxin-like protein. This protein has a C-terminal domain that closely resembles (52% over 80 amino acids) that of the rabbit, chicken and human thioredoxin proteins (55–57). It is questionable whether this *Arabidopsis* gene codes for an active thioredoxin. In spite of the striking similarity with established thioredoxins the encoded protein lacks the pentapeptide sequence WCGPC typical of the active centre of thioredoxins (58). We are confident that this gene is actively transcribed because of two non-overlapping ESTs; AT267 (97% over 400 bp) and ATTS 3379 (97% over 350 bp). The first EST also confirms the absence of the WCGPC sequence. The most 3′ located incomplete open reading frame has no counterpart in the sequence databases and could, therefore, not be identified yet. Also to this gene we could assign an *Arabidopsis* EST (ATTS5002; 96% identical over 250 bp).

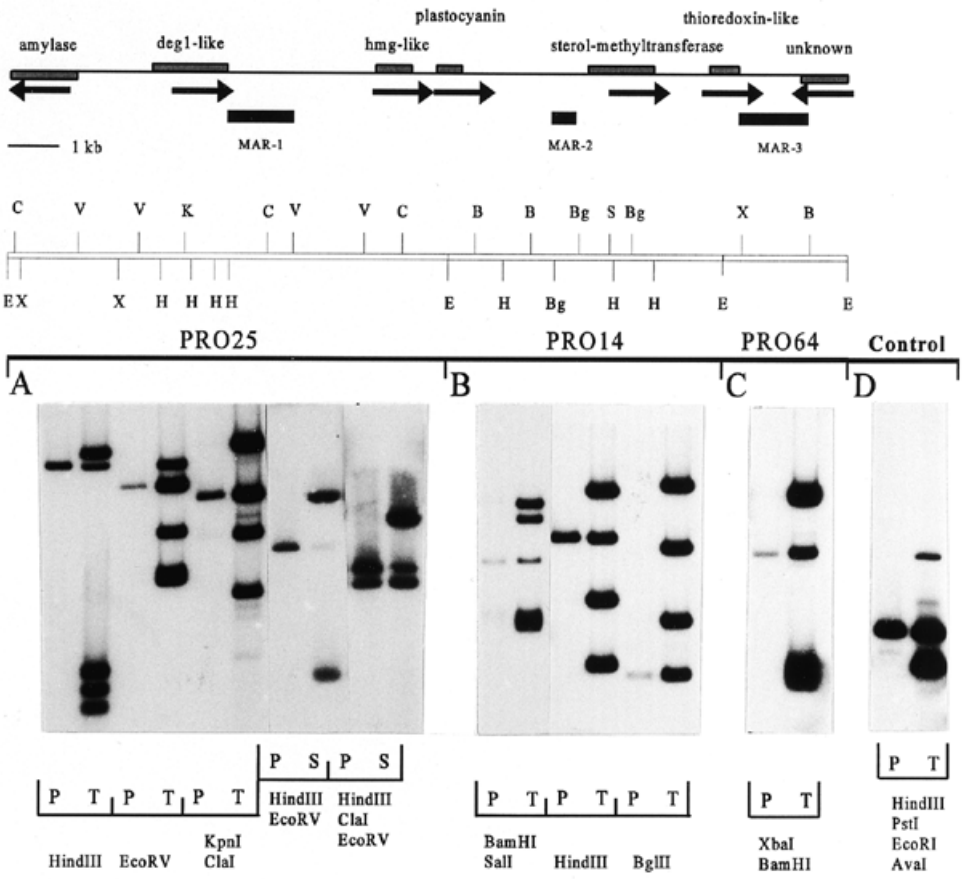

Figure 2. Identification of the S/MARs of the plastocyanin region. (**A**) Binding assay of pRO25 and the cloned MAR. (**B**) Binding assay of pRO14. (**C**) Binding assay of pRO64. (**D**) Binding of the *Drosophila melanogaster* histone H1–H3 intergenic MAR. In the top panel we have indicated the position of the identified S/MARs (black boxes) in relation to the genes. In the central panel are the position of the relevant restriction sites (B, *Bam*HI; Bg, *Bgl*II; C, *Cla*I; E, *Eco*RI; H, *Hin*dIII; K, *Kpn*I; S, *Sal*I; V, *Eco*RV; X, *Xba*I) used to obtain overlapping restriction fragments for the *in vitro* binding assay (P, Pellet; S, Supernatant; T, Total).

Finally, our sequence analysis showed that the PC region is devoid of repetitive sequences and does not contain any CpG-islands.

Mapping of S/MARs in the plastocyanin region

We have used a heterologous matrix binding assay for our screen of S/MARs, as this procedure will identify evolutionary conserved S/MARs, using the histone H1-H3 intergenic MAR of *Drosophila* as a positive control (45). These are the S/MARs that are considered to mark potential boundaries of chromatin domains. Together, the constructs pRO25, pRO14 and pRO64 cover the entire PC region (Fig. 1). Each of these constructs was tested separately for the presence of S/MAR sequences by analysing the binding of overlapping restriction fragments to rat liver matrices.

Figure 2 depicts the results of the S/MAR screen of sequences upstream of the PC gene in pRO25. Within this construct we identified matrix association of a 5500 bp *Hin*dIII fragment and a 3000 bp *Eco*RV fragment (Fig. 2A). These sequences share a 1300 bp region, suggesting the presence of a S/MAR in this overlapping *Hin*dIII–*Eco*RV restriction fragment. A weaker association was observed for a 1700 bp *Kpn*I–*Cla*I (Fig. 2A). This fragment partially overlaps with the *Hin*dIII–*Eco*RV region. From these observations we concluded that the upstream S/MAR is located overlapping the *Cla*I site in the *Hin*dIII–*Eco*RV fragment. To confirm these conclusions we cloned the *Hin*dIII–*Eco*RV fragment and showed strong binding of this fragment to the nuclear matrices (Fig. 2A). The *Hin*dIII–*Cla*I and *Cla*I–*Eco*RV subfragments also bound to the nuclear matrix, albeit with lower affinity.

In a similar screen of pRO14 we analysed sequences directly downstream of the PC gene for matrix association. Here we observed binding of a 1500 bp *Bam*HI–*Sal*I fragment and of an overlapping 2000 bp *Hin*dIII fragment (Fig. 2B). The binding activity of this S/MAR could be assigned to a 500 bp *Bgl*II restriction fragment (Fig. 2B) that resides in both fragments. A second S/MAR downstream of the PC gene was identified in pRO64. Reproducibly, we observed binding of the central 1400 bp *Xba*I–*Bam*HI fragment (Fig. 2C) to the nuclear matrix. Binding persisted even at competitor DNA concentrations that exceeded the 5000-fold molar excess we routinely use in our assay (data not shown). We never observe any binding for the flanking sequences, placing MAR-3 within the 1400 bp *Xba*I–*Bam*HI fragment.

The three constructs pRO25, pRO14 and pRO64, which we used in our screen for S/MARs, span the entire PC region, but do not overlap. A consequence is that we may overlook S/MARs that are located at the junctions between the constructs. Therefore, we also screened pRO12, which overlaps with both junctions, but no additional S/MARs were found. We conclude that the 16 kb PC region contains three S/MARs. The first two S/MARs of the PC

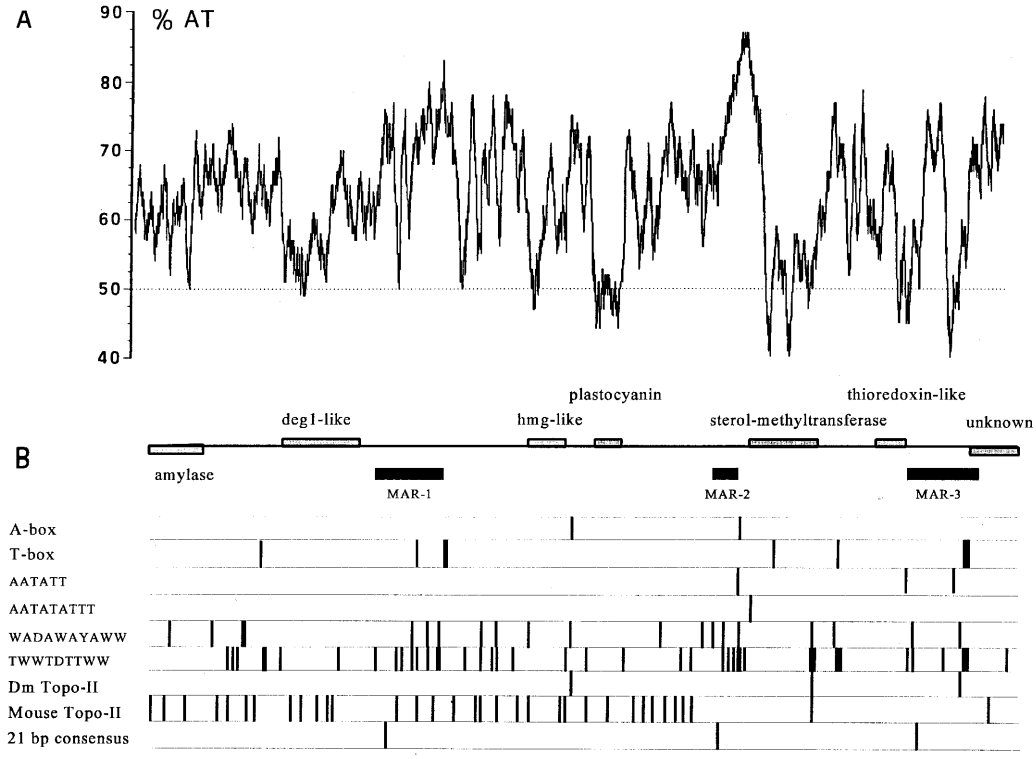

Figure 3. Structural features of the plastocyanin region. (**A**) The percentage of A and T bases (window size = 100) in relation to the position of the genes and S/MARs (central panel) of the plastocyanin region. (**B**) Location of a set of repeated sequences known to be enriched in S/MARs and the position of the 21 bp MAR consensus sequence in relation to the position of the identified S/MARs.

region have a high affinity for rat liver nuclear matrices, whereas the more distal downstream S/MAR may have a somewhat lower affinity. The three S/MARs define two adjacent chromatin domains, each ~5 kb in size and containing two genes.

Sequence characteristics of the S/MARs in the PC region

Over 40 S/MARs have been cloned from a variety of organisms. Although there is no obvious sequence homology between S/MARs, they do share a number of structural characteristics. These include a high A+T content and the presence of several repeat sequences (59). In Figure 3A we compare the A+T-profile of the PC region to the position of the S/MARs. Although all three S/MARs are A+T-rich (>70%), this feature does not uniquely discriminate them from non-S/MAR sequences. There are parts of the PC region that are equally A+T-rich as S/MARs, but do not bind to the nuclear matrix. Evidently, in *Arabidopsis* like in other organisms, S/MARs are A+T-rich, but not all A+T-rich sequences have an affinity for nuclear matrices. As we compared coding regions with non-coding regions it became clear that the non-coding regions are somewhat more A+T-rich than the coding regions. A striking dip in the A+T profile is observed at the start of each of the open reading frames (Fig. 3A). In the case of the PC gene a low A+T level persists throughout the coding region. In the other genes the overall A+T content is only slightly lower than that of the intergenic regions. We conclude that, although the S/MARs confirm the general rule of being A+T-rich, the A+T profile by itself is not a suitable indicator for the localisation of S/MARs in *Arabidopsis*.

S/MARs have also been reported to be enriched in a number of repeated sequences, like AATAAAYAAA (A-box), TTWTWTTW-TT (T-box), WADAWAYAWW, TWWTDTTWW, topoisomerase-II binding sites, and the DNA strand unwinding sequences (AATATT and AATATATTT) (14,34,60–63). In Figure 3B the sequence of the *Arabidopsis* S/MARs indeed show clustering of some of these repeats, but not of all. The S/MAR directly upstream and the one downstream of the PC gene show the highest level of clustering of the repeated sequences TTWTWT-TWTT (T-box), WADAWAYAWW and TWWTDTTWW. Interestingly, these S/MARs have a higher binding affinity for rat liver nuclear matrices than the S/MAR located near the 3′-edge of the cloned area, containing a smaller number of these repeats.

These sequence elements cannot be used to predict the presence of a S/MAR. In Figure 3B we indicated the position of these sequences in the cloned PC region. Some of the repeated sequences do indeed seem to distinguish S/MARs from their environment in the *Arabidopsis* genome. The WADAWAYAWW and TWWTDTTWW repeats are highly clustered and enriched in the S/MARs directly upstream and downstream of the PC gene. However, they are not strictly confined to the S/MARs. Repeat sequences are found mostly in the intergenic regions throughout the cloned PC region, also outside S/MARs. The A-box, T-box and the unwinding sequences are likewise not unique for S/MARs.

A unique *A.thaliana* S/MAR recognition signature

We have aligned the sequences of the three PC S/MARs in search of elements that may be specific for *Arabidopsis* S/MARs.

```
PC-S/MAR1          TAAAAAATGATTATAAGGAAG
PC-S/MAR2          TATAATTTAAAAGTAATTTTG
PC-S/MAR3          TAAATTAGTAAAGTAATGTAG
ATB2-S/MAR         TATATTAT-TTTATAAAAATG
ATH1-S/MAR1        TATATTA--GTAGTAATATTG
ATH1-S/MAR2        TATAAAA--AAAATAAATTTG
ATH1-S/MAR3        TAAATAAAAATGATAAGAAAG

Consensus          TAWAWWWNNAWWRTAANNWWG
```

Figure 4. Potential S/MAR sequence signature. Sequence alignment of the *A.thaliana* S/MARs from the PC region (PC-S/MAR 1, 2 and 3), from the ATB2 region (ATB2-S/MAR) and from the ATH1 region (ATH1-S/MAR 1, 2 and 3). The consensus sequence (W = A or T; R = A or G; N = G, A, T or C) is based on the alignment of the *Arabidopsis* S/MARs and where the bases shared by the overlapping binding sites (TAWAWWW and AWWRTAANNWWG) are in bold.

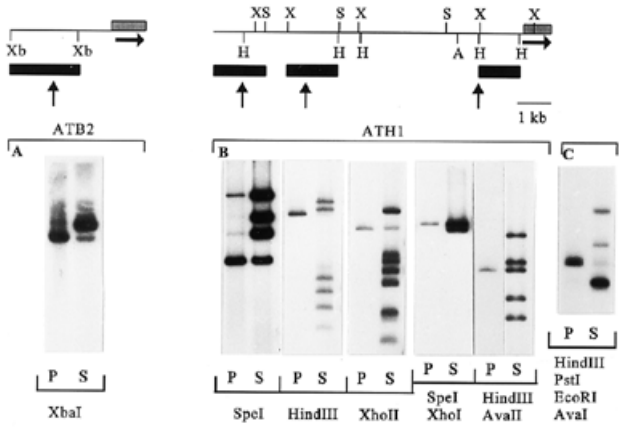

Figure 5. Identification of the S/MARs in the ATB2 and ATH1 genomic regions. (**A**) Binding assay of ATB2. (**B**) Binding assay of ATH1. (**C**) Binding of the *Drosophila melanogater* histone H1–H3 intergenic MAR. In the top panel we have indicated the position of the identified S/MARs (black boxes) in relation to MRSs (arrows) and the position of the relevant restriction sites (A, *Ava*II; H, *Hin*dIII; S, *Spe*I; Xb, *Xba*I; X, *Xho*II) used in the *in vitro* binding assay (P, Pellet; S, Supernatant).

Several short A+T-rich sequences of varying lengths can be observed that are shared by the S/MARs, but none of these discriminate S/MAR from non-S/MAR regions. Such sequences probably simply reflect the overall high A+T content of a region, rather than the presence of a S/MAR. Interestingly, however, the alignment in Figure 4 revealed a degenerated 21 bp sequence: TAWAWWWNNAWWRTAANNWWG. This sequence is unique for the three S/MARs and is not found elsewhere in the PC region (Fig. 3B).

This observation prompted us to investigate whether this sequence could be used to predict the location of a S/MAR in the *Arabidopsis* genome. To this end we screened two non-related genomic regions: a 4.3 kb fragment around the leucine zipper-type transcription factor gene ATB2 (S.C.M.Smeekens, unpublished results) and a 11.2 kb fragment containing the ATH1 transcription factor gene (64) and its upstream region. Figure 5 depicts these results. Matrix binding assays show that the ATB2 clone contains a single S/MAR in the 2.1 kb *Xba*I fragment just upstream of the gene, whereas the rest of the ATB2 region had no affinity for the nuclear matrix (Fig. 5A). Interestingly, we find a DNA element in this S/MAR that strongly resembles the 21 bp consensus sequence. The DNA element in the ATB2 S/MAR has a single N at position eight/nine, where the PC S/MARs have two (Fig. 4). This suggested that the 21 bp PC consensus sequence could be a compound sequence comprised of a closely spaced 7 bp sequence (TAWAWWW) and 12 bp sequence (AWWRTAANNWWG). Given this refinement of the consensus sequence, which does not introduce additional predicted S/MARs in the PC region, we also screened the 11.2 kb genomic fragment around ATH1. The ATH1 region contains two sequences with a closely spaced TAWAWWW and AWWRTAANNWWG (ATH1 S/MAR-1 and -3 in Fig. 4), resembling the S/MARs in the PC region. In addition, ATH1 also contains a region where both sequences partly overlap (ATH1 S/MAR-2 in Fig. 4). A detailed matrix binding analysis of ATH1 region revealed three S/MARs that precisely correspond to the positions of the postulated consensus sequence. The first S/MAR is located in the 1500 bp *Spe*I fragment located at the 5′-end (Fig. 5B). Digestion of this region with either *Xho*II or *Hin*dIII (position 1191 and 993) abolishes matrix binding of the first S/MAR. This locates the S/MAR overlapping these restriction sites, in close proximity to the consensus sequence (position 490). The second S/MAR is located in a *Xho*II–*Hin*dIII restriction fragment (position 2046 and 3762), as is evident from the matrix association of two partially overlapping *Hin*dIII and *Xho*II restriction fragments (Fig. 5B). Notably, this region contains the second consensus sequence (ATH1 S/MAR-2 in Fig. 4) we had identified at position 2813. The prominent association of the *Xho*II and *Hin*dIII fragments contrasts the weaker association of the corresponding *Spe*I fragment that overlaps with these restriction fragments. Although we did not explore this difference in binding behaviour, we believe that it is due to competition for binding to the nuclear matrix between the different S/MARs located on the same construct. Inactivation of binding of the first S/MAR, as a consequence of digestion with either *Xho*II or *Hin*dIII, could conceivably favour association of the second S/MAR. A third S/MAR is located in the most 3′ *Spe*I restriction fragment, containing the promoter region and upstream sequences of the ATH1 transcription factor, as is evident from the weak association of the upper SpeI fragment (Fig. 5B). To refine the position of this third S/MAR we sub-cloned and assayed this region and could show binding of a 1200 bp *Hin*dIII fragment (Fig. 5B). This maps the third S/MAR close to the third MRS at position 7859.

Our results show that all seven S/MARs we have identified in *Arabidopsis* contain a closely spaced combination of TAWAWWW and AWWRTAANNWWG, strongly suggesting that this combination constitutes a S/MAR-specific sequence signature for *Arabidopsis*.

DISCUSSION

The genomic organisation of the plastocyanin region

Although it is generally accepted that the genomic organisation of eukaryotes is an important aspect of gene regulation, so far the topic has not been addressed in *A.thaliana*. In order to investigate this one needs a well defined genomic region in which all genes and S/MARs have been mapped. This paper describes the analysis of the genomic organisation of such a region: a 16 kb genomic domain around the plastocyanin (PC) gene of *A.thaliana*.

To identify S/MAR sequences we made use of the fact that S/MAR–matrix interactions are evolutionary conserved. We

employed a rat liver matrix preparation to detect restriction fragments that bind to the nuclear matrix with high affinity and specificity. The quality of our nuclear matrix preparations was checked using the well studied histone H1–H3 intergenic MAR in *Drosophila melanogaster* as a positive control (Figs 2D and 5C). In this way we have located three S/MARs in a 16 kb region that contains, in addition to the PC gene, six genes and one additional unidentified open reading frame. Each of the three S/MARs is located in an intergenic region. The S/MARs define two potential chromatin domains. Each of the two domains contains two genes: HMG-like and PC in the first and sterol methyl transferase and the thioredoxin-like gene in the second. Evidently, with seven genes in 16 kb the PC region constitutes an area with high coding potential. The *Arabidopsis* genome is small, especially when compared with other plants. Therefore, a high coding density is expected. Preliminary data from the ongoing *Arabidopsis* genome project suggests that highly transcribed regions are organised in clusters (Dr I.Bancroft, personal communication). Starting from the PC gene we probably have isolated part of such a gene cluster.

This detailed description of the two chromatin domains confirms the overall organisation that has been observed in studies on *Drosophila* and maize (35,37,38,65,66). In both these organisms domains of variable sizes (5–100 kb) have been found, in which the smaller contain one up to a small number of genes, whereas the larger seem to contain more repetitive DNA. This distribution leads to an average size of roughly 90 kb, typical for higher eukaryotes (4). In the PC region the domains are rather small (~5 kb), which probably is a consequence of the clustering of genes. As the *Arabidopsis* PC region is devoid of any repetitive DNA sequences, the observed domain size is probably biased and probably does not reflect the *Arabidopsis* average domain size.

Sequence characteristics of the S/MARs in the PC region

S/MARs have been cloned and sequenced from a wide variety of eukaryotes. These S/MARs are characterised by a high A+T content (typically >70%) and are enriched in a specific set of sequences: the repeat sequences AATAAAYAAA (A-box), TTWT-WTTWTT (T-box), WADAWAYAWW and TWWTDTTWWT, one or more topoisomerase II binding sites and the DNA unwinding sequences (AATATT and AATATATTT) (14,34, 59–63). All these sequence elements, including a high A+T content, are found in the three *Arabidopsis* S/MARs of the PC region. Moreover, like almost all known S/MARs, these are located in non-transcribed intergenic regions. This shows that these S/MARs belong to a class of evolutionary conserved sequence elements that specifically bind to the nuclear matrix. However, like in other eukaryotes, these features are not sufficiently unique to recognise a S/MAR from sequence data alone.

An *A.thaliana*-specific S/MAR sequence signature

A comparison of the sequences of the three S/MARs in the PC region revealed a common degenerated 21 bp sequence: TA-WAWWWNNAWWRTAANNWWG. This sequence is unique for the three S/MARs and does not occur elsewhere. To investigate whether this DNA element is specific for *Arabidopsis* S/MARs we screened two non-related genomic regions for S/MAR binding activity. In the ATB2 and the ATH1 regions we identified four additional S/MARs. Only in one case (Fig. 5, ATH1-S/MAR 3) we found the homologous 21 bp sequence. Alignment of the seven *Arabidopsis* S/MARs revealed that the original 21 bp sequence is composed of two closely spaced individual sequence elements. The close combination of TAWAWWW and AWWRT-AANNWWG elements can be found in all seven S/MARs. In all these S/MARs they are found either separated by a small number of base pairs or partly overlapping. We suggest that this configuration of sequence elements is a S/MAR recognition sequence (MRS) that constitutes a landmark for *Arabidopsis* S/MARs. A database survey of genomic *Arabidopsis* DNA revealed that the MRS occurs on average once every 10 kb. To our knowledge, this is the first time that a sequence element has been identified that is truely unique for S/MARs.

We can only speculate what function this MRS fulfils and what proteins bind to this sequence. The configuration of two closely spaced or partially overlapping sequence elements is compatible with two proteins binding adjacently or with one protein with two separate DNA binding domains. As this DNA element occurs only once in each of the S/MARs tested and S/MAR binding activity is known to be dispersed over a region of several hundreds of base pairs, it is unlikely that this sequence by itself is responsible for binding of the S/MAR to the nuclear matrix. Indeed, preliminary experiments have shown that a MRS containing oligo does not bind to the nuclear matrix. As the MRS is specific for *Arabidopsis* S/MARs, but not involved in the physical association of a S/MAR with the nuclear matrix, it is likely that the MRS is required for some S/MAR specific function other than matrix binding. The notion of a modular configuration of S/MARs is compatible with the observation that in transgenic plants the ability of S/MARs to shield against position effects can be uncoupled from their ability to direct copy number-dependent expression (21).

Proteins that contain the AT-rich interaction domain (ARID) are a plausible candidate for the class of proteins that bind to the MRS (67). S/MAR binding proteins, like SatB1 (68) and Bright (69), bind to AT-rich sequence of the ATC-type. Their binding sites are loosely defined but share a common feature: one of the DNA strands is lacking G residues. A similar asymmetrical distribution of C and G residues can be observed in the MRS sequence elements identified in the *Arabidopsis* S/MARs (Fig. 5).

The evolutionary conserved nature of S/MARs suggests that S/MAR binding proteins must be commonly and ubiquitously expressed. This is the case for SAF-A (70), but not for SatB1 and Bright. These latter proteins are tissue specific (68,69). We find this MRS only in *Arabidopsis* S/MARs and not in S/MARs from other organisms, suggesting that the MRS is a binding site for an *Arabidopsis*-specific protein. The observation that SatB1, although specifically expressed in thymus, is able to bind to a large variety of other S/MARs would point to a widespread distribution of ARID proteins with similar but not identical binding sites.

Our analysis of the 16 kb *Arabidopsis* plastocyanin region gives insight into the organisation of a region that contains a cluster of genes. An important question that can now be tackled is whether this organisation is related to the spatial and temporal expression pattern of the genes. Moreover, the identification of a S/MAR-specific sequence element allows (i) the analysis of the chromatin domain organisation of the *A.thaliana* genome from the genomic sequence data, (ii) the identification of proteins that specifically bind to S/MARs, and (iii) the identification of similar S/MAR-specific sequence elements in other organisms.

ACKNOWLEDGEMENTS

This work is supported by a NWO/SLW grant (16.261) to both C.M.v.D. and R.W.O. The authors would like to thank E.van Weezep for technical assistance; Drs J.Mulligan and R.Davis (Stanford University) for their kind gift of the lambdaGEM-11 *A.thaliana* genomic library; Drs D.Bouchez and C.Camilleri (CEPH/INRA/CNRS) for their chromosome mapping data; and the Martinsried Institute for Protein Sequencing for their help in the sequence data analysis. Sequence analysis was supported by the European Scientists Sequencing *Arabidopsis* (ESSA) project of the European Union.

REFERENCES

1 Paranjape,S.M., Kamakaka,R.T. and Kadonaga, J.T. (1994) *Annu. Rev. Biochem.* **63**, 265–297.
2 Wolffe,A.P. (1992) *FASEB J.* **6**, 3354–3361.
3 Laemmli,U.K., Kas,E., Poljak,L. and Adachi,Y. (1992) *Curr. Opin. Genet. Dev.* **2**, 275–285.
4 Jackson,D.A., Dickinson,P. and Cook,P.R. (1990) *EMBO J.* **9**, 567–571.
5 Garrard,W.T. (1990) In Eckstein,F. and Lilley,D.M.J. (eds), *Nucleic Acids and Molecular Biology.* Springer Verlag, Heidelberg. Vol. 4, pp. 163–175.
6 Razin,S.V., Hancock,R., Iarovaia,O., Westergaard,O., Gromova,I. and Georgiev,G.P. (1993) *Cold Spring Harbor Symp. Quant. Biol.* **58**, 25–35.
7 Bode,J., Pucher,H.J. and Maass,K. (1986) *Eur. J. Biochem.* **158**, 393–401.
8 Craddock,C.F., Vyas,P., Sharpe,J.A., Ayyub,H., Wood,W.G. and Higgs,D.R. (1995) *EMBO J.* **14**, 1718–1726.
9 Amati,B. and Gasser,S.M. (1988) *Cell* **54**, 967–978.
10 Amati,B. and Gasser,S.M. (1990) *Mol. Cell. Biol.* **10**, 5442–5454.
11 Avramova,Z. and Bennetzen,J.L. (1993) *Plant Mol. Biol.* **22**, 1135–1143.
12 Avramova,Z. and Paneva,E. (1992) *Biochem. Biophys. Res. Commun.* **182**, 78–85.
13 Bode,J. and Maass,K. (1988) *Biochemistry* **27**, 4706–4711.
14 Gasser,S.M. and Laemmli,U.K. (1986) *Cell* **46**, 521–530.
15 Slatter,R.E., Dupree,P. and Gray,J.C. (1991) *Plant Cell* **3**, 1239–1250.
16 Cockerill,P.N. and Garrard,W.T. (1986) *FEBS Lett.* **204**, 5–7.
17 Breyne,P., Vanmontagu,M. and Gheysen,G. (1994) *Transgen. Res.* **3**, 195–202.
18 Breyne,P., Vanmontagu,M., Depicker,A. and Gheysen,G. (1992) *Plant Cell* **4**, 463–471.
19 Dietz,A., Kay,V., Schlake,T., Landsmann,J. and Bode,J. (1994) *Nucleic Acids Res.* **22**, 2744–2751.
20 Klehr,D., Maass,K. and Bode,J. (1991) *Biochemistry* **30**, 1264–1270.
21 Mlynarova,L., Jansen,R.C., Conner,A.J., Stiekema,W.J. and Nap,J.P. (1995) *Plant Cell* **7**, 599–609.
22 Talbot,D., Descombes,P. and Schibler,U. (1994) *Nucleic Acids Res.* **22**, 756–766.
23 Bode,J., Schlake,T., RiosRamirez,M., Mielke,C., Stengert,M., Kay,V. and Klehr-Wirth,D. (1995) In Berezney,R. and Jeon,K.W. (eds), *Structural and Functional Organization of the Nuclear Matrix*. Academic Press Inc., San Diego,Vol. 162A, pp. 389–454.
24 Cai,H. and Levine,V. (1995) *Nature (London)* **376**, 533–536.
25 Dorsett,D. (1993) *Genetics* **134**, 1135–1144.
26 Kellum,R. and Schedl,P. (1992) *Mol. Cell. Biol.* **12**, 2424–2431.
27 Udvardy,A., Maine,E. and Schedl,P. (1985) *J. Mol. Biol.* **185**, 341–358.
28 Spana,C., Harrison,D.A. and Corces,V.G. (1988) *Genes Dev.* **2**, 1414–1423.
29 Chung,J.H., Whiteley,M. and Felsenfeld,G. (1993) *Cell* **74**, 505–514.
30 Zhao,K., Hart,C.M. and Laemmli,U.K. (1995) *Cell* **81**, 879–889.
31 Levy-Wilson,B. and Fortier,C. (1989) *J. Biol. Chem.* **264**, 21196–21204.
32 Razin,S.V., Petrov,P. and Hancock,R. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 8515–8519.
33 Vandergeest,A.H.M., Hall,G.E., Spiker,S. and Hall,T.C. (1994) *Plant J.* **6**, 413–423.
34 Mirkovitch,J., Gasser,S.M. and Laemmli,U.K. (1988) *J. Mol. Biol.* **200**, 101–109.
35 Brun,C., Qi,D. and Miassod,R. (1990) *Mol. Cell. Biol.* **10**, 5455–5463.
36 Brun,C., Surdej,P. and Miassod,R. (1993) *Exp. Cell Res.* **208**, 104–114.
37 Surdej,P., Got,C., Rosset,R. and Miassod,R. (1990) *Nucleic Acids Res.* **18**, 3713–3722.
38 Avramova,Z., Sanmiguel,P., Georgieva,E. and Bennetzen,J.L. (1995) *Plant Cell* **7**, 1667–1680.
39 Fisscher,U., Weisbeek,P. and Smeekens,S. (1994) *Plant Mol. Biol.* **26**, 873–886.
40 Vorst,O., Kock,P., Lever,A., Weterings,B., Weisbeek,P. and Smeekens,S. (1993) *Plant J.* **4**, 933–945.
41 Vorst,O., Oosterhoff-Teertstra,R., Van Kan,P., Smeekens,S. and Weisbeek,P. (1988) *Gene* **65**, 59–69.
42 Altschul,S.F., Gish,W., Miller,W., Myers,W.E. and Lipman,P.J. (1990) *J. Mol. Biol.* **215**, 403–410.
43 Izaurralde,E., Mirkovitch,J. and Laemmli,U.K. (1988) *J. Mol. Biol.* **200**, 111–125.
44 Ludérus,M.E.E., Degraaf,A., Mattia,E., Den Blaauwen,J.L., Grande,M.A., De Jong,L. and Van Driel,R. (1992) *Cell* **70**, 949–959.
45 Mirkovitch,J., Mirault,M.-E. and Laemmli,U.K. (1984) *Cell* **39**, 223–232.
46 Leutwiler,L.S., Hough-Evans,B.R. and Meyerowitz,E.M. (1984) *Mol. Gen. Genet.* **194**, 15–23.
47 Pruitt,R.E. and Meyerowitz,E.M. (1986) *J. Mol. Biol.* **187**, 169–183.
48 Carbone,M.L., Solinas,M., Sora,S. and Panzeri,L. (1991) *Curr. Genet.* **19**, 1–8.
49 Wilson,R., Ainscough,R., Anderson,K., Baynes,C., Berks,M., Bonfield,J., Burton,J., Connell,M., Copsey,T., Cooper,J., *et al.* (1994) *Nature (London)* **168**, 32–38.
50 Kolodrubetz,D. and Burgum,A. (1990) *J. Biol. Chem.* **265**, 3234–3239.
51 Kun,J.F. and Anders,R.F. (1995) *Mol. Biochem. Parasitol.* **71**, 249–253.
52 Dalrymple,B.P. and Peters,J.M. (1992) *Biochem. Biophys. Res. Commun.* **184**, 31–35.
53 Paull,T.T. and Johnson,R.C. (1995) *J. Biol. Chem.* **270**, 8744–8754.
54 Hardwick,K.G. and Pelham,H.R. (1994) *Yeast* **10**, 265–269.
55 Jones,S.W. and Luk,C.K. (1988) *J. Biol. Chem.* **263**, 9607–9611.
56 Tonissen,K.F. and Wells,J.R. (1991) *Gene* **102**, 221–218.
57 Wollman,E.E., D'Auriol,L., Rimsky,L., Shaw,A., Jacquot,J.-P., Wingfield,P., Graber,P., Dessarps,F., Robin,P., Galibert,F., *et al.* (1988) *J. Biol. Chem.* **263**, 15506–15512.
58 Holmgren,A. (1989) *J. Biol. Chem.* **264**, 13963–13966.
59 Boulikas,T. (1995) In Berezney,R. and Jeon,K.W. (eds), *Structural and Functional Organization of the Nuclear Matrix.* Academic Press Inc, San Diego. Vol. 162A, pp. 279–388.
60 Bode,J., Kohwi,Y., Dickinson,L., Joh,T., Klehr,D., Mielke,C. and Kohwi-Shigematsu,T. (1992) *Science* **255**, 195–197.
61 Glazkov,M.V. (1995) *Mol. Biol.* **29**, 561–571.
62 Hall,G., Allen,G.C., Loer,D.S., Thompson,W.F. and Spiker,S. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 9320–9324.
63 Roberge,M. and Gasser,S.M. (1992) *Mol. Microbiol.* **6**, 419–423.
64 Quaedvlieg,N., Dockx,J., Rook,F., Weisbeek,P. and Smeekens,S. (1995) *Plant Cell* **7**, 117–129.
65 Mirkovitch,J., Spierer,P. and Laemmli,U.K. (1986) *J. Mol. Biol.* **190**, 255–258.
66 Surdej,P., Brandli,D. and Miassod,R. (1991) *Biol. Cell.* **73**, 111–120.
67 Nakagomi,K., Kohwi,Y., Dickinson,L.A. and Kohwi-Shigematsu,T. (1994) *Mol. Cell. Biol.* **14**, 1852–1860.
68 Cunningham,J.M., Purucker,M.E., Jane,S.M., Safer,B., Vanin,E.F., Ney,P.A., Lowrey,C.H. and Nienhuis,A.W. (1994) *Blood* **84**, 1298–1308.
69 Herrscher,R.F., Kaplan,M.H., Lelsz,D.L., Das,C., Scheuermann,R. and Tucker,P.W. (1995) *Genes Dev.* **9**, 3067–3082.
70 Fackelmayer,F.O., Dahm,K., Renz,A., Ramsperger,U. and Richter,A. (1994) *Eur. J. Biochem.* **221**, 749–757.