ELSEVIER

Biochimica et Biophysica Acta 1517 (2001) 351–364

BIOCHIMICA ET BIOPHYSICA ACTA
BBA

www.elsevier.com/locate/bba

Comparative study and prediction of DNA fragments associated with various elements of the nuclear matrix

Galina V. Glazko [a,b,*], Igor B. Rogozin [a,c], Mikhail V. Glazkov [d]

[a] *Institute of Cytology and Genetics, Novosibirsk 630090, Russia*
[b] *Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, 311 Mueller Laboratory, University Park, PA 16802, USA*
[c] *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA*
[d] *Institute of General Genetics, Moscow 117809, Russia*

Received 13 July 2000; received in revised form 7 November 2000; accepted 1 December 2000

Abstract

Scaffold/matrix-associated region (S/MAR) sequences are DNA regions that are attached to the nuclear matrix, and participate in many cellular processes. The nuclear matrix is a complex structure consisting of various elements. In this paper we compared frequencies of simple nucleotide motifs in S/MAR sequences and in sequences extracted directly from various nuclear matrix elements, such as nuclear lamina, cores of rosette-like structures, synaptonemal complex. Multivariate linear discriminant analysis revealed significant differences between these sequences. Based on this result we have developed a program, ChrClass (Win/NT version, ftp.bionet.nsc.ru/pub/biology/chrclass/chrclass.zip), for the prediction of the regions associated with various elements of the nuclear matrix in a query sequence. Subsequently, several test samples were analyzed by using two S/MAR prediction programs (a ChrClass and MAR-Finder) and a simple MRS criterion (S/MAR recognition signature) indicating the presence of S/MARs. Some overlap between the predictions of all MAR prediction tools has been found. Simultaneous use of the ChrClass, MRS criterion and MAR-Finder programs may help to obtain a more clearcut picture of S/MAR distribution in a query sequence. In general, our results suggest that the proportion of missed S/MARs is lower for ChrClass, whereas the proportion of wrong S/MARs is lower for MAR-Finder and MRS.

Keywords: Nuclear matrix; Scaffold/matrix-associated region; Computer-based prediction; Nucleotide motifs; Discriminant analysis; Satellite DNA

1. Introduction

The model of loop-domain organization of eukaryotic chromosomes is now widely accepted [1–4]. According to this model topologically independent chromatin loops are attached to the nuclear matrix/scaffold. Various nuclear matrix/scaffold proteins, which potentially participate in the loop organization of chromosomes, have been identified and some of their characteristics have been studied [1,4–8]. Sites at which chromosomal DNA is attached to the nuclear matrix/scaffold are called S/MARs (scaffold/matrix-associated regions) [2,9]. A number of possible functions have been discussed earlier for S/MARs, which include forming boundaries of chromatin domains, changing of chromatin conformations, participating in initiation of DNA replication and organizing the chromatin structure of a chromosome [10–12]. S/MARs are common in centromere-associated DNA [13] and telomeric arrays [14,15], and appear to be important in mitotic chromosome assembly and maintenance of chromosome shape during metaphase [16]. Thus, S/MARs are involved in multiple independent processes during different stages of the cell cycle, and indeed this functional plethora may be the key reason for the high heterogeneity in S/MAR sequences. At first glance a typical S/MAR is built from several nucleotide motifs, none of which is highly specific (reviewed in [17]). The recent identification of the S/MAR

Abbreviations: LDA, linear discriminant analysis; S/MAR, scaffold/matrix-associated region; scDNA, fragments of chromosomal DNA extracted from synaptonemal complex; rsDNA, fragments of chromosomal DNA extracted from the protein cores of rosette-like structures; nlDNA, fragments of chromosomal DNA extracted from the nuclear lamina; 5′flDNA, 5′ flanking regions of eukaryotic genes; randDNA, random sequences; MDS, minimal diagnostic set; CCPQ, ChrClass prediction quality; MRS, S/MAR recognition signature

* Corresponding author. Fax: +1-814-863-7336;
E-mail: gvg2@psu.edu

0167-4781/01/$ – see front matter
PII: S0167-4781(00)00297-9

recognition signature (MRS) simultaneously revealed S/MARs without this motif [18], similar to the case of the DNA topoisomerase II consensus sequence [19]. Many characterized S/MARs are AT-rich, but simply being AT-rich does not make a DNA fragment a S/MAR [13,20,21].

Another S/MAR identification problem is the variability of the nuclear matrix itself. Matrix attachment regions are identified by in vitro binding assays that measure the affinity of a cloned DNA fragment with the nuclear matrix. However, different methods of isolation (2 M NaCl, LIS detergent, electroelution) can yield different nuclear matrix contents, thus, a nuclear matrix is a functionally defined structure. A typical nuclear matrix from somatic cells consists of various elements, including the nuclear pores–lamina complex, residual nucleoli, and a fibrillar-granular network (inner matrix) [22]. The latter is considered to be the most variable [23,24]. In addition, in meiotic cells the synaptonemal complex (SC) proteins are an integral part of the nuclear matrix in prophase I [25]. Therefore, it has been generally assumed that preexisting nuclear matrix/lamina proteins participate in the molecular organization of the synaptonemal complex during meiotic prophase and presumably coevolve with some synaptonemal complex protein (SCP1) [26–28]. Consequently, properties of nucleotide sequences associated with the nuclear matrix depend on the method of their isolation, i.e. they could be defined by the matrix elements that these sequences are associated with. All S/MARs known in the chromosomes of somatic cells were identified through specific interactions of DNA with the nuclear matrix, however, experimental conditions may vary considerably [9,10]. To date there is no experimental evidence to suggest which particular elements of the nuclear matrix the S/MARs are associated with, although there are examples of chromosomal DNA fragments directly extracted from chromosomal/nuclear substructures, i.e. elements of the nuclear matrix, such as the nuclear lamina, cores of rosette-like structures (which are presumably the part of the fibrillar-granular network, i.e. the inner matrix), and synaptonemal complex (in meiotic cells) [29–32]. Initially we suggested there should be some kind of correspondence (at the nucleotide level) between these sites and S/MAR sequences. In what follows, we will discuss mainly the characteristics of S/MARs; however, the other classes of sites mentioned above may possess very important functions and should also be taken into consideration.

We used a multivariate linear discriminant analysis (LDA) to compare the nucleotide sequences of chromosomal DNA fragments extracted from various chromosomal/nuclear substructures and S/MARs. Also, to obtain a clearcut picture of the differences between samples of sequences we use 5′ flanking regions (first 1500 bp) of some tissue-specific human genes and random sequences as an 'outgroup' (it was assumed that these sequences are not related to any elements of the nuclear matrix). Significant differences among all DNA samples were observed. Based on the results of the LDA, we developed the ChrClass program to predict sites associated with various elements of the nuclear matrix. To estimate the prediction accuracy of this ChrClass program, several test samples were analyzed using two computer programs, a ChrClass and MAR-Finder. The latter predicts S/MAR sequences [33]. Test samples were also analyzed for the presence of MRS. We found that the proportion of missed S/MARs is lower for ChrClass, whereas the proportion of wrong S/MARs is lower for MAR-Finder and the MRS criterion. Thus, predictions based on these tools should be interpreted with caution, and the problem of S/MAR prediction will require further analysis. The choice of a method to analyze a query sequence suitable to the problem being studied depends on the sequence length and preferences in the proportion of missed and wrong S/MARs. We will discuss the choice problem in Section 3.

2. Materials and methods

A complete description of samples is presented in the Appendix. Here we describe them briefly.

2.1. Training samples

1. S/MARs: sample volume $n_1 = 27$ (S/MARs obtained from GenBank, EMBL, S/MARt-DB databases and original papers; 16 and 11 S/MARs were from vertebrates and plants, respectively).
2. rsDNA: sample volume $n_2 = 16$ (DNA fragments were extracted from the cores of rosette-like structures of mouse interphase chromosomes) [30].
3. sc1DNA: sample volume $n_3 = 17$ (DNA fragments were extracted from the synaptonemal complex of Chinese hamster) [31].
4. sc2DNA: sample volume $n_4 = 18$ (DNA fragments were extracted from the synaptonemal complex of rat) [32].
5. nlDNA: sample volume $n_5 = 25$ (DNA fragments were extracted from the nuclear lamina of mouse oocytes) [29].
6. 5′flDNA: sample volume $n_6 = 24$ (5′ flanking regions of tissue-specific eukaryotic genes).
7. randDNA: sample volume $n_7 = 116 = n_1+n_2+...+n_6$ ('random' DNA fragments obtained using a random generator assuming a uniform distribution of all four nucleotides and lengths of real sequences from samples 1–6).

2.2. Test samples

1. Relatively short sequences annotated as S/MARs in GenBank to investigate the ChrClass and MRS prediction power:
   - hsDNA: 12 fragments associated with the nuclear matrix from the human chromosome 19 [34];

- tbDNA: 12 fragments of the chromosomal DNA associated with the nuclear matrix of tobacco [35];
- chDNA: 4 fragments of the chromosomal DNA associated with the nuclear matrix of Chinese hamster [21].

2. Relatively long (cosmid-sized) sequences to investigate the ChrClass, MAR-Finder and MRS prediction power simultaneously:
   - glDNA: 3 β-globin gene clusters from different species: human, 73 308 bp; galago, 57 113 bp; rabbit, 44 594 bp. The β-globin gene cluster and 16 609 bp locus of control region (LCR) were manually merged together for rabbit, since they are presented in GenBank by two different entries. The members of the β-globin gene cluster from different species as well as revealed S/MARs were positioned according to the beginning of the first exon in the human ε-globin gene (locus: HUMHBB);
   - ptDNA: 30 035 bp and 42 447 bp of the colinear Sh2/A1 homologous regions of rice and sorghum respectively [19].
3. A hundred random sequences (with the total length of 100 000 nucleotides) were generated based on the assumption of equal nucleotide frequencies.
4. Human alphoid DNA of different centromeric regions (obtained from GenBank, 1998) and an 'ideal' telomeric sequence $(TTAGGG)_{500}$ were also included in our analysis to test the experimentally observed ability of such sequences to associate with various elements of the nuclear matrix.

2.3. Statistical data processing (linear discriminant analysis)

The training samples were compared using pairwise and multivariate linear discriminant analysis (LDA). For each sample of the chromosomal DNA fragments (training samples), frequencies of simple nucleotide motifs $\{X_1,...,X_n\}$ (described in Section 3) were calculated using an ad hoc program. Then, LDA was used to construct the discrimination functions between different samples of sequences.

Briefly, the strategy of LDA includes two steps. Firstly, the probability P of a successful discrimination between different classes of sequences (each class corresponds to one training sample) is estimated. Secondly, if there is a set of discriminating variables ($P < 0.05$), a classification step is carried out. In the classification step, a discriminant function is constructed such that each sequence is attributed to one of the classes. The discriminant function h_k is a linear combination of the input characteristics $\{X_1,...,X_n\}$:

$$h_k = b_{k0} + b_{k1}X_1 + ... + b_{kn}X_n \quad (1)$$

The discriminant function coefficients b_{ki} are calculated as inverse to the matrix intraclass variance:

$$b_{ki} = (N-K)\sum_{j=1}^{n} a_{ij}X_{jk} \quad (2)$$

where a_{ij} is the matrix element, N is the sample volume and K is the number of classes. Each sequence is attributed to a class that gives a maximal value of h_k.

To estimate the Mahalanobis distances between various classes of sequences and posterior classification of the training samples we used the *Statistica* 5.0 software (the 'discriminant' analysis module). The samples were clustered based on the obtained matrix of Mahalanobis distances using the 'nearest neighbor' technique (the 'cluster analysis' module).

3. Results

3.1. Analysis of simple nucleotide motifs in S/MAR, rsDNA, nlDNA, sc1DNA, sc2DNA, 5′flDNA and randDNA samples

It is well known that some nucleotide motifs influence DNA conformation under certain conditions. For example, both purine and pyrimidine tracts can form triple strands [36], whereas $(AT)_n$ and $(GC)_n$ tracts can form the Z-DNA structure [37]. In addition, $(AT)_n$ tracts can form cruciform DNA [17] and cause an unfolding of the DNA duplex [38]. Certain combinations of TG, TA, CA can result in a curved 'kinked' DNA [39]. To compare the chromosomal DNA fragments some simple nucleotide motifs were chosen. In the beginning of our study a set of only 13 sequence characteristics was used, which included the 'kinked DNA' motifs, short palindromes, tracts of $poly(A)_n$, $poly(T)_n$, $poly(C)_n$, and $poly(G)_n$ ($n \geq 4$); tracts of $(AT)_m$ and $(GC)_m$ ($m = 5, 6, > 6$; m means total length of the motif, but not the number of the repeated units; for example, if $m = 5$ the motif will be WATAT or SGCGC, W = A or T, S = G or C), and also tracts consisting of R (A or G), Y (T or C), S (G or C), W (A or T), K (G or T), M (A or C) (more than six nucleotides). Below, we refer to this set of characteristics as a 'minimal diagnostic set' (MDS). Our choice is based on the generally accepted viewpoint on various context features of S/MARs [2,17,40]. Such features include (1) the narrow minor groove that occurs in AT-rich DNA, (2) base-unpairing regions which have been determined in many cases to be just AATATATTT [38], (3) curved DNA motifs $A_4N_7A_4N_7A_4$ or T_3A_3, (4) (A+T)-rich motifs shared by the origins of replication and S/MARs [41]. All these motifs may be described as a superposition of AT, TA, poly(A) and poly(T) tracts. The other commonly used S/MAR motif is the TopoII consensus sequence which

can be described by an $(RY)_n$ tract according to the experimental data of Spitzner et al. [42]. This way we attempted to construct an MDS that takes into account the context features of S/MARs and at the same time possesses some flexibility.

In general, such an MDS should be sufficient to obtain a reliable prediction for S/MAR sequences. However, a preliminary analysis with the above motifs gives accuracies for posterior classification of 75%, 56%, 36%, 29%, 33% and 75% for S/MARs, rsDNA, nlDNA, sc1DNA, sc2DNA, and 5′flDNA, respectively. These results suggest that modifications to the MDS are required. Unfortunately, there is not much information about nucleotide or structural motifs within the chromosomal fragments associated with either the nuclear lamina, the inner matrix (i.e. protein cores of rosette-like structures), or the synaptonemal complex. DNA sequences associated with the nuclear lamina bear some specific motifs recognized by several lamina proteins. Lamins A, B, and C can specifically bind to homopolytracts, whereas lamins A and C can associate with the telomeric repeat of vertebrates [43]. On the other hand, the telomeric repeat $(TTAGGG)_n$ has been shown to be associated with the nuclear matrix [14,15]. Alphoid DNA also was found to be associated with the nuclear matrix [13], the nucleolus, and the nuclear lamina [44]. However, this repetitive telomeric and alphoid DNA does not contain MDS motifs (results not shown). Therefore, these sequences contain some specific motifs (different from MDS motifs) that are able to associate with the nuclear matrix. Thus some unknown context characteristics of the telomeric and alphoid DNA can be taken into account in order to improve the prediction. In such cases of uncertainty, it is more sensitive to characterize sequences by the statistics of short oligonucleotides [45], e.g. the telomeric repeat $(TTAGGG)_n$ can be described as a combination of six triplets (TTA, GGG, TAG, AGG, GGT, GTT). MDS has been enriched by the triplet frequencies that were also used in the LDA procedure.

3.2. Heterogeneity of simple nucleotide motifs

All training samples used have their own a priori classification such as DNA fragments that are extracted from the nuclear lamina, the synaptonemal complex, the protein cores of rosette-like structures, S/MARs, and 5′ flanking sequences. Some samples (S/MARs and scDNA) include DNA fragments extracted by different techniques and from various regions of genes (for example, introns and flanking regions). To take this into account, at the first stage of the analysis we divided the S/MAR sample into three subsamples. S/MARs located in animal genes in (1) the flanking regions (S/MARs1) and (2) intronic regions (S/MARs3), and since all of our initial plant S/MARs were found outside the intronic regions, (3) S/MARs located in the flanking regions of plant genes (S/MARs2). The sample of sequences associated with the synaptonemal complex was naturally divided into two subsamples, sc1DNA (from Chinese hamster) and sc2DNA (from rat).

The analysis of simple nucleotide motifs in the examined chromosomal DNA samples shows that each sample might have a set of characteristic features that distinguish it from the other samples (Table 1). For example, S/MAR fragments in all three subsamples are saturated with AT motifs and 'kinked' DNA, whereas low frequencies of these motifs are typical for scDNA and rsDNA fragments (Table 1).

Based on the frequencies of the sequence context characteristics, we carried out a pairwise LDA between all the samples of interest in order to determine the relative heterogeneity of S/MARs and scDNA. The LDA-obtained Mahalanobis distance between plant and animal S/MARs that are localized in flanking regions appears to be the smallest among all pairwise distances (results not shown). A possible explanation for the similarity between plant and animal flanking regions is that S/MARs have some highly similar context features in different phyla. As long as the chromosomes are anchored via specific sequences to the nuclear matrix, these anchoring sequences may have some kind of similarity to interact with nuclear matrix proteins. One such possible common feature of many S/MARs is a very high A+T content. It has been demonstrated that nuclear matrices prepared from the yeast specifically bind a MAR sequence derived from the mouse κ light chain immunoglobulin gene [46]. Our results also show a low intertaxon S/MARs heterogeneity. This fact might support the hypothesis that some S/MAR-specific motifs are evolutionarily conserved and different types of nuclear matrix proteins have evolved without changing the context specificity of interaction with S/MARs.

S/MARs in introns are closer to S/MARs in flanking regions than to any other sample of DNA fragments. Thus the internal heterogeneity of the S/MAR sample is lower than the differences between S/MARs and other samples. For this reason all S/MARs were subsequently analyzed together. The distance between the sc1DNA and sc2DNA samples appears to exceed the distance between them and the samples of nlDNA fragments (results not shown). These samples were considered separately. The revealed difference between sc1DNA and sc2DNA could be due to the differences in experimental conditions, however, we cannot exclude that the set of context characteristics used in this study is not optimal for scDNA recognition since MDS was constructed for S/MAR sequences specifically.

3.3. Comparative analysis of DNA samples

We used a multivariate LDA technique in order to analyze differences among all the samples simultaneously. Sig-

Table 1
The frequency distribution for some simple nucleotide motifs in the samples of sequences associated with various elements of the nuclear matrix, in the 5′flDNA and in randDNA samples

	S/MARs1	S/MARs2	S/MARs3	S/MARs	5′flDNA	sc1DNA	sc2DNA	nlDNA	rsDNA	randDNA
'Kinked' DNA	0.105[a]	0.107	0.107	0.106	0.090	0.090	0.084	0.096	0.083	0.084
	0.009[b]	0.006	0.004	0.007	0.020	0.020	0.016	0.021	0.029	0.056
$Poly(A)_3$	0.053	0.048	0.049	0.049	0.027	0.024	0.024	0.032	0.026	0.013
	0.024	0.015	0.021	0.021	0.015	0.026	0.018	0.028	0.022	0.003
$Poly(T)_3$	0.058	0.054	0.057	0.057	0.025	0.038	0.022	0.019	0.039	0.014
	0.023	0.019	0.020	0.020	0.011	0.050	0.018	0.017	0.043	0.005
$Poly(C)_3$	0.006	0.007	0.007	0.007	0.021	0.012	0.020	0.001	0.013	0.018
	0.004	0.006	0.006	0.005	0.013	0.013	0.024	0.010	0.015	0.005
$Poly(G)_3$	0.008	0.005	0.008	0.007	0.025	0.014	0.016	0.016	0.013	0.014
	0.007	0.003	0.002	0.005	0.013	0.001	0.014	0.013	0.017	0.004
$(AT)_2$	0.030	0.024	0.023	0.027	0.008	0.012	0.005	0.013	0.007	0.006
	0.027	0.014	0.006	0.020	0.008	0.013	0.006	0.014	0.007	0.002
$(GC)_2$	0.001	0.001	0.0	0.0005	0.004	0.014	0.002	0.002	0.0002	0.010
	0.001	0.001	0.0	0.001	0.010	0.010	0.008	0.004	0.001	0.004
TC[c]...	0.010	0.008	0.010	0.009	0.011	0.010	0.010	0.009	0.013	0.006
	0.006	0.001	0.001	0.005	0.004	0.008	0.007	0.007	0.010	0.002
AG[c]...	0.008	0.007	0.005	0.007	0.010	0.010	0.007	0.009	0.013	0.008
	0.004	0.003	0.002	0.004	0.005	0.008	0.006	0.007	0.010	0.002
TG[c]...	0.006	0.007	0.009	0.007	0.007	0.006	0.008	0.006	0.006	0.008
	0.005	0.001	0.003	0.004	0.004	0.007	0.006	0.005	0.009	0.002
AT[c]...	0.023	0.023	0.024	0.022	0.010	0.009	0.006	0.013	0.007	0.006
	0.009	0.008	0.007	0.010	0.007	0.008	0.006	0.008	0.009	0.002
GC[c]...	0.001	0.001	0.0	0.001	0.005	0.001	0.001	0.002	0.001	0.007
	0.001	0.001	0.0	0.001	0.008	0.002	0.003	0.002	0.002	0.002
AC[c]...	0.006	0.007	0.006	0.006	0.008	0.009	0.009	0.001	0.007	0.006
	0.003	0.002	0.002	0.003	0.003	0.006	0.006	0.009	0.005	0.001

[a]The expectation (E_x).
[b]The variance (D_x).
[c]Tracts consist of these arbitrary combined dinucleotides and contain more than six bases.

nificant differences between the samples were observed (Wilk's $\Lambda = 0.002$, $P < 0.001$). This suggests that all samples are highly heterogeneous in terms of the frequencies of the context characteristics and cannot be merged together in one sample. The result of the cluster analysis (the 'nearest neighbor' method), based on the Mahalanobis distances (Table 2), is shown in Fig. 1. Here one can see that the examined chromosomal DNA fragment samples can be divided into at least four classes: (1) S/MARs; (2) nlDNA, rsDNA, sc1DNA, sc2DNA; (3) 5′flDNA; (4) randDNA. It should be noted that the obtained Mahalanobis distances show greater similarity of S/MARs with sc1DNA and 5′flDNA than with all other samples.

3.4. Posterior classification

During the posterior classification, the homogeneity of each sample was examined and classification errors ('outsiders') were identified (Table 3). Homogeneity of 93% for the S/MARs sample was remarkably high (Table 3). One outsider was the 3′ MAR of the chicken α-globin gene, which was classified as a 5′ flanking region. Another outsider was the MAR in the distal mouse chromosome 7 imprinted domain, which was classified as a DNA fragment from nuclear lamina. The least heterogeneity was observed for the rsDNA fragments (100% of true classification). The nlDNA sample can be considered the most heterogeneous of all samples examined. There were five

Table 2
Pairwise Mahalanobis distances between the samples of sequences associated with various elements of the nuclear matrix, the 5′flDNA sample and the randDNA sample

	S/MARs	rsDNA	sc1DNA	nlDNA	sc2DNA	5′flDNA	randDNA
S/MARs	0.00	46.21	19.99	29.62	35.12	20.13	41.50
rsDNA	46.21	0.00	29.64	18.21	25.70	22.09	54.89
sc1DNA	19.99	29.64	0.00	14.83	23.05	16.03	37.88
nlDNA	29.62	18.21	14.83	0.00	21.56	17.19	41.95
sc2DNA	35.12	25.70	23.05	21.56	0.00	19.75	46.01
5′flDNA	20.13	22.09	16.03	17.19	19.75	0.00	40.79
randDNA	41.50	54.89	37.88	41.95	46.01	40.79	0.00

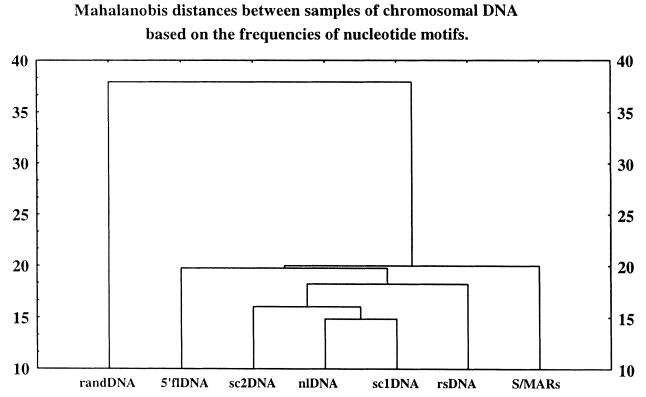

Fig. 1. Clustering of the samples of sequences associated with various elements of the nuclear matrix, the 5′flDNA sample and the randDNA sample.

outsiders in 25 sequences. Three of the five were attributed to the sc1DNA sample, and two to either sc2DNA or the 5′flDNA sample. In the sc1DNA sample, only one outsider was attributed to the 5′flDNA sample. In the sc2DNA sample one outsider was attributed to the nlDNA sample and another one was attributed to the rsDNA sample. 100% of random DNA fragments were classified correctly. In general, LDA has revealed a high internal homogeneity of all original samples.

3.5. The ChrClass program to predict DNA fragments associated with various elements of nuclear matrix

Significant differences between samples of chromosomal DNA fragments ($\Lambda = 0.002$) allowed us to develop a classification program, the ChrClass. It predicts regions associated with different elements of the nuclear matrix in a query sequence based on the results of the LDA. The program analyzes 300–1000 bp regions of the query sequence at a time. The output highlights four classes of regions in a query sequence: (1) S/MARs, (2) regions associated the inner matrix (with the protein cores of rsDNA), (3) regions associated with the synaptonemal complex (scDNA), and (4) regions associated with the nuclear lamina (nlDNA). 5′-Flanking regions and random sequences were not included in the program output.

The ChrClass program (Win95/NT version 1.1) is available from Galina V. Glazko (gvg2@psu.edu), Igor B. Rogozin (rogozin@bionet.nsc.ru) or from an anonymous ftp site (ftp.bionet.nsc.ru/pub/biology/chrclass/chrclass.zip).

3.6. Analysis of relatively short sequences annotated as S/MARs in GenBank to investigate the ChrClass and MRS prediction power

We have tested ChrClass on a number of sequence samples in order to evaluate its prediction accuracy. Test sample 1 includes entries annotated in GenBank as S/MARs. These relatively short sequences were extracted in vitro using various matrix binding assays and are 300–4000 bp long (the limits of S/MARs length). For this reason the MAR-Finder program, developed for cosmid-sized sequences, cannot be used to analyze these sequences [33]. For comparison, test sample 1 was also analyzed by the MRS criterion [18]. The sensitivity for ChrClass is estimated as a proportion (%) of the most representative class (S/MARs, rsDNA, nlDNA or scDNA) in a given sample. For example, in the sample of DNA fragments (hsDNA) obtained by Nikolayev et al. [34], only three out of 12 fragments were classified as S/MARs (Table 4). Two were classified as sequences presumably associated with the inner matrix (rsDNA), another two were classified as sequences presumably associated with the nuclear lamina (nlDNA), and one was classified as a sequence presumably associated with the synaptonemal complex (sc2DNA). The remaining four fragments were not classified at all. Thus we obtained the heterogeneous classification. However, S/MAR was the most representative class and the sensitivity in this case was estimated to be 25% ($3/12 \times 100\%$). As was mentioned above, different methods of isolation can yield different nuclear matrix contents. This could lead to a heterogeneous classification of sequences as in the case of the hsDNA sample. The other reason can be attributed to sequence length. A ChrClass query sequence should be longer than 300 bp (the minimal S/MAR length), but some fragments from the hsDNA sample were shorter. In order to solve this problem each short hsDNA fragment was tandemly duplicated. However, such a heuristic procedure may cause serious problems in the prediction, and, in general, the analysis of data from Nikolaev et al. [34] is extremely complicated due to short sequence lengths. For the tobacco and Chinese hamster sequences, the sensitivity

Table 3
The posterior classification of the samples of sequences associated with various elements of the nuclear matrix, the 5′flDNA sample and the randDNA sample

	%TRUE	S/MARs	rsDNA	sc1DNA	nlDNA	5′fl DNA	sc2DNA	randDNA
S/MARs	92.6	25	0	0	1	1	0	0
rsDNA	100.0	0	16	0	0	0	0	0
sc1DNA	94.11	0	0	16	0	1	0	0
nlDNA	80.76	0	0	3	21	1	1	0
5′flDNA	82.60	4	0	0	0	19	0	0
sc2DNA	89.47	0	1	0	1	0	17	0
randDNA	100.0	0	0	0	0	0	0	116
Total	91.36	29	17	19	23	22	18	116

Table 4
Classification of the test samples by the ChrClass program and MRS criterion

Sample ident.	MRS criterion	Classification of queries obtained with ChrClass program					% of truly predicted
		rsDNA	nlDNA	scDNA	Unclassified	S/MARs (true classification)	
hsDNA[a]		Z35288 Z54222	Z35280 Z35283	Z35282	Z35290 Z35279 Z54220 Z54223	Z35291 Z54221 Z54224	25
tbDNA[b]	AF065880:278[d];324 AF065884:352;398		AF065886			AF065877-AF065885; AF065887-AF065888	90
chDNA[c]						X96546-X96549	100
Average sensitivity (%)							72

[a]Twelve DNA fragments from human chromosome 19 associated with the nuclear matrix.
[b]Twelve chromosomal DNA fragments associated with the nuclear matrix of tobacco.
[c]Four chromosomal DNA fragments associated with the nuclear matrix of hamster.
[d]First number corresponds to the AWWRTAANNWWGNNNC position of the MRS consensus; second number, after ';', corresponds to the AATAAYA position.

was 90% and 100%, respectively (Table 4). The average ChrClass sensitivity for all test samples was 72% ((25%+90%+100%)/3). A proportion of the predicted matrix-associated regions in the random sequences was about 3% at the nucleotide level. Because S/MARs should be absent or very rare in random sequences, the rate-predicted S/MARs in a random sequence could be a measure of false positives. A proportion of predicted matrix-associated regions was on the same level for random sequences generated based on the nucleotide frequencies of the overall training set (training samples 1–5; $F(A) = 0.3$, $F(T) = 0.3$, $F(C) = 0.2$, $F(G) = 0.2$).

The ChrClass prediction quality, CCPQ (normalized to fall between 0 and 1), reflects the quality of predicted S/MARs. CCPQ was constructed based on the training S/MAR sample (all discriminant function values, $H_{S/MAR}$, were calculated for the sequences from this sample). The obtained value $h_{S/MAR}$ for a query sequence is compared with the $H_{S/MAR}$ array. CCPQ was tested on a sample of 12 tobacco S/MARs (Table 4), where the values of binding strength (the affinity of S/MARs to the nuclear matrix) are obtained for each tobacco S/MAR experimentally by Michalowski et al. [35]. We analyzed the correlation between the binding strength and CCPQ values (data not shown). The linear correlation coefficient r of 0.66 ($P < 0.05$) suggests that the CCPQ could reflect S/MAR affinity to the nuclear matrix, although further investigation is necessary to confirm this finding.

Only two MRSs were found in the test sample 1, clones ps202-1 (AF065880, 635 bp) and ps211-1 (AF065884, 685 bp) in the tbDNA. It is not surprising that sequence length is also crucial for MRS recognition, since the MRS was defined as the region where 16 and 8 bp consensus sequences are < 200 bp apart. That means we should test for their presence on the set of overlapping windows < 200 bp apart. Obviously for short sequences this number is much smaller than for cosmid-sized ones.

3.7. Analysis of relatively long (cosmid-sized) sequences to investigate the prediction power of ChrClass, MAR-Finder and MRS

The best way to test the capability of different approaches for S/MAR prediction is to carry out a analysis using relatively long sequences.

3.7.1. S/MARs in the β-globin locus

The human β-globin locus (HUMHBB, 7 308 bp) has been used for MAR-Finder testing [33,47], since the location of S/MARs was experimentally determined for this locus [48]. However, the conditions for the experimental detection of S/MARs by Jarman and Higgs [48] were much more stringent than conditions that are usually used for S/MAR detection [49]. S/MARs were found in K562 cells, which express only ε- and fetal γ-globins, and thus some S/MARs can be detached from the nuclear matrix in these cells. Seven S/MARs were determined experimentally in the HUMHBB sequence. However, the precise location was described for only four S/MARs, whereas the locations of the remaining three S/MARs were described with respect to the globin genes (Table 5). Even in the face of this problem, the experimental and computer predictions appear to be consistent. Six out of seven experimentally determined S/MARs were predicted correctly by MAR-Finder (results from Walter et al. [47]) and five were predicted correctly by ChrClass (Table 5). For the MRS motif we recalculated the results of van Drunen et al. [18]. They found six MRSs in the β-globin cluster, all of which map to biochemically identified S/MARs. No MRS was found only for the S/MAR sequence from the first intron of the β-globin gene [18]. MRS locations recalculated here are listed in Table 5. The probability of observing an overlap for the four precisely known S/MARs at random, calculated using a Monte Carlo technique, is very low ($P = 0.02$ for

MAR-Finder and $P = 0.05$ for ChrClass). We repeated the MAR-Finder analysis of the HUMHBB sequence in order to obtain the average strength values for the MAR-Finder prediction by Walter et al. [47]. However, only three real S/MARs were predicted by MAR-Finder using a threshold value of '0.5' and a default set of other parameters (Table 5). Although the four real S/MARs in the human β-globin locus cannot be predicted, even with a threshold value of '0.5', the average strength value 0.6–0.75 was considered to yield reasonable results by Singh et al. [33]. Thus, the choice of threshold value for the MAR-Finder prediction is somewhat ambiguous. Results of the β-globin locus suggest that threshold values less than '0.5' also can be recommended for some sequences. Analysis with the ChrClass program reveals that the relative location (relative to genes constituting the β-globin locus) of the predicted S/MARs is conserved among the β-globin loci from different mammalian species (galago and rabbit [50]), suggesting that some experimentally undiscovered S/MARs may exist there (Fig. 2). The sequence homology between human–rabbit, human–galago, galago–rabbit β-globin loci is limited mostly to coding regions, with some minor extensions in flanking regions [50]. The relative locations of S/MARs revealed with ChrClass almost coincide with the MRS distribution within these orthologous regions in different mammalian species, which was reported by van Drunen et al. [18] (Fig. 3 in the original paper). The distribution looks like each developmentally regulated gene in the β-globin locus as well as the LCR sequence is positioned in an individual 'loop' or 'functional unit', attached to the nuclear matrix from both sides (plant genes are organized in the similar manner, see below). One speculation is that this organization supports the interaction between the LCR and promoter regions of distal genes, which provides in turn the developmentally regulated gene expression. The interaction between HSs (the DNase hypersensitive sites) in LCR and promoter regions has been postulated by many authors to be crucial for correct tissue- and stage-specific gene expression in the β-globin locus. Experimental investigations are needed to address this hypothesis.

Table 5
Analysis of S/MARs in the human β-globin locus (HUMHBB)

Experimentally revealed S/MARs	MAR-Finder prediction by Walter et al. [47]	MAR-Finder prediction (average strength)	ChrClass prediction (prediction score)	Description in the Feature Table	MRS prediction by van Drunen et al. [18]
1–2 500	–	–	–	–	–
	a: 7 500–9 700	–	8 700–9 000 (0.39)	Alu 8 019–8 314	–
			10 000–10 600 (0.08)		
	b: 11 700–13 900	–	11 300–11 900 (0.20)	L1 12 912–13 066	–
				HS1 12 752–13 769	
14 000–16 000	c: 13 900–16 500	–	15 200–16 100 (0.14)	L1 14 836–15 701	17 653[a]; 17 586
			23 300–24 600 (0.43)	L1 23 118–31 136	
			24 900–25 800 (0.25)	L1 23 118–31 136	
			26 700–27 700 (0.53)	L1 23 118–31 136	
			28 000–28 500 (0.45)	L1 23 118–31 136	
			31 700–32 700 (0.20)	Alu 32 407–32 711	
				34 531–35 982 G-γ-globin	
				39 467–40 898 A-γ-globin	41 107; 40 955
			42 500–42 800 (0.24)		42 232, 42 231 (overlap)
			44 200–44 500 (0.23)		
In β-globin	d: 46 600–49 300	–	46 300–46 800 (0.42)	β-globin pseudogene	
				45 710–47 124	
				intron 2 46 146–46 996	
Somewhere between pseudogene and δ-globin*	e: 52 700–54 500	53 900–54 500 (0.51)	48 000–48 400 (0.41)	δ-globin gene 54 740–56 389	47 833; 47 958
				intron 2 55 233–56 130	
	f: 54 500–56 300	55 500–55 800 (0.51)	55 600–56 000 (0.33)		
Somewhere between δ- and β-globin*	g: 58 900–61 900	60 300–60 700 (0.64)	58 100–58 500 (0.40)		–
62 632–63 481	h: 61 900–64 100	–	–	β-globin gene 62 137–63 742	–
				intron 2 62 632–63481	
65 610–66 100	i: 64 100–67 400	66 000–66 600 (0.74)	64 900–66 600 (0.26)	Alu 65 503–65 757	65 111; 65 138
				Alu 66 761–67 070	65 947; 65 946 (overlap)

HS1 is the DNase hypersensitive site 1. A threshold value '0.5' and a default set of parameters were used for the second MAR-Finder prediction. Asterisk indicates that for a marked S/MAR a precise location was not described [48]. The ChrClass program predicted the location of the nuclear lamina-associated site between positions 9000 and 9700.

[a]First number corresponds to the AWWRTAANNWWGNNNC position of the MRS consensus; second number, after ';', corresponds to the AATAAYA position.

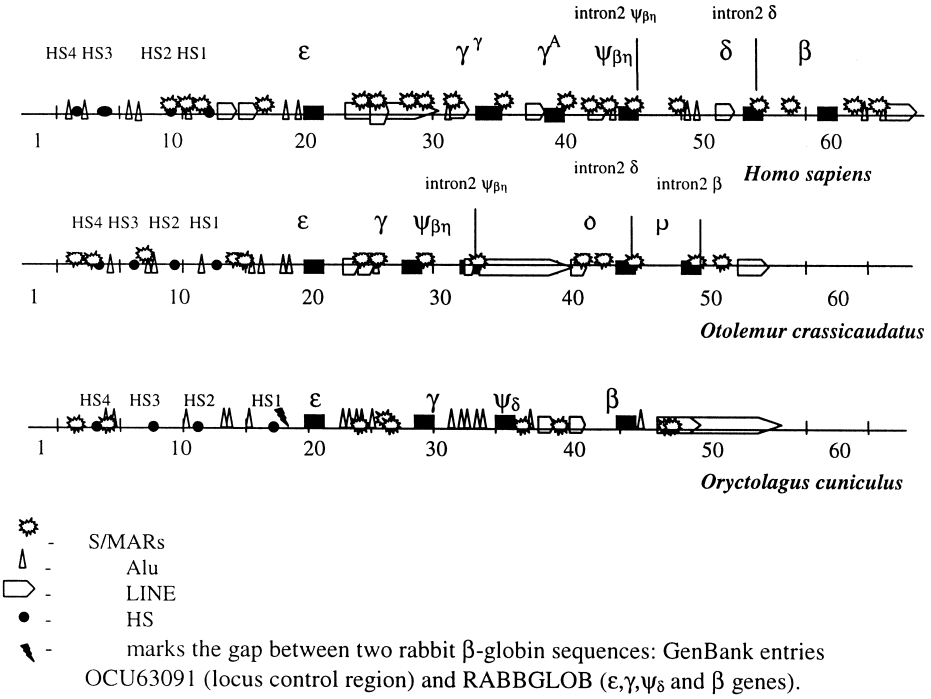

Fig. 2. Distribution of S/MARs predicted by the ChrClass program in the β-globin loci from three different mammalian species. All genes in the loci are positioned in reference to the human ε-globin gene. HS1, 2, 3 and 4 are the DNase hypersensitive sites.

3.7.2. S/MARs in the colinear Sh2/A1 homologous regions of rice and sorghum

About 50 and 30 kb of sorghum and rice DNA respectively, which contain A1/Sh2 homologous regions, were screened for the location of S/MARs by the in vitro binding method and seven S/MARs in sorghum and four in rice were identified [19]. The 30 kb region of rice contains three functional genes: Sn1, X (putative transcription factor) and A1 (for more details see Fig. 1 in [19]). All these genes are flanked by S/MARs on the 5′ as well as on the 3′ ends (except for 3′ end of A1) in rice. The presence, order and direction of transcription of Sh2, X and A1 are the same in sorghum as in rice and in addition a direct duplication of A1, A1a, is localized 10 kb downstream of A1 [19]. All four rice S/MARs have their orthologous S/MAR counterparts in sorghum. There are three additional S/MARs in sorghum: two flanking the 3′ end of A1, A1a and one weakly binding fragment, overlapping the strongly binding 5′-Sn1 flanking S/MAR [19]. However, hybridization and sequence analysis of the two colinear genomic regions indicated that the sequence homology between them is limited to the coding regions. This 'evolutionary conservation', i.e. the similar location of S/MARs in different species relative to the orthologous genes, was confirmed in this case by in vitro binding assay [19]. We also analyzed A1/Sh2 homologous regions with ChrClass, MAR-Finder programs and MRS criterion. The results are presented in Table 6. Four and three S/MARs were predicted by ChrClass and MAR-Finder for four experimentally identified S/MARs in rice; five and two for five S/MARs in sorghum. As one can see the results are partially overlapping, but each S/MAR prediction tool has its own peculiarities. With regard to the MRS criterion, it was supposed that the presence of MRS in the genomic region is a not necessary but sufficient condition to find an S/MAR in close proximity. In general this is true. For example, 13 MRSs were found in the rice Sh2/A1 region, and all of them were inside the S/MAR sequences. However, seven MRSs were found in S/MARs which were predicted by the ChrClass program only (Nos. 7–13, Table 6), two MRSs in S/MARs which were predicted by both ChrClass and MAR-Finder (Nos. 5 and 6) and four MRSs in S/MARs which were predicted by both programs and by in vitro MAR-binding method (Nos. 1–4). These results reveal a very interesting problem. Three different computer tools predict an S/MAR sequence where the experimental method does not (the case of Nos. 5 and 6). Possible reasons may include dynamically regulated DNA–matrix association, which could not be revealed on the experimental level. To account for all possible tissue- and stage-specific associations between the nuclear matrix and S/MARs, the in vitro binding assay should be done for various stages of the cell cycle and different tissues. The Sh2/A1 region of sorghum was also tested, but in this case only two MRSs were found and only one of those had S/MARs in close proximity. Thus for the detection of S/MAR sequences the presence of MRS is not always a sufficient condition. In addition, ChrClass usually identifies all S/MARs that are found experimentally, but also reveals many S/MARs experimentally 'silent'. In some

Table 6
Analysis of S/MARs in the colinear Sh2/A1 homologous regions of rice and sorghum

Experimentally revealed S/MARs	ChrClass prediction (prediction score)	MAR-Finder prediction by Singh et al. [33] (average strenght)	Description in the Feature Table	MRS prediction by van Drunen et al. [18]
Rice				
1–1 173	1–900 (0.05)	–	1 996–6 023 putative Sh2 subunit	–
	1 200–1 600 (0.22)		2 203–2 594 intron 1	
	2 200–2 500 (0.49)		3 323–3 493, exon 3	
	3 400–3 700 (0.2)		3 591–3 680, exon 4	
5 400–7 425	6 400–7 200 (0.33)	7 200–7 400 (0.6)	6 479–6 651, 7 042–7 191 transposons	–
			9 722–10 006 ORF 3	
	9 700–10 600 (0.36)	10 800–11 200 (0.85)	10 191–10 328 ORF 4	
			10 543–10 590 tandem repeat AT	
	13 700–14 400 (0.13)		15 665–16 105 snapback region Snabo-1	
	15 300–16500 (0.33)			
17 300–18 460	17 400–2 1300 (0.06)	19 500–20 400 (0.9)	21 151–21 345 transposon	N1 18 315;18 364
20 020–23 080	21 600–22 200 (0.48)	20 700–21 600 (0.8)	21 644–21 977 transposon	N2 20 104;19 958
	22 800–23 200 (0.25)		22 843–23 225 snapback region Snabo-2	N3 20 667;20 711
				N4 20 864;20 711
				N5 24 213;24 263
				N6 24 414;24 263
	23 500–26 800 (0.10)	24 200–24 600 (0.6)	5′flank NADPH-dependent reductase A1	N7 25 197;25 108
				N8 25 464;25 678
	28 500–28 900 (0.41)		26 910–28 526 CDS 3′flank Adh1	N9 25 679;25 729
				N10 25 679;25 678
				N11 25 907;25 729
				N12 26 592;26 625
				N13 26 644;26628
Sorghum				
1–1 495	0–700 (0.18)	–	2 362–6 325 putative Sh2 subunit	–
	2 900–3 400 (0.26)		3 046–3 168, 3 433–3 603 exon 2, 3	
	3 700–4 200 (0.45)		3 836–3 925, 4 017–4 103, 4 187–4 242 exon 4, 5, 6	
7 050–9 710	7 200–7 800 (33)	10 000–10 300 (0.7)	7 118–7 458, 7 464–7 763 transposons	13 273;13 148
	9 600–10 000 (0.21)	11 900–12 300 (0.7)	10 465–11 142 S-1 transposon	
	15 100–15 900 (0.35)			
22 400–24 680	23 000–23 300 (0.30)	–	24 370–24 715 transposon	–
	23 600–23 900 (0.50)			
	24 200–24 700 (0.53)			
	26 900–27 900 (0.21)		5′flank NADPH-dependent reductase A1-a	
			28 349–29 769 CDS	
32 500–33 660	31 900–32 400 (0.05)		31 925–36 403 solo LTR	33 740;33 839
	33 400–33 900 (0.32)	33 860–34 260 (0.8)		
	36 600–37 300 (0.28)			
			39 521–41 042 CDS NADPH-dependent reductase A1-b	
41 600–42 280	–	–		–

cases the results may be attributed to 'noise', but in some others this 'noise' could consist of real S/MARs not revealed experimentally for various reasons (as was mentioned above in the discussion about MRS Nos. 5 and 6). In principle, in the case of β-globin genes when only stage-specific expression should be observed [48] and some real S/MARs could not be associated with the nuclear matrix, the obtained S/MAR 'overprediction' has a biological meaning. Computer predictions might reveal 'silent' S/MARs that function in a stage- or tissue-specific manner. This problem needs further experimental analysis.

3.8. Recommendations on the use of S/MAR prediction programs

All listed examples support that the proportion of missed S/MARs is lower for ChrClass, whereas the proportion of wrong S/MARs is lower for MAR-Finder and MRS criterion. That is, ChrClass has a tendency to overpredict the presence of S/MARs, while the MRS criterion and MAR-Finder have a tendency to underpredict this, depending on the testing sequence (MAR-Finder usually reveals about 50–70% of experimentally confirmed S/MARs). It is impossible to resolve which is better,

Table 7
ChrClass classification of telomeric and alphoid DNA

GenBank identifier, subunit length (bp)	Chromosome	Result of classification with ChrClass
HSAJ1561, 1427	7	S/MAR
HSAJ1558, 2482	7	rsDNA
HSAJ1560, 2010	7	scDNA
HSAJ2432, 535	11	scDNA
HUMSATLS, 718	17	scDNA
HSAJ2431, 405	17	rsDNA
D29750, 1868	21	scDNA
$(TTAGGG)_{500}$		S/MAR

over- or underprediction. In each case, the choice should depend on the problem under investigation. We recommend analyzing short sequences with the ChrClass program, since for other tools the proportion of missed S/MARs appears to negatively correlate with sequence length (MAR-Finder and MRS frequently miss S/MARs in short sequences). The higher overall A+T content ($>70\%$) increases the proportion of wrong S/MARs in ChrClass output (results not shown) and in such cases the use of Mar-Finder would be better. Also, we propose that the presence of S/MARs predicted by *any tool* may be verified by analyzing the orthologous regions between different species with the *same tool*, if such data are available. The cases of simultaneous S/MAR presence in such regions should support the reliability of prediction. Finally, using the ChrClass, MRS criterion and MAR-Finder programs together may help to obtain the more clearcut picture of S/MAR distribution in a query sequence.

3.9. Classification of telomeric and alphoid DNA

We also tested some interesting examples of experimentally observed matrix association in tandemly repeated DNA. Human alphoid DNA consists of a monomer (171 bp) tandemly repeated thousands of times at each centromere. These regions are configured into higher order structures, where several tandemly repeated monomers form a subunit (divergence of each repeat from the consensus sequence is about 15–20%). Each subunit is, in turn, repeated hundreds of times [51]. Here we consider some of the subunits of alphoid DNA from chromosomes 7, 17, 11, and 21 (GenBank, 1998). The result of the prediction with the ChrClass program suggests that different subunits of alphoid DNA may contain fragments associated with different chromosomal/nuclear substructures (Table 7). In the three subunits from chromosome 7 the ChrClass program identified three different kinds of sequences, that is presumably associated with the synaptonemal complex, presumably associated with the core of rosette-like structures, and presumably S/MARs. Two subunits from chromosome 17 might also be associated with the rosette-like structure cores and the synaptonemal complex. Two subunits from chromosomes 11 and 21 contain fragments presumably associated with the synaptonemal complex. Telomeric repeats have been attributed to S/MARs. We will discuss these results below.

4. Discussion

Computer tools for the accurate prediction of S/MARs and other sites associated with the nuclear matrix are very important for molecular biology, since experimental mapping of such sites is very slow compared to sequencing of eukaryotic genomes. However, the prediction of S/MARs is a necessary step for the successful functional mapping of nucleotide sequences, since these sites can bring genes into association with the nuclear matrix significantly changing their transcription level, and thus, marking transcriptionally active regions (reviewed in [17]). But the problem of S/MAR prediction is complicated for two reasons. First, the sequence divergence between different S/MARs is very high (such that they cannot be aligned). Second, operationally defined nuclear matrix includes various components for which the correspondence of cytologically determined nuclear substructures to biochemically obtained matrix-associated regions is largely unknown.

We have carried out a comparative analysis of the nucleotide sequence of fragments of chromosomal DNA associated with various elements of the nuclear matrix in animal and plant chromosomes (in both somatic and meiotic cells). The fragments of chromosomal DNA extracted from either chromosomal or nuclear substructures (nuclear lamina, cores of rosette-like structures, synaptonemal complex) are much more similar to each other than to the chromosomal DNA fragments obtained by specific DNA–protein interaction assays (S/MARs). The analysis of distances among the samples from cytological structures (nlDNA, rsDNA, scDNA) suggested some similarity in the mechanisms responsible for the formation of chromosome loop structures in somatic and meiotic cells. Interestingly, four out of 25 fragments of nlDNA, which participates in spatial chromosome organization during the interphase, were classified as fragments extracted from the synaptonemal complex, which participates in structural chromosome organization at stage I of meiosis prophase (Table 3). Previous cytological studies have suggested that certain chromosomal DNA sites may participate in the

association of interphase chromosomes either with the nuclear lamina in somatic cells or with the synaptonemal complex in prophase I of meiosis.

Interestingly, S/MAR sequences are closer to the 5′ flanking regions of eukaryotic protein-coding genes and to DNA fragments from the synaptonemal complex than to other fragments of chromosomal DNA extracted from 'cytological' structures. However, the reasons for such differences are not clear. The nuclear pores–lamina complex is known as a component of the nuclear matrix/scaffold [22]. Thus, chromosomal DNA fragments extracted from the nuclear lamina should be classified as S/MARs. However, the S/MAR sample does not overlap with the nlDNA sample (Table 3). This result is consistent with several previous observations [52]. Unfortunately, our data do not clarify if there is any correspondence between S/MARs and cytologically observed morphological elements of the nuclear matrix. Hancock [53] recently discussed the possibility that S/MARs actually bind to DNA binding proteins or multiprotein complexes that are incorporated into the nuclear matrix during preparation. This point of view does not change the existence and properties of S/MARs as a sequence family but imply the complete revision of supposed S/MAR functions. Thus, the problem of a formal definition for S/MARs needs further analysis [54].

The association of telomeric repeats with the matrix of the interphase nucleus was shown by de Lange [14] and Luderus et al. [15]. Approximately every 1000 bp of the telomeric repeats is associated with the inner matrix [15]. However, the context characteristics of S/MARs are absent in this simple hexameric repeat $(TTTAGG)_n$. Since the association of telomeric repeats with the nuclear matrix has already been proven, we have included in our analysis some additional context characteristics (triplet frequencies). After combining triplet frequencies and MDS, we were able to achieve a more accurate posterior classification for S/MARs, rsDNA, nlDNA, sc1DNA, sc2DNA and 5′flDNA (the sensitivity achieved was 94, 100, 94, 80, 83, 89, and 100%; for MDS it was 75, 56, 36, 29, 33, and 75% respectively, Table 3). This result suggests that MDS alone is not sufficient for an accurate LDA classification of fragments associated with various elements of the nuclear matrix.

Previous experimental data show that the 1.7 kb human alphoid DNA subunit from chromosome 16 and the 1.9 kb subunit from chromosome 1 [13] contain S/MARs. Interestingly, different subunits of alphoid DNA might be associated with different chromosome/nuclear substructures (Table 7). This result suggests that different tandemly repeated subunits might interact with various elements of the nuclear matrix at a certain stage of the cell cycle. However, the predicted association with different chromosome/nuclear substructures should be interpreted with caution since this prediction has no confirmation using control sets. The former is impossible today because of the lack of large samples of chromosomal fragments associated with various elements of nuclear matrix. Thus, the prediction of these fragments in a query sequence can only point to some unusual and presumably interesting protein-binding regions.

Some overlap between the predictions of all MAR prediction tools has been found. Using the ChrClass, MRS criterion and MAR-Finder programs together may help to obtain a more clearcut picture of S/MAR distribution in a query sequence. The real predictive power and functional importance of all employed context characteristics in MAR-Finder and ChrClass (described above) are not so obvious. Thus the problem of S/MAR prediction requires further analysis. The accumulation of new experimentally confirmed S/MARs will help to improve existing approaches and to develop better models for the prediction of sites associated with various elements of the nuclear matrix. The significant correlation between the experimentally determined binding strength and the S/MAR prediction quality measure CCPQ needs further investigation.

Acknowledgements

The authors wish to thank M.S. Gelfand, N.A. Kolchanov, F.A. Kodrashov and D.N. Nguyen for help and stimulating discussions and E.Y. Nagornikh for help with translation of the manuscript. This work was partly supported by the Russian Fund of Fundamental Investigation (Grant 99-04-49535) and by research grants from NIH and NASA to M. Nei.

Appendix. Description of samples

A1. Training samples

1. S/MARs:
   Thirteen S/MAR sequences were obtained from GenBank: accession numbers X54282 (complex of human β-globin genes), X98408 (chicken lysozyme gene), M62716 (5′ flanking region of human CSP-B gene), M83137 (human β-interferon gene), X06654 (intron of Chinese hamster DHFR gene), L23999 (fourth intron of the human DNA topoisomerase I gene), L23998 (second intron of the human DNA topoisomerase I gene), X60225 (*Drosophila* histone genes), X07690 (first intron of the human HPRT gene), M77843 (pea plastocyanin gene), U29136 (corn Adh1 gene), U67919 (anonymous MAR of tobacco), X67164 (petunia MAR near the T-DNA integration site);
   Four S/MAR sequences were obtained from EMBL: accession numbers U71191–U71193 (S/MARs in the distal mouse chromosome 7 imprinted domain), U71190 (S/MAR near mouse rpl23 gene);
   Seven S/MAR sequences were obtained from the

S/MARt database (URL: transfac.gbf.de/SMARtDB/index.html): accession numbers SM0000061, SM0000062 (two anonymous MARs of tobacco), SM0000063, SM0000064 (5′ and 3′ flanking regions of rice ADP-glucose pyrophosphorylase large subunit gene), SM0000065 (3′ flanking region of rice X gene), SM0000066 (5′ flanking MAR of rice A1 gene), SM0000069 (5′ flanking region of the *Sorghum bicolor* dihydroflavonol-4-reductase gene);
Three S/MAR sequences were obtained from original papers: 3′ MAR of the human apolipoprotein B gene [55], MAR from the J-C intron of the mouse Ig κ immunoglobulin gene [56], 3′ MAR of chicken α-globin gene [17].

2. rsDNA fragments were extracted from cores of rosette-like structures of interphase mouse chromosomes, clones pChrM1–pChrM16 [30].
3. sc1DNA fragments were extracted from synaptonemal complex of Chinese hamster [31]; GenBank accession numbers Z32801–Z32803, Z32798, Z32808, Z32807, Z32805, Z32810, Z32811, Z86071, Z86085, Z86086, Z32797, U09289, U09301, MASC4L, MASC9L.
4. sc2DNA fragments were extracted from synaptonemal complex of rat [32] (GenBank accession numbers X61772–X51786, X61789–X61792).
5. nlDNA fragments were extracted from nuclear lamina of mouse oocytes [29] (GenBank accession numbers X55461–X55472, X55474, X55475, X55477–X55485, X55487, X55488).
6. A dataset of 5′flDNA fragments (5′ flanking regions of tissue-specific eukaryotic genes 1–3 kb long) was provided by N.V. Milshina.

A2. Test samples

1. Sequences annotated as S/MARs in GenBank, 1998:
   - hsDNA: 12 chromosomal DNA fragments associated with the nuclear matrix from human chromosome 19 [34] (accession numbers Z35288, Z35290, Z35291, Z35279, Z35280, Z35282, Z35283, Z54220–Z54224);
   - tbDNA: 12 chromosomal DNA fragments associated with the nuclear matrix of tobacco [35] (accession numbers AF065877–AF065888);
   - chDNA: four chromosomal DNA fragments associated with the nuclear matrix of hamster [21] (accession numbers X96546–X96549).
2. Relatively long (cosmid-sized) sequences to investigate the ChrClass, MAR-Finder and MRS prediction power simultaneously:
   - ptDNA: two DNA fragments (30 035 bp, U70541 and 42 447 bp, AF010283) of the colinear Sh2/A1 homologous regions of rice and sorghum respectively [19];
   - glDNA: three clusters of β-globin genes: human (HUMHBB, 73 308 bp); galago (OCU6090, 57 113 bp); rabbit (RABBGLOB, 44 594 bp cluster of β-globin genes and OCU63091, 16 609 bp LCR were considered together).
3. Human alphoid DNA (accession numbers AJ001516, HSAJ1558, HSAJ1560, HSAJ1559, HSAJ2432, HUMSATLS, HSAJ2431, D29750); an 'ideal' telomeric motif of eukaryotic chromosomes $(TTAGGG)_{500}$.

References

[1] J.W. Bodnar, J. Theor. Biol. 132 (1988) 479–507.
[2] S.M. Gasser, B.B. Amati, M.E. Cardenas, J.F.X. Hofmann, Int. Rev. Cytol. 119 (1989) 57–96.
[3] U.K. Laemmli, E. Kas, L. Poljak, Y. Adachi, Curr. Opin. Genet. Dev. 2 (1992) 275–285.
[4] P. Breyne, M. Van Montagu, G. Gheysen, Transgenic Res. 3 (1994) 195–202.
[5] H. Buhrmester, J.P. von Kries, W.H. Stratling, Biochemistry 34 (1995) 4108–4117.
[6] L.A. Dickinson, T. Kohwi-Shigematsu, Mol. Cell. Biol. 15 (1995) 456–465.
[7] L.A. Dickinson, C.D. Dickinson, T. Kohwi-Shigematsu, J. Biol. Chem. 272 (1997) 11463–11470.
[8] T. Kohwi-Shigematsu, K. Maas, J. Bode, Biochemistry 36 (1997) 12005–12010.
[9] P.N. Cockerill, W.T. Garrard, Cell 44 (1986) 273–282.
[10] J. Mirkovitch, M-E. Mirault, U.K. Laemmli, Cell 39 (1984) 223–232.
[11] P.A. Dijkwel, J.L. Hamlin, in: R. Berezney, W.L. Kwang (Eds.), Structural and Functional Organization of the Nuclear Matrix, Academic Press, New York, 1995, pp. 455–484.
[12] J.W. Bodnar, M.K. Bradley, J. Theor. Biol. 183 (1996) 1–7.
[13] P.L. Strissel, R. Espinosa III, J.D. Rowley, H. Swift, Chromosoma 105 (1996) 122–133.
[14] T. De Lange, EMBO J. 11 (1992) 717–724.
[15] M.E. Luderus, B. van Steensel, L. Chong, O.C. Sibon, F.F. Cremers, T. de Lange, J. Cell Biol. 135 (1996) 867–881.
[16] R. Strick, U.K. Laemmli, Cell 83 (1995) 1137–1148.
[17] T. Boulikas, Int. Rev. Cytol. 162A (1995) 279–388.
[18] C.M. van Drunen, R.G.A.B. Sewalt, R.W. Oosterling, P.J. Weisbeek, S.C.M. Smeekens, R. van Drial, Nucleic Acids Res. 27 (1999) 2924–2930.
[19] Z. Avramova, A. Tikhonov, M. Chen, J.L. Bennetzen, Nucleic Acids Res. 26 (1998) 761–767.
[20] C. Mielke, Y. Kohwi, T. Kohwi-Shigematsu, J. Bode, Biochemistry 35 (1990) 2239–2252.
[21] M.A. Fernandez, B. Baron, M. Prigent, F. Toledo, G. Buttin, M. Debatisse, J. Cell Biochem. 67 (1997) 541–551.
[22] R. Berezney, D.S. Coffey, J. Cell Biol. 73 (1977) 616–637.
[23] S.H. Kaufman, D.S. Coffey, J.H. Shaper, Exp. Cell Res. 132 (1981) 105–123.
[24] J.M. Craig, S. Boyle, P. Perry, W.A. Bickmore, J. Cell Sci. 110 (1997) 2673–2682.
[25] S. Li, M.L. Meistrich, W.A. Brock, T.C. Hsu, M.T. Kuo, Exp. Cell Res. 144 (1983) 63–72.
[26] L. Sudhakar, M.R. Rao, J. Biol. Chem. 265 (1990) 22526–22532.
[27] L. Gil-Alberdi, J. del Mazo, Cytogenet. Cell. Genet. 59 (1992) 1–5.
[28] R.L. Meuwissen, H.H. Offenberg, A.J. Dietrich, A. Riesewijk, M. van Iersel, C. Heyting, EMBO J. 11 (1992) 5091–5100.
[29] R. Christova, I. Bach, Z. Galcheva-Gargova, DNA Cell Biol. 11 (1992) 627–636.
[30] M.V. Glazkov, A.B. Poltaraus, I.A. Lebedeva, Genetika (Moscow) 30 (1994) 1146–1154.
[31] O.I. Karpova, M.V. Penkina, S.Y. Dadashev, N.V. Milshina, X.

Ernandes, I.V. Radchenko, U.F. Bogdanov, Mol. Biol. (Moscow) 29 (1995) 512–521.
[32] R.E. Pearlman, N. Tsao, P.B. Moens, Genetics 130 (1992) 865–872.
[33] G.B. Singh, J.A. Kramer, S.A. Krawetz, Nucleic Acids Res. 25 (1997) 1419–1425.
[34] L.G. Nikolaev, T. Tsevegiyn, S.B. Akopov, L.K. Ashworth, E.D. Sverdlov, Nucleic Acids Res. 24 (1996) 1330–1336.
[35] S.M. Michalowski, G.C. Allen, G.E. Hall Jr., W.F. Thompson, S. Spiker, Biochemistry 38 (1999) 12795–12804.
[36] H. Htun, J.E. Dahlberg, Science 241 (1988) 1791–1796.
[37] P. Vogt, Hum. Genet. 84 (1990) 301–336.
[38] J. Bode, Y. Kohwi, L. Dickinson, T. Joh, D. Klehr, C. Mielke, T. Kohwi-Shigematsu, Science 255 (1992) 195–197.
[39] P.T. McNamara, A. Bolshoy, E.N. Trifonov, R.E. Harrington, J. Biomol. Struct. Dyn. 8 (1990) 529–538.
[40] C. Benham, T. Kohwi-Shigematsu, J. Bode, J. Mol. Biol. 274 (1997) 181–196.
[41] B.B. Amati, S.M. Gasser, Cell 54 (1988) 967–978.
[42] J.R. Spitzner, I.K. Chung, M.T. Muller, Nucleic Acids Res. 18 (1990) 1–11.
[43] R.L. Shoeman, P. Traub, J. Biol. Chem. 265 (1990) 9055–9061.
[44] Y. Moroi, M.L. Hartman, P.K. Nakane, E.M. Tan, J. Cell Biol. 90 (1981) 254–259.
[45] J.M. Claverie, I. Sauvaget, L. Bougueleret, Methods Enzymol. 183 (1990) 237–251.
[46] P.N. Cockerill, W.T. Garrard, FEBS Lett. 204 (1986) 5–7.
[47] W.R. Walter, G.B. Singh, S.A. Krawetz, Biochem. Biophys. Res. Commun. 242 (1998) 419–422.
[48] A.P. Jarman, D.R. Higgs, EMBO J. 7 (1988) 3337–3344.
[49] J. Bode, K. Maas, Biochemistry 27 (1988) 4706–4711.
[50] J.L. Slightom, J.H. Bock, D.A. Tagle, D.L. Gumucio, M. Goodman, N. Stojanovic, J. Jackson, W. Miller, R. Hardison, Genomics 39 (1997) 90–94.
[51] H.F. Willard, J.S. Waye, Trends Genet. 3 (1987) 192–198.
[52] W.F. Marshall, A.F. Dernburg, B. Harmon, D. Agard, J.W. Sedat, Mol. Biol. Cell 7 (1996) 825–842.
[53] R. Hancock, Chromosoma 109 (2000) 219–225.
[54] I.B. Rogozin, G.V. Glazko, M.V. Glazkov, Briefings Bioinformatics 1 (2000) 33–44.
[55] B. Levy-Wilson, C. Fortier, J. Biol. Chem. 264 (1989) 21196–21204.
[56] C. Blasquez, A.O. Sperry, P.N. Cockerill, W.T. Garrard, Genome 31 (1989) 503–509.