Computer prediction of sites associated with various elements of the nuclear matrix

Igor B. Rogozin, Galina V. Glazko and Mikhail V. Glazkov

Date received (in revised form): 29th October 1999

Abstract

Attachment regions of the eukaryotic chromosomal DNA to the nuclear scaffold/matrix (S/MARs) participate in various important cellular processes. However, no obvious characteristics common for these nucleotide sequences have been revealed, except that S/MARs are non-coding sites containing putative regulatory elements and binding sites of DNA-topoisomerase II. Heterogeneity among S/MARs can be caused by a variety of biological factors. In this paper, the accuracy of two S/MARs prediction programs, MAR-Finder (Singh, Kramer and Krawetz, 1997) and ChrClass (Glazkov, Rogozin and Glazko, 1998) are compared and it is concluded that both programs can be recommended for analysis of eukaryotic genomes. However, results of their prediction should be interpreted with caution since estimation of prediction accuracy of both programs needs further analysis. Problems of S/MARs prediction are illustrated on several examples of human protein-coding genes, repeated elements and the beta-globin locus from different mammalian species. Results of our analysis suggest that the proportion of missed S/MARs is lower for ChrClass, whereas the proportion of wrong S/MARs is lower for MAR-Finder (a default set of parameters).

Keywords: nuclear matrix, chromatin, S/MAR, beta-globin, discriminant analysis, repetitive elements

SCAFFOLD/MATRIX ATTACHMENT REGIONS

The model of loop–domain organisation of eukaryotic chromosomes is now generally accepted.1–5. According to this model, topologically independent chromatin loops are attached to the nuclear matrix/scaffold. A number of proteins of the nuclear matrix/scaffold presumably participating in the loop organisation of chromosomes have been identified and some of their characteristics are known.2,5 Regions of attachment of the chromosomal DNA to the nuclear scaffold/matrix (S/MARs) were identified as those involved in very important cellular processes: transcription, replication and recombination.6 However, attempts to reveal some common characteristics of nucleotide sequences for various S/MARs have only demonstrated that they are usually non-coding regions highly saturated with putative protein-binding sites. Some S/MARs are AT-rich but not all AT-rich sequences are S/MARs. Even the presence of DNA-topoisomerase II consensus sequences or A and T-blocks2,6–8 cannot be considered to be highly valuable features of S/MARs. These features are described in more detail in the description of an MAR-Finder program below.9 An MAR/SAR recognition signature (MRS) which is common to a large group of matrix and scaffold attachment regions has been identified.10 The MRS is composed of two consensus sequences (AATAAYAA and AWWRTAANNWWGNNNC) within close proximity. Although all revealed MRSs are associated with a SAR, not all known S/MARs contain an MRS, suggesting that at least two classes exist, one containing an MRS, the other not.10 Thus this promising observation needs further analysis. Various classes of S/MARs have been defined.11 The diversity of...
S/MARs could be caused by a number of reasons. For example, S/MARs may be associated with different proteins and elements of the nuclear matrix. Different methods of extraction (2 M NaCl, lithium diiodosalicylate (LIS) detergent, electroelution) can yield various nuclear matrix contents. The nuclear matrix of somatic cells consists of various elements including (1) the nuclear pores–nuclear lamina complex, (2) residual nucleoli and (3) a fibrillar–granular network (inner matrix). Among these, the inner matrix is considered as the most variable. During prophase I in meiotic cells the synaptonemal complex is also an integral part of the nuclear matrix. Characteristics of the nucleotide sequences extracted as ‘DNA fragments, associated with the nuclear matrix’ evidently depend on the method of extraction, i.e., they are defined by those matrix elements with which these sequences were associated. All the S/MARs known in the chromosomes of somatic cells were revealed through the identification of specific interaction of DNA with the nuclear matrix; however, experimental conditions can be quite different. Meanwhile, several samples of the chromosomal DNA fragments have been extracted from different chromosomal/nuclear substructures, such as the nuclear lamina, cores of rosette-like structures and the synaptonemal complex. The S/MARs prediction is mainly discussed here; however, the other classes of sites associated with the nuclear matrix may possess very important functions and also should be taken into consideration.

### Computer tools for the accurate prediction of S/MARs are important for molecular biology

Two programs for prediction of S/MARs are described: MAR-Finder and ChrClass. These programs are based on different approaches: a statistical inferencing (MAR-Finder) and a multivariate linear discriminant analysis (ChrClass). These programs complement other S/MAR identification techniques, including stress-induced duplex destabilisation (SIDD) profiles and the MRS. The S/MAR database SMARt-DB provides an invaluable new resource and can be found at the GBF website (URL: transfac.gbf.de/SMARtDB/index.html).

### Programs for prediction of S/MARs

The MAR-Finder prediction of S/MARs is based on several S/MAR analysis rules:

- It is generally accepted that DNA replication is associated with the nuclear matrix. It has also been shown that S/MARs and the origins of replication share the ATTA, ATTTA and ATTTTA motifs.
- Some S/MARs contain TG-rich spans.
- Intrinsically curved DNA has been revealed at or near several S/MARs. Curved DNA is thought to play an important role in many molecular processes that involve the interaction of DNA and proteins, such as recombination, replication and transcription. Significant curvature can be expected for sequences with repeats of two motifs: AAAA(N)AA (N)AAAA and TTAAA.
- Kinked DNA has generally been associated with the presence of copies of DNA that are necessary for the accurate prediction of S/MARs and other sites associated with the nuclear matrix.
of the dinucleotide TG, CA or TA that are separated by 2–4 or 9–12 base pairs (bp).

- Topoisomerase II binding sites are concentrated at the sites of nuclear attachment. Both vertebrate and Drosophila topoisomerase II consensus sequences are used for prediction.

- Many S/MARs contain significant stretches of AT-rich sequences, but it has been suggested that the simple occurrence of isolated AT-rich regions is not sufficient to cause matrix association. Several such regularly spaced motifs may be required for matrix association.

Each of these properties of S/MARs has been considered in MAR-Finder. The task of detecting sites of matrix association has been treated by Singh et al.9 as a problem of hypothesis testing. Since matrix association is a property of a span of the sequence, a sliding window algorithm was considered appropriate for detecting S/MARs. This algorithm uses two parameters: the window width and a window slide distance. S/MAR finding progresses by analysing 1,000–2,000 bp regions of the sequence at a time. Successive windows obtained by sliding a distance of 100–200 bp are examined. The patterns detected in the region are used to associate a statistical significance to the region’s potential for being a matrix attachment site (average strength). S/MAR is defined by the presence of three or more successive windows with a high statistical significance. The MAR-Finder average strength (MFAS) values are normalised to fall between 0 and 1. This parameter specifies the minimum normalised potential above which there is a high likelihood that the underlying region has an S/MAR. The default has been set to 0.6, although a threshold value less than 0.6 was used for analysis of the β-globin locus.22 Values of 0.6 and 0.5 and a default set of other MAR-Finder parameters were used in our analysis.

**ChrClass**

The ChrClass program25,28 is based on comparative analysis of various context characteristics in the S/MARs sequences and the sequences extracted from the three nuclear matrix elements described below. Two subsets of context characteristics were joined together. The first subset included 13 motifs similar to the MAR-Finder analysis rules: the ‘kinked DNA’ motifs, short palindromes; poly(A)n, (T)n, poly(C)n and (G)n tracts (n ≥ 4); (AT)n and (GC)n tracts (n = 2, 3, >3); TC- (or CT-), AG- (or GA-), TG- (or GT-), AT- (or TA-) GC- (or CG-), AC- (or CA-) tracts (the tracts include these dinucleotides, arbitrarily combined; a tract length is more than six nucleotides). The second subset included 43 triplets with the least dispersion within training samples. Seven sets of sequences were analysed:

- 16 S/MAR sequences were taken from GenBank.
- 16 DNA fragments were extracted from the cores of rosette-like structures of interphase mice chromosomes.23
- 17 DNA fragments were extracted from the synaptonemal complex of Chinese hamster.24
- 18 DNA fragments were extracted from the synaptonemal complex of rat.22
- 25 DNA fragments were extracted from the nuclear lamina of mice oocytes.31
- 24 5′-flanking regions of tissue-specific eukaryotic genes of 1,000–3,000 bp.
- 116 DNA fragments (lengths of sequences were taken from the above
samples) were obtained by using a random generator assuming a uniform distribution of all the four nucleotides.

A multivariate linear discriminant analysis (LDA) technique has been used for simultaneous analysis of the differences among all seven samples. The discriminant function \( h_k \) is a linear combination of the input characteristics \( \{X_1, \ldots, X_n\} \):

\[
h_k = b_{k0} + b_{k1}X_1 + \ldots + b_{kn}X_n
\]

where \( b_{ki} \) (\( i = 1, n \)) are the discriminant function coefficients. Statistica 5.0 software (the ‘discriminant’ analysis module) was used for LDA. Significant differences between the samples were revealed (Wilk’s lambda \( \Lambda = 0.002, P < 0.001 \)). This means that the samples are highly heterogeneous in terms of the frequencies of the context characteristics, and cannot be merged together in one sample. The ChrClass program25,28 for prediction of the regions associated with different elements of the nuclear matrix in a query sequence is based on the results of LDA. A query sequence analysis progresses by analysing 300–1,000 bp regions of the sequence at a time. Four classes of regions are predicted in a query sequence: (1) S/MARs; (2) regions presumably associated with the protein cores of rosette-like structures (crDNA); (3) regions associated with the synaptonemal complex (scDNA); (4) regions associated with the nuclear lamina (nlDNA). A ChrClass prediction score (CCPS; normalised to fall between 0 and 1) indicates how likely it is that a predicted region is a true one.

COMPARATIVE STUDY OF NUCLEOTIDE SEQUENCES

Problems of accuracy evaluation

Evaluating the accuracy of prediction is an important issue. Experiments that are planned on the basis of computer prediction often need a considerable amount of effort and resources. Accuracy is also important for making a comparison between programs. Accuracy evaluation is extremely complicated in the case of the MAR-Finder and ChrClass programs, in part because there is no well-established control sequence sets. It is also difficult to define a set of sequences that do not associate with the nuclear matrix. In this paper properties of the MAR-Finder and ChrClass programs are discussed using 10 human protein-coding genes, a set of random sequences, the \( \beta \)-globin locus, L1 elements and satellite DNA.

Analysis of protein-coding genes

The results of analysis of 10 protein-coding genes are shown in Table 1. Considering all the data, ChrClass predicted 14 putative S/MARs, 4 regions presumably associated with the protein cores of rosette-like structures, 2 regions presumably associated with the synaptonemal complex and 5 regions presumably associated with the nuclear lamina, whereas MAR-Finder revealed 11 putative S/MARs (8 S/MARs for the threshold value 0.6). Seven putative S/MARs were found by both programs simultaneously (Table 1). The probability of observing such overlap at random calculated using a Monte-Carlo technique (randomly shuffling putative S/MARs predicted by MAR-Finder and ChrClass) is very low (\( P = 0.01 \)). This result suggests that the two programs gave consistent results. This consistency does not necessarily mean that the prediction is correct, but this is a positive feature of programs. Most putative S/MARs were found in non-coding regions, and such distribution can be expected for real S/MARs. Suggesting that some predicted regions are not real (such an assumption seems to be reasonable), more false S/MARs may be predicted.
The proportion of wrong S/MARs may be lower for MAR-Finder by ChrClass in comparison with MAR-Finder, and thus the proportion of wrong S/MARs may be lower for MAR-Finder. Analysis of random sequences Each of the ten protein-coding sequences from Table 1 was randomly shuffled. Because S/MAR should be absent or very rare in random sequences, the rate of predicted S/MARs in a random sequence is a measure of the false positive rate. Four putative S/MARs and three regions presumably associated with the nuclear lamina were found by ChrClass, while MAR-Finder predicted two S/MARs for the threshold value 0.5 and only one for the threshold value 0.6. This result supports our suggestion that the proportion of wrong S/MARs is lower for MAR-Finder.
Analysis of S/MARs in the β-globin locus

The human β-globin locus (HUMHBB, 73,308 bp) has been used for the MAR-Finder testing,17 since the location of S/MARs was experimentally determined for this locus.21 However, the conditions for the experimental detection of S/MARs by Jarman and Higge20 were much more stringent than conditions that are usually used for S/MAR detection.21 S/MARs were revealed in the K562 cells, which expressed only α- and foetal γ-globins,21 and thus some S/MARs can be detached from the nuclear matrix in these cells. Seven S/MARs were determined

---

**Table 2: Analysis of S/MARs in the human β-globin locus (HUMHBB)**

<table>
<thead>
<tr>
<th>Experimentally revealed S/MARs</th>
<th>MAR-Finder prediction by Walter et al. (1998)</th>
<th>MAR-Finder prediction (average strength)</th>
<th>ChrClass prediction (prediction score)</th>
<th>Description in the Feature Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–2,500</td>
<td>7.5–9,700</td>
<td>8,700–9,000</td>
<td>L1</td>
<td>8,019–8,314</td>
</tr>
<tr>
<td></td>
<td>11,700–13,900</td>
<td>11,300–11,900</td>
<td>L1</td>
<td>12,912–12,066</td>
</tr>
<tr>
<td></td>
<td>13,900–16,500</td>
<td>15,200–16,100</td>
<td>L1</td>
<td>14,836–15,701</td>
</tr>
<tr>
<td></td>
<td>16,000–16,100</td>
<td>23,300–24,600</td>
<td>L1</td>
<td>23,118–31,136</td>
</tr>
<tr>
<td>46,600–49,300</td>
<td>44,300–46,800</td>
<td>45,710–47,124</td>
<td>L1</td>
<td>46,146–46,996</td>
</tr>
<tr>
<td>Somewhere between α-globin and β-globin18</td>
<td>54,500–56,300</td>
<td>55,600–56,000</td>
<td>L1</td>
<td>54,740–56,389</td>
</tr>
<tr>
<td></td>
<td>58,900–61,900</td>
<td>58,100–58,500</td>
<td>L1</td>
<td>55,233–56,130</td>
</tr>
<tr>
<td></td>
<td>62,632–43,481</td>
<td>64,100–67,400</td>
<td>L1</td>
<td>65,533–65,757</td>
</tr>
</tbody>
</table>

HSI1 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.

1 For a marked S/MAR indicates that a precise location was not described.21 The ChrClass program predicted the location of the nuclear lamina associated site between positions 9,000 and 9,700.
experimentally in the HUMHBB sequence; however a 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 genes2 (Table 2). Even in the face of this problem, the experimental and computer predictions appear consistent: six out of seven experimentally determined S/MARs were predicted correctly by Mar-Finder (results from Walter et al.) and five S/MARs were predicted correctly by ChrClass (Table 2). The probability of observing the 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 MAR-Finder analysis of the HUMHBB sequence in order to obtain values of the average strength for the MAR-Finder prediction by Walter et al.2 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 2). Although four real S/MARs in the human β-globin locus cannot be predicted even for the threshold value 0.5, the average strength value 0.6–0.75 was considered to yield reasonable results by Singh et al.9 Thus, a choice of the threshold value for the MAR–Finder prediction is somewhat ambiguous. Results of the β-globin locus suggest that threshold values of less than 0.5 also can be recommended for some sequences. Interestingly, the location of the predicted S/MARs is conservative among the β-globin loci from different mammalian species (Figure 1), suggesting that some experimentally undiscovered S/MARs may exist there.

Analysis of L1 element

The presence of regions associated with the nuclear matrix in repeated SINEs and LINE (Short/Long INterspersed Elements) might be a very important choice of threshold value for MAR-Finder is somewhat ambiguous. MAR-Finder reveals putative S/MAR is a member of recently amplified L1 subfamily.

Figure 1: 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 human α-globin gene. HS1, 2, 3 and 4 are the DNase hypersensitive (HS) sites (LINE = Long Interspersed Element).
biological function of non-coding repeated DNA, since S/MARs can be quickly amplified and inserted in different genome regions as a part of repeated element. SINEs and LINEs have been found near or at some S/MARs. The human β-globin locus (Table 2) illustrates this property of S/MARs. MAR-Finder revealed putative S/MAR in positions 7,000–7,200 of L1Md-A2 element (AC M13002; 3' non-coding region of L1 element; MAR-Finder average strength = 0.68), which is a member of recently amplified L1 subfamily. ChrClass did not confirm this observation; however the region 1–900 (5' flanking sequence), regions 1,900–2,800 (ORF1) + 4,300–5,000 (ORF2) and region 5,000–5,500 (ORF2) have been predicted as region presumably associated with the protein cores of rosette-like structures (CCPS = 0.15), regions presumably associated with the nuclear lamina (CCPS = 0.20 and 0.12) and region presumably associated with the synaptonemal complex (CCPS = 0.06), respectively. The reality of both predictions is not clear, and these features of L1 may be interesting for further analysis.

Analysis of satellite DNA

It was found that centromeres are associated either with the inner surface of the nuclear envelope or with nucleoli. Human alphoid DNA consist of a monomer repeat (171 bp) tandemly repeated thousands of times at each centromere. These regions are configured into higher-order structures: several tandemly repeated monomers form a subunit (divergence of each repeat from consensus sequence is about 15–20 per cent); each subunit is, in turn, repeated hundreds of times. Experimental results have shown that 1.7 kb subunit of chromosome 16 and 1.9 kb subunit of chromosome 1 contain S/MARs. Several available sequences of alphoid DNA were analysed for chromosomes 7 and 21. From the three sequences available from the chromosome 7, one sequence contains a fragment presumably associated with the synaptonemal complex, two sequences contain a region presumably associated with the cores of rosette-like structures, and one sequences contains a putative S/MAR. The sequence from the chromosome 21 contains a fragment that is presumably associated with the synaptonemal complex; two sequences contain a region presumably associated with the cores of rosette-like structures, and one sequences contains a putative S/MAR. The sequence from the chromosome 21 contains a fragment that is presumably associated with the synaptonemal complex; thus, different subunits of alphoid DNA might be associated with different chromosome/nuclear substructures (Table 3).

It was found that lamin A and C are able to associate with the telomeric...
repeats of vertebrates. On the other hand, the telomeric repeat (TTAGGG)_n was shown to associate with the nuclear matrix. Furthermore, the telomeric repeats were shown to be associated with the inner matrix at a frequency of at least one per 1,000 nucleotides. An 'ideal' telomeric repeated sequence, (TTAGGG)_500, and a real telomeric sequence, (HSUBTERR), were predicted as putative S/MARs by the ChrClass program (Table 3).

No S/MARs were found by the MAR-Finder program in alphoid DNA and telomeric repeats. These results suggest that the proportion of missed S/MARs may be lower for ChrClass, since the presence of S/MARs in these sequences has experimental confirmation. The failure of MAR-Finder to detect S/MARs in alphoid DNA and telomeric repeats simply may reflect the dependence of MAR-Finder on the average base composition in a query sequence.

COMPARISON OF MAR-FINDER AND CHRCLASS

One obvious difference between MAR-Finder and ChrClass is the complexity of the statistical models: ChrClass uses more context characteristics in comparison with MAR-Finder. However, higher complexity of a model does not always guarantee better prediction since accuracy of prediction depends on both quality of these characteristics and methods of their combining. The problem of model parameterisation is well studied in molecular evolution. It has been found that simple models often give a better estimation of a phylogenetic tree in cases of high sampling variance, correlation among parameters, and high branch length heterogeneity. In S/MAR prediction, increasing model complexity can lead to unpredictable behaviour for sequences with an unusual oligonucleotide composition, and this could happen with the ChrClass analysis of the HSU04737 sequence (Table 1).

However, it appears that results of the ChrClass satellite DNA analysis (Table 3) are more consistent with the experimental data than the MAR-Finder prediction. The presence of the putative S/MAR in L1Md-A2 elements is an interesting feature of the MAR-Finder prediction; however, this observation needs experimental support.

A default set of parameters was used for the MAR-Finder searches. It should be noted that MAR-Finder was originally developed for cosmid-sized sequences, and thus results obtained from shorter sequences need to be considered carefully. When analysing sequences less than 40 kb in size, optimal results may be obtained when the window and step size are appropriately adjusted. This means that using MAR-Finder for fast multiple sequence analysis is inadequate.

The ChrClass program has been designed in order to predict sites associated with different elements of the nuclear matrix including S/MARs. Operationally S/MARs are defined as fragments able to bind with the nuclear matrix. At least two methods of S/MARs detection are considered 'standard', and conditions of experiments can vary significantly. These methods are different from approaches for revealing nucleotide sequences associated with the protein cores of rosette-like structures, with the synaptonemal complex and with the nuclear lamina. There are no data about the specificity of S/MAR, binding to various elements of the nuclear matrix. A putative S/MAR in positions 7,501–9,700 of the β-globin locus (MAR-Finder) overlaps with a region 9,000–9,700 that is predicted to be presumably associated with the nuclear lamina (ChrClass) (Table 2), and a putative S/MAR in positions 1,100–1,900 of the HSFOA sequence (MAR-Finder) was predicted by ChrClass to be presumably associated
with the protein cores of rosette-like structures (Table 1). These observations suggest that S/MARs may be associated with a variety of elements of the nuclear matrix. However, significant differences among sites associated with various elements of the nuclear matrix have been found by Glazkov et al., thus suggesting that a problem of a S/MAR formal definition is not yet solved and needs further analysis.

CONCLUSION
Results of our analysis suggest that both programs can be recommended for analysis of eukaryotic genomes. Our interpretation of the MAR-Finder/ChrClass comparison is that the proportion of missed S/MARs is lower for ChrClass, whereas the proportion of wrong S/MARs is lower for MAR-Finder. A considerable overlap between predictions of these programs has been found, and it can be suggested that regions predicted by both programs are better candidates for S/MAR. However, results of their prediction should be interpreted with caution since reliable estimates of prediction accuracy of both programs are not yet achieved. The real predictive power and functional importance of all employed context characteristics in MAR-Finder and ChrClass (described above) are not so obvious, and thus the problem of S/MARs prediction requires further analysis. Accumulation of new experimentally confirmed S/MARs will help to improve existing approaches and develop better models for the prediction of sites associated with various elements of the nuclear matrix.

AVAILABILITY
The MAR-Finder program is resident on the NCGR server (URL: www.ncgr.org/MARFinder/). The ChrClass program (Win95/NT version) is available from Galina V. Glazko (glazko@biotech.relc.com or glazko@biotech.kiev.ua) or Igor B. Rogozin (rogozin@bionet.nsc.ru).

Acknowledgements
We are grateful to J. Lyons-Weiler for helpful comments and for helpful recommendations of two anonymous referees. This work was partly supported by research grants from NIH and NSF to M. Nei and by the Russian Fund of Fundamental Investigation (Grant N 99-04-4935).

References


