Nucleic Acids Research, 1997, Vol. 25, No. 7 **1419–1425**

Mathematical model to predict regions of chromatin attachment to the nuclear matrix

Gautam B. Singh*, Jeffrey A. Kramer[1] and Stephen A. Krawetz[1]

Bioinformatics Algorithms Research Division, National Center for Genome Resources, 1800 Old Pecos Trail, Santa Fe, NM 87505, USA and [1]Department of Obstetrics and Gynecology, Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48202, USA

Received November 20, 1996; Revised and Accepted February 11, 1997

ABSTRACT

The potentiation and subsequent initiation of transcription are complex biological phenomena. The region of attachment of the chromatin fiber to the nuclear matrix, known as the matrix attachment region or scaffold attachment region (MAR or SAR), are thought to be requisite for the transcriptional regulation of the eukaryotic genome. As expressed sequences should be contained in these regions, it becomes significant to answer the following question: can these regions be identified from the primary sequence data alone and subsequently used as markers for expressed sequences? This paper represents an effort toward achieving this goal and describes a mathematical model for the detection of MARs. The location of matrix associated regions has been linked to a variety of sequence patterns. Consequently, a list of these patterns is compiled and represented as a set of decision rules using an AND–OR formulation. The DNA sequence was then searched for the presence of these patterns and a statistical significance was associated with the frequency of occurrence of the various patterns. Subsequently, a mathematical potential value, *MAR-Potential*, was assigned to a sequence region as the inverse proportion to the probability that the observed pattern population occurred at random. Such a MAR detection process was applied to the analysis of a variety of known MAR containing sequences. Regions of matrix association predicted by the software essentially correspond to those determined experimentally. The human T-cell receptor and the DNA sequence from the Drosophila bithorax region were also analyzed. This demonstrates the usefulness of the approach described as a means to direct experimental resources.

INTRODUCTION

Recent studies have established that human somatic cell chromatin is organized in loops that span ~50–100 kb (1). The points of attachment of these chromatin loops serve as specific sequence landmarks as they anchor the DNA sequence to the fibers of chromosomal scaffold. These sites of DNA attachment to the nuclear scaffold are termed scaffold (metaphase) or matrix (interphase) attachment regions (SAR or MAR). They are known to facilitate the expression of genes and may function as the origins of replication.

The matrix or scaffold attachment regions are relatively short (100–1000 bp long) sequences that anchor the chromatin loops to the nuclear matrix. MARs often include the origins of replication (ORI) and can possess a concentrated area of transcription factor binding sites (2). Approximately 100 000 matrix attachment sites are believed to exist in the mammalian nucleus of which ~30 000–40 000 serve as ORIs (3). MARs have been observed to flank the ends of genic domains encompassing various transcriptional units. It has also been shown that MARs bring together the transcriptionally active regions of chromatin such that the transcription is initiated in the region of the chromosome that coincides with the surface of nuclear matrix (3,4).

Matrix attachment regions have been categorized as constitutive (permanent) or facultative (cell-type specific) (2). The constitutive MARs occur in all types of cells irrespective of the tissue in which they are found. In contrast, the presence of a facultative MAR is tissue specific and its use is governed by that tissue. MARs have been experimentally defined for several gene loci, including the chicken lysozyme gene (5), human interferon-β gene (6), human β-globin gene (7), chicken α-globin gene (8), p53 (9) and the human protamine gene cluster (10).

It is widely accepted that the next phase of the Human Genome Project will focus on completing the transcript map. This will entail the mapping of the transcribed sequences to the appropriate regions of the chromosomes. To help identify these regions, some sequence identifiers, such as promoters, enhancers and locus control regions (LCR) are typically used. One of the clearest indicators of functional sequences are MARs. In light of the key role of MARs in genetic processes, and their localization to functional chromatin domains, a means to model these markers so that they could be placed on the map from sequencing data was sought. The results of our studies to computationally define MARs for experimental validation are presented.

Characterizing the regions of matrix attachment

MARs are polymorphic and appear to be distributed throughout the genome. There is no known consensus sequence that is characteristic of a MAR. Biologists have physically identified

* To whom correspondence should be addressed. Tel: +1 505 982 7840; Fax: +1 505 982 7690; Email: gbs@ncgr.org

Motif Index	*Motif Name*	*Sequence Predicate*
m_1	ORI Signal	$ATTA$
m_2	ORI Signal	$ATTTA$
m_3	ORI Signal	$ATTTTA$
m_4	TG-Rich Signal	$TGTTTTG$
m_5	TG-Rich Signal	$TGTTTTTTG$
m_6	TG-Rich Signal	$TTTTGGGG$
m_7	Curved DNA Signal	$AAAAn_7AAAAn_7AAAA$
m_8	Curved DNA Signal	$TTTTn_7TTTTn_7TTTT$
m_9	Curved DNA Signal	$TTTAAA$
m_{10}	Kinked DNA Signal	TAn_3TGn_3CA
m_{11}	Kinked DNA Signal	TAn_3CAn_3TG
m_{12}	Kinked DNA Signal	TGn_3TAn_3CA
m_{13}	Kinked DNA Signal	TGn_3CAn_3TA
m_{14}	Kinked DNA Signal	CAn_3TAn_3TG
m_{15}	Kinked DNA Signal	CAn_3TGn_3TA
m_{16}	mtopo-II Signal	$RnYnnCnnGYnGKTnYnY$
m_{17}	dtopo-II Signal	$GTnWAYATTnATnnR$
m_{18}	AT-Rich Signal	$WWWWWW$

Figure 1. The set of motifs characterizing MARs constitute DNA sequence signals or predicates upon which rules the defining higher level patterns are constructed. Note that the IUPAC characters R,Y, W and K are defined as follows. R = A or G, Y = T or C, W = A or T and K = G or T.

MARs and have tried to correlate their presence with the occurrence of several DNA sequence motifs, including the ORI, curved and/or kinked DNA. A description of some of the motifs that have been identified within several MARs is as follows.

Origin of replication. It is known that DNA replication is associated with the nuclear matrix. It has also been demonstrated that nuclear matrix attachment sites, homeotic protein recognition and binding sites and the origins of replication share the ATTA, ATTTA and ATTTTA motifs. This implies that differential activation of origins of replication (important for development) are regulated while part of the nuclear matrix (2). ORI motifs $m_1 \ldots m_3$ in Figure 1 have been used to formulate Rule 1 in Figure 2.

TG-rich sequences. Some matrix attachment regions have been characterized by T-G rich spans (2). These regions are abundant in the 3′ UTR of a number of genes, and may act as signals at the recombination sites, e.g. immunoglobulin genes. Motifs, $m_4 \ldots m_6$ in Figure 1 are used to compose Rule 2 (Fig. 2) that identifies the T-G rich spans.

Curved DNA. Intrinsically curved DNA has been identified at or near several matrix attachment sites (2,11). Curved DNA is also considered to play an important role in nuclear processes that involve the interaction of DNA and proteins, such as recombination, replication and transcription. Optimal curvature is expected for sequences with repeats of the motif, $AAAAn_7AAAn_7AAAA$ as well as the motif $TTTAAA$ (motifs $m_7 \ldots m_9$ in Fig. 1). Rule 3 will identify curved DNA.

Kinked DNA. Kinked DNA has generally been associated with the presence of copies of the dinucleotide TG, CA or TA that are separated by 2–4 or 9–12 nt. For example, kinked DNA will be produced by the motif TAn_3TGn_3CA, with TA, TG and CA occurring in any order (motifs $m_{10} \ldots m_{15}$ in Fig. 1). Rule 4 will identify kinked DNA.

Topoisomerase II sites. Topoisomerase II binding and cleavage sites are concentrated at the sites of nuclear attachment. Both vertebrate and Drosophila topoisomerase II consensus sequences have been identified (12,13), and are fashioned as Rule 5 in Figure 2.

AT-rich sequences. Many MARs contain significant stretches of AT-rich sequences. 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 are required for matrix association. Periodicity was considered while formulating Rule 6 (2), although consideration of local nucleotide concentration above a threshold may be required.

Several other characteristics, some of which have not been included in the current analysis, have been proposed for MARs. For example, MARs have been shown to contain palindromic sequences, Z-DNA and DNase I hypersensitive sites (2). A few *Alu* elements have also been identified within MARs. Specifically, *Alu*s with a high AT content may interact with the nuclear matrix. With the exception of *Alu* elements, whose role in matrix attachment is unclear (and their occurrence limited to primates), the above elements cannot be easily fashioned into a definitive consensus pattern. Moreover, these elements may not be appropriate or necessary for the mathematical determination of MAR-potential. For example, many bacterial, viral and mammalian ORI sequences, which are characteristic of MARs, are also palindromic. The occurrence of palindromic sequences at sites of nuclear matrix attachment may thus reflect the presence of origins of replication, which are already identified by Rule 1 in Figure 2. DNase I hypersensitivity, while possibly a characteristic of MARs, is likely a result of the interaction with the nuclear matrix and not its cause. Thus, DNase I hypersensitivity may represent a useful method for identifying MARs experimentally rather than computationally.

There is ample evidence to suggest that transcription factor binding sites, promoter regions and other regulatory regions of the genome may be nuclear matrix associated. While there are a myriad regulatory elements and promoter sequences of known composition, these have not been utilized in the model presented. The computational model proposed in this report will most likely

Rule	*Name*	*Definition*	*Probability*
R_1	ORI Rule	$m_1 \vee m_2 \vee m_3$	$p_1 = \sum_{i=1}^{3} Pr(m_i)$
R_2	TG-Richness Rule	$m_4 \vee m_5 \vee m_6$	$p_2 = \sum_{i=4}^{6} Pr(m_i)$
R_3	Curved DNA Rule	$m_7 \vee m_8 \vee m_9$	$p_3 = \sum_{i=7}^{9} Pr(m_i)$
R_4	Kinked DNA Rule	$m_{10} \vee m_{11} \vee m_{12} \vee m_{13} \vee m_{14} \vee m_{15}$	$p_4 = \sum_{i=10}^{15} Pr(m_i)$
R_5	Topoisomerase Rule	$m_{16} \vee m_{17}$	$p_5 = \sum_{i=16}^{17} Pr(m_i)$
R_6	AT-Richness Rule	$m_{18} \wedge_8^{12} m_{18}$	$p_6 = Pr(m_{18}) \cdot (1 - \exp(-5 \cdot Pr(m_{18})))$

Figure 2. The set of biological rules defining patterns that were used for detecting structural MARs. The table also specifies the relationship between the DNA motif probabilities, $Pr(m_i)$, and the rule probabilities, p_j.

identify constitutive or class 2 MARs, although, when adjusted, it has been successful in detecting facultative class 3 MARs (10).

The computational model for detecting MARs

Known consensus sequences for eukaryotic transcription factors and promoters can be identified using algorithms such as SIGNAL SCAN (14) and PROMOTER SCAN (15). These programs look for singular patterns, and are thus not very useful for determining the significance of the co-occurrence of many patterns. Similarly, 'gene grammar' has been utilized (16) to capture relations between the promoters, introns and exons. While this approach has its merits for detecting functional units within a sequence, its application is limited to recognizing patterns where the relationship between the component motifs is known *a priori*. A neural network can be utilized to recognize higher level patterns when such a relationship between the component motifs is not formally defined but learned by the computational agents. Such a system has been utilized by Grail where the network was trained to detect genes (≥100 bases in size) in raw DNA sequence (17). However, the lack of an appropriate training set (there are very few MAR regions experimentally mapped) makes neural network an impractical tool for this problem. Thus, our approach based on assigning a statistical significance to the co-occurrence of several MAR specific patterns represents a unique and, as we demonstrate below, viable solution to the MAR detection problem.

MAR patterns definition

As a first step toward the algorithmic detection of regions of probable matrix association (MARs), an effective mathematical framework for representing such patterns must be adopted. In our approach, the underlying architecture used to represent patterns is based on an AND–OR Boolean decision tree. As shown in Figure 3, such a tree represents a disjunction (OR) of the conjunctions (AND) on motifs detected in the sequence. Thus, sequence level motifs serve as the lowest level predicates used to detect the presence of a higher level pattern in the sequence. Note that the lowest level predicates may be negated before being used in the AND layer. In such an instance, the absence of motifs is sought to satisfy the conditions for the occurrence of the higher level pattern.

As an example, consider a simpler instantiation of such an AND–OR decision tree. A rule to define the origin of DNA replication (R_1) can be based on an OR or the $\vee$ operator applied to the three motifs m_1 = ATTA, m_2 = ATTTA and m_3 = ATTTTA. The motif detectors bypass the AND layer in this case, and directly feed into the OR layer.

$$R_1 = m_1 \vee m_2 \vee m_3 \qquad \mathbf{1}$$

Similarly, the requirement for co-occurrence of multiple motifs can be specified using the AND or the $\wedge$ operator. In the AND rule for multiple patterns, an additional parameter is incorporated to constrain the allowable gap between the two co-occurring motifs. For example, the AT-richness (R_6) rule has been formulated as the occurrence of two hexanucleotide strings, m_{18} = WWWWWW (note: the IUPAC code W denotes an ambiguous base A or T), that are separated by distance of 8–12 nt. Thus, the AT-richness rule can be written as:

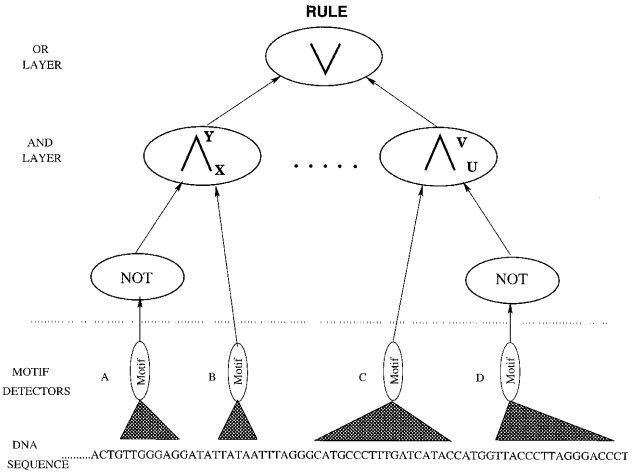


Figure 3. A rule is defined by utilizing logical connectives on the occurrence of underlying patterns. The DNA motifs (which may be overlapping) that are present or absent in the raw sequence serve as the input predicates upon which rule the logical definition of a rule is based. Such a generalized rule architecture is modeled after a AND–OR decision tree with the AND or $\wedge$ operator constraining the motifs separation distance. The highest level of this match hierarchy entails the OR or $\vee$ operation applied to the output of the AND layer. The hypothetical rule represented in the above figure is $(\overline{A} \wedge^{Y}_{X} B) \vee (C \wedge^{V}_{U} \overline{D})$.

$$R_6 = m_{18} \overset{12}{\underset{8}{\wedge}} m_{18} \qquad \mathbf{2}$$

Such a formulation uses an augmented AND operator, $\wedge^{high}_{low}$, to define the acceptable distance between the two motifs.

Rule database and probability assignment

As depicted in Figure 1, sequence motifs serve as predicates in the modeling of the MAR-detection rules. These predicates essentially represent the various sequence motifs that have been known to occur in the vicinity of MARs. With the motif indices thus defined, a set of rules for detecting higher level MAR patterns were developed and shown in Figure 2. This rule database represents the core set of rules needed to identify MARs in a DNA sequence. The core set of MAR pattern rules are considerably simpler than what can be represented by the general rule format (18,19). However, as the experimental determination of some additional regions of matrix association continues, other related patterns are likely to emerge. The more generalized framework for pattern representation will facilitate the incorporation of new (and potentially complex) pattern rules.

Associated with each of the motifs is a probability of its random occurrence. This is derived using the base composition of the sequence being analyzed (20). For example, the probability of finding the motif ATTA in a sequence with composition {A,C,G,T} = (0.2, 0.2, 0.3, 0.3) is equal to $(0.2^2 \cdot 0.3^2) = 3.6 \times 10^{-3}$.

In order to calculate the random probability of occurrence of a rule, the motif probabilities are multiplied across an AND layer when the motifs are independent, and added across an OR layer when their occurrence is mutually exclusive. Furthermore, assuming a Poisson density function for motif occurrences, the probability of finding at least one motif within an acceptable distance from the reference motif can be computed. Thus, the random occurrence probabilities for rules R_1 and R_6 in Equations **1** and **2** will be given by:

$$Pr(R_1) = Pr(m_1) + Pr(m_2) + Pr(m_3)$$
$$Pr(R_6) = Pr(m_{18}) \times \{1 - \exp[-(12 - 8 + 1) \cdot Pr(m_{18})]\} \quad \mathbf{3}$$

$$= Pr(m_{18}) \cdot \{1 - \exp[-5 \cdot Pr(m_{18})]\} \quad \mathbf{4}$$

As we demonstrate, these probability values for random occurrences of the various rules can be used for assigning statistical significance to the set of complex patterns that are detected in a given region of the sequence.

Statistical significance of MAR motifs

The task of detecting regions of matrix association is modeled 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 MARs. The sliding window algorithm uses two parameters, W and δ to measure a local property of the DNA sequence. The statistic of interest is measured in a window of size W centered at location x along that sequence. Successive window measurements are then made by sliding the window in increments of δ nucleotides. If δ is small, linear interpolation can be used to join the individual window statistics gathered at x, $x + \delta$, . . . , $x + k\delta$. In this manner, a continuous distribution of the parameter of interest is obtained as a function of x.

The null hypothesis, H_0, tested in each window is as follows, H_0: the frequency of the MAR patterns observed is not significantly different from the frequency expected in a random W nt sequence of the same composition as the sequence being analyzed.

Thus, large deviation from the expected frequency of patterns in a window will force the rejection of H_0, which in turn will imply the presence of a MAR. Under H_0, the cumulative probability, p, of observing a frequency vector with each of its components greater than or equal to f_i is essentially the probability that the null hypothesis will be erroneously rejected. In other words, a small value for p signifies that the observed event would be a rare occurrence under the null hypothesis and hence qualify the window sample as a candidate for containing a site for matrix attachment.

In order to quantify the significance of this deviation, the statistic measured for each window is $[-log(p)]$. The value $[-log(p)]$ is also referred to as MAR-potential, and denoted as ρ. The value of ρ is computed for both the forward and the reverse strands of the DNA sequence and the average of the two values is considered to be the potential at a given location. We now describe the mathematical model for calculating the value of ρ.

Let us assume that we are searching for k distinct types of MAR patterns within a given window of the sequence. Typically, these patterns are defined as rules $R_1, R_2, \ldots, R_k$. Using the probability formulation defined in Equation **3**, the random probability of the occurrence of the various rules is calculated. Let these values be $p_1, p_2, \ldots, p_k$.

Next, a random vector of pattern frequencies, F, is constructed. F is a k-dimensional vector with components, $F = \{x_1, x_2, \ldots, x_k\}$, where each component x_i is a random variable representing the frequency of the pattern R_i in the W nt window. Furthermore, the component random variables x_i are assumed to be independently distributed Poisson processes, each with the parameter λ_i. The joint probability of observing a frequency vector $F_{obs} = \{f_1, f_2, \ldots, f_k\}$ purely by chance is given by Equation **5**.

$$P(F_{obs}) = \prod_{i=1}^{k} \frac{exp^{-\lambda i}\lambda_i^{fi}}{f_i!} \quad where\ \lambda_i = p_i \cdot W \quad \mathbf{5}$$

The steps required for computation of the p, the cumulative probability that the observation satisfies H_0, is given by Equation **6** below.

$$\begin{aligned} p &= Pr(x_1 \geq f_1, x_2 \geq f_2, ..., x_k \geq f_k) \\ &= Pr(x_1 \geq f_1) \cdot P_r(x_2 \geq f_2) \cdot ... \cdot Pr(x_k \geq f_k) \\ &= \sum_{x_1=f_1}^{\infty} \frac{\exp^{-\lambda_1}\lambda_1^{x_1}}{x_1!} \cdot \sum_{x_2=f_2}^{\infty} \frac{\exp^{-\lambda_2}\lambda_2^{x_2}}{x_2!} \cdot ... \cdot \sum_{x_k=f_k}^{\infty} \frac{\exp^{-\lambda_k}\lambda_k^{x_k}}{x_k!} \end{aligned} \quad \mathbf{6}$$

The p-value in Equation **6** is next utilized to compute the value of ρ or the MAR-potential as given by Equation **7** below.

$$\begin{aligned} \rho &= \ln\frac{1.0}{p} = -\ln(p) \\ &= \sum_{i=1}^{k}\lambda_i + \sum_{i=1}^{k}\ln f_i! - \sum_{i=1}^{k} f_i \ln\lambda_i \\ &\quad - \sum_{i=1}^{k}\ln(1 + \frac{\lambda_i}{f_i + 1} + \frac{\lambda_i^2}{(f_i + 1)(f_i + 2)} + ... + \frac{\lambda_i^t}{(f_i + 1)(f_i + 2)...(f_i + t)}) \end{aligned} \quad \mathbf{7}$$

It is not surprising that a pattern's contribution to the overall MAR-potential is strongest when its observed frequency is high while its probability of random occurrence is low. The infinite summation term in Equation **7** quickly converges and thus can be adaptively calculated to the precision desired. For small values of λ_i, the series may be truncated such that the last term satisfies the following condition:

$$[\frac{\lambda_i^t}{(f_i + 1)(f_i + 2)...(f_i + t)}] \leq \epsilon \quad \mathbf{8}$$

Depending upon the level of accuracy desired, the value of ϵ is typically set to a small positive number. In our implementation, ϵ was set at 10^{-4}.

Inferring MAR locations

After computing the values for ρ, the next task requires the interpretation of the statistical potential values to infer the location of matrix attachment sites. The following provides a discussion of the basis for selecting some key parameters that can influence the determination of these sites.

Run length. A true MAR will be characterized by a run-length of high potential values. In other words, if the average span of a MAR is M_{span} bases, we should expect to see successive high potential values in an average run length of:

$$r_L = \frac{W - M_{span} - \delta/2}{\delta} \quad \mathbf{9}$$

Thus, if W = 1000 bp and δ =100 bp, and if M_{span} is assumed to be ~600 bp, an average run-length of three high potential values is expected.

The use of this parameter is illustrated in Figure 4. The top panel shows the analysis of the sequence using a window size W = 1000 bp, while the lower panel shows the same analysis done with W = 2000 bp. In both these cases, however, the windows were stepped by δ = 100 bp. Using the formula in Equation **9**, the expected value for r_L is 3 in the first case and 13 in the second.

The r_L cut-offs specified above were used in establishing the locations of the various MARs. Although, locations of MARs are

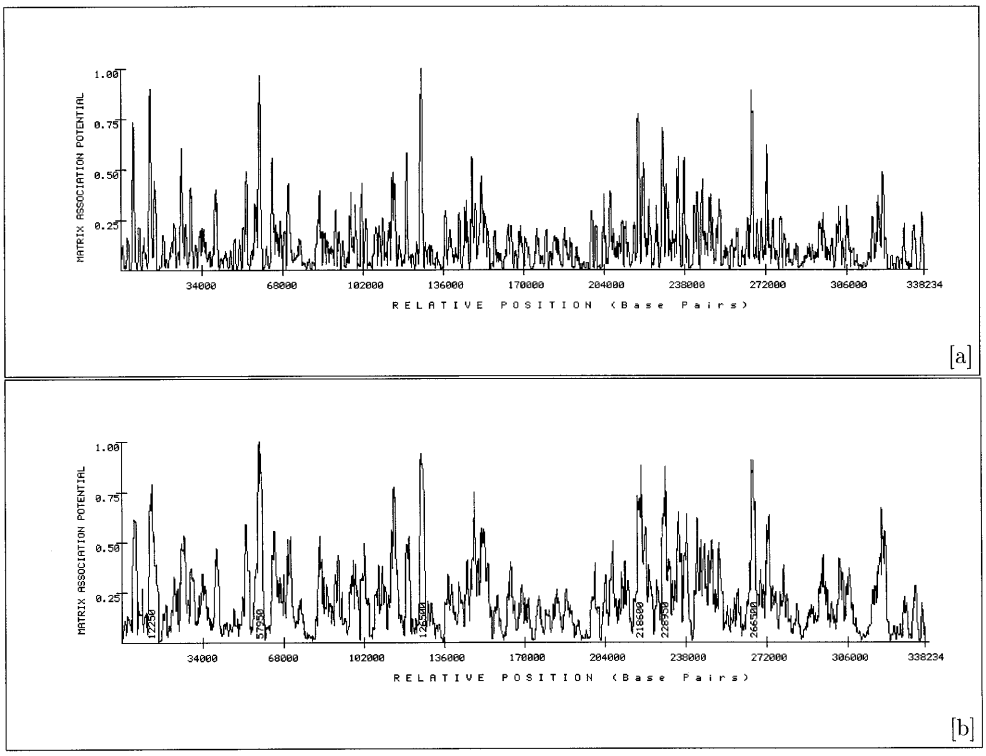


Figure 4. Detection of MARs in a 338 234 bp bithorax complex of *Drosophila* (U31961). MAR predictions are done with (**a**) W = 1 kb and (**b**) W = 2 kb. A window step size of δ = 100 bp was used in both the cases.

only plotted on Figure 4b, the values were identical in both cases. It may be noted that in general, the larger window size tends to smooth the potential values and widen any peaks. This is to be expected since a larger value of λ_i shifts the Poisson density function to the right and increases the probability of observing a high f_i under the null hypothesis.

Normalization. From the formulation of ρ in Equation **7**, it is clear that value for MAR-potential is unbounded, i.e. $0 \leq \rho \leq \infty$. This warrants that the raw potential be normalized in some manner. A linear normalization technique is to scale every value by $[1.0/\rho_{max}]$. After such a scaling, the value of ρ is restricted to fall between [0..1] interval. However, there is a concern when using this simple procedure. If the value of ρ_{max} happens to be extremely high, it will tend to attenuate the other, albeit statistically significant, potential values. This becomes apparent from observing the very high value in the potential graph shown in Figure 5a, where the potential value at location ~438 kb is substantially higher than the other peaks. In order to avoid this scenario, a saturation value can be set on ρ_{max}. Such a saturation value can either be pre-set to an absolute value based on statistical constraints, or be dynamically chosen based on the peaks observed in the sequence. In our implementation, a hybrid approach was adopted. If ρ_{max} was larger than a specified threshold, a saturation value equal to a fraction of the largest peak potential is used. In the potential graph shown in Figure 5b, the peaks were saturated to $0.4 \times \rho_{max}$.

Intelligent zoom. While processing a large sequence, caution must be exercised in the interpretation of a graphical plot of the potential values. Specifically, if a large number of bases are packed into a single pixel location, multiple peaks will have the tendency to overlap one another. This could possibly leave one with a false impression that the results are noisy.

To visualize the locations of these MARs, the analysis window can be zoomed to the area of interest. This is illustrated in Figure 5c where the displayed region spans locations 360 kb through 460 kb. The process is referred to as being intelligent since the normalization of the potential values occurs only on the basis of the potential values present in the zoomed region. That is, the saturation values for the entire sequence may be different than that used in the viewing window.

RESULTS

The ability of the MAR prediction algorithm described above to identify MARs is shown in Figure 6. When physical evidence is available, as in the human-globin and PRM1→PRM2→TNP2 domains, the matrix attachment regions predicted by the software closely match those established experimentally (18). This analysis also successfully identified the MAR at the 5′ end of the human apolipoprotein B locus (18,19). Based on these analyses, a normalized MAR-potential value of 0.6–0.75 was considered to yield reasonable results. The higher value is generally used for smaller window size.

The algorithm was then applied to identify candidate MARs in the *Drosophila melanogaster* bithorax region and the human T-cell receptor beta locus, for which the location of MARs are not known. The results for the Drosophila bithorax region are shown in Figure 4 while Figure 5 shows the analysis for the human β T cell receptor locus. Potential MARs spaced ~60–100 kb apart were identified, which is in accord with data suggesting that MARs in somatic nuclei occur at intervals of 60 kb to >100 kb (1).

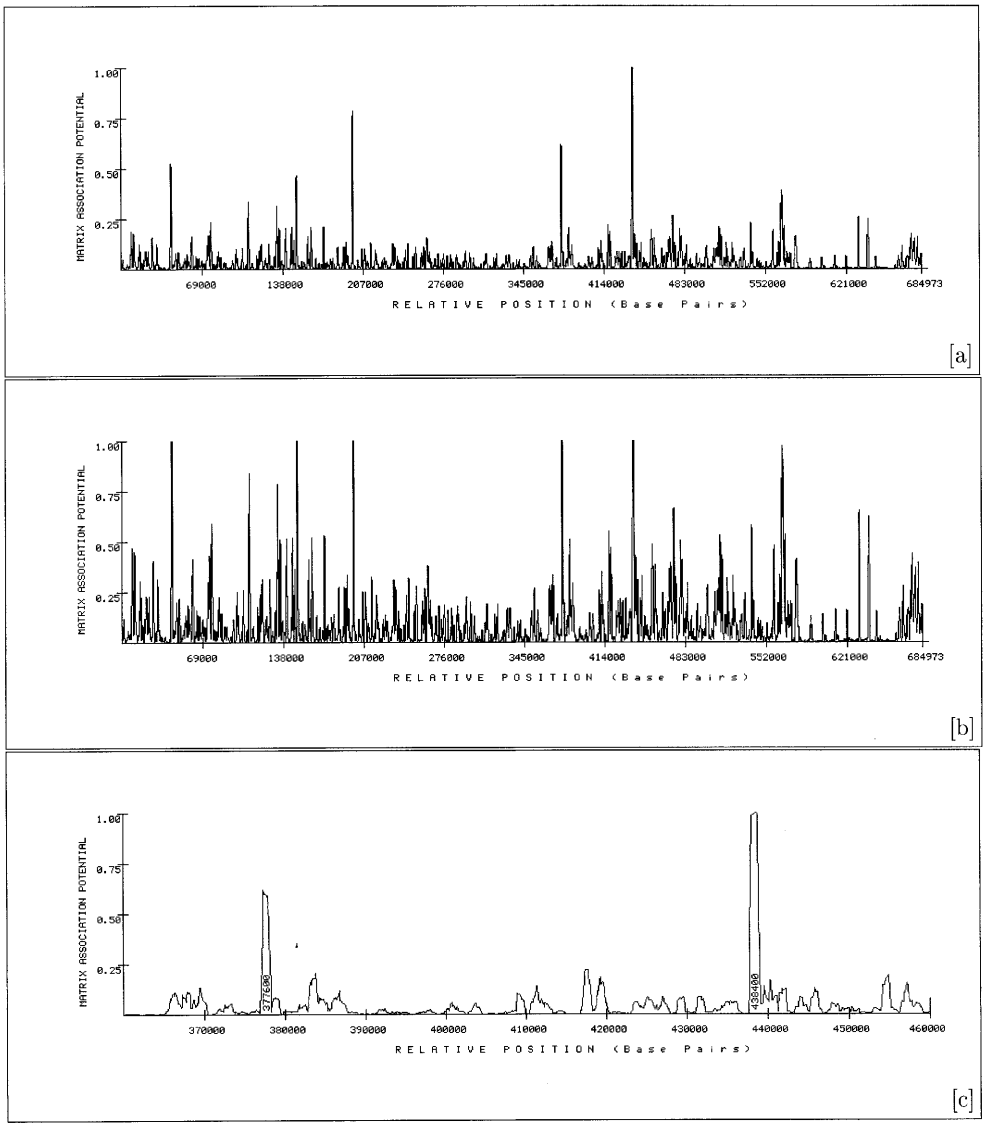


Figure 5. The MARs detected in the 684 973 bp sequence of the human β T cell receptor locus (21). (**a**) The normalized plot of potential values without any saturation, (**b**) The MAR-potential with a saturation value of $0.4 \times \rho_{max}$, (**c**) A zoomed view of the 360–460 kb region.

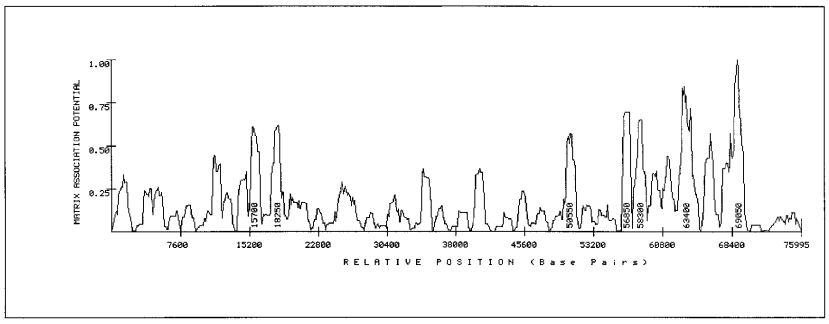
This suggests that reasonable candidate regions of interest for physical analysis were identified.

To the best of our knowledge, there is no data to define the spacing of MARs in the Drosophila genome. It is reasonable to assume that the spacing would be similar to that seen in other eukaryotic nuclei. The ~338 kb region queried in Figure 4 constitutes a region of the Drosophila genome that contains several homeotic genes. This has been suggested to be rich in constitutive type MARs. The regions identified by this analysis need to be verified experimentally, as they may represent characteristic constitutive or class 2 MARs (10).

Conclusion

The method of analysis described above has successfully identified known MARs in several well characterized loci. Furthermore, it has determined MAR candidates in a reasonable fashion in large sequences where sites of attachment to the nuclear matrix are unknown. This is despite the lack of a clear consensus sequence for MARs. However, care should be taken when applying this approach, as the results can vary when different combinations of rules are applied. This may reflect the beginning of the mathematical resolution of the four classes of MARs. The algorithm described should be readily applicable for the identification of any functional element for which a consensus sequence is unclear or unknown. The utility of this approach is immediately obvious for the identification of regulatory regions of locus control, which may contain multiple individual sequence motifs. Such regulatory motifs, which may occur at random throughout the genome, would be in close association with other regulatory and promoter motifs at regions of locus control, like the well characterized locus control region of the multigenic β-globin domain. Use of this algorithm in conjunction with a database containing known regulatory motifs should enable the identification of potential regions of locus control in large sequences, just as the current analysis identified MARs.

[a]



[b]

Sequence	*Length*	Software Predicted MARs		Experimentally Determined	
		No.	*Position*	*No.*	*Position*
β-globin [7]	75,995	–	–	1	~1.5 kb
(L22754 + HUMHBB)		2	~15–18 kb	1	~15–18 kb
		1	50,550	1	~46–54 kb
		4	~56–69 kb	2–4	~58–70 kb
*PRM1→PRM2	40,573	1	8,700	1	~9.3 kb
→TNP2 (U15422)		1	34,350	1	~33 kb
		1	38,150	1	~38 kb
Apolipoprotein B	7,274	1	225	1	~200 bp
D. Melanogaster	338,234	1	12,250		
Bithorax (U31961)		1	57,950		
		1	126,500		not known
		2	~218–229 kb		
		1	266,500		
Human β T-Cell	684,973	1	42,600		
Receptor (L36092)		1	149,900		
		1	198,000		not known
		1	377,600		
		1	438,400		
		1	565,400		

Figure 6. Veracity of MAR prediction: (**a**) MARs detected in the human β-globin locus. (**b**) Predicted and experimental MARs: comparison of the MAR locations predicted and those determined experimentally. *Note that all MARs predicted for the protamine locus excluded the AT-richness rule, R_6, and all were experimentally verified.

Software

A beta-version of the software described in this paper may be obtained by sending an e-mail message to gbs@acm.org.

REFERENCES

1 Vogelstein,B., Pardoll,D. and Coffey,D. (1980) *Cell*, **22**, 79–85.
2 Boulikas,T. (1993) *J. Cell. Biochem.*, **52**, 14–22.
3 Bode,J., Stengert–Iber,M., Kay,V., Schlake,T. and Dietz–Pfeilstetter,A. (1996) *Crit. Rev. Euk. Gene Exp.*, **6**, 115–138.
4 Nikolaev,L., Tsevegiyn,T., Akopov,S., Ashworth,L. and Sverdlov,E. (1996) *Nucleic Acids Res.*, **24,** 1330–1336.
5 Phi–Van,L. and Strätling,W. (1988) *EMBO J.*, **7**, 655–664.
6 Jade,J., Rios–Ramirez,M., Mielke,C., Stengert,M., Kay,V. and Klehr–Wirth,D. (1995) *Int. Rev. Cytol.*, **162A**, 389–454.
7 Jarman,A. and Higgs,D. (1988) *EMBO J.*, **7**, 3337–3344.
8 Farache,G., Razin,S., Targa,F. and Scherrer,K. (1990) *Nucleic Acids Res.*, **18**, 401–409.
9 Deppert,W. (1996) *J. Cell. Biochem.*, **62**, 172–180.
10 Kramer,J. and Krawetz,S. (1996) *J. Biol. Chem.*, **271,** 11619–11622.
11 von Kries,J., Phi–Van,L., Diekmann,S. and Strätling,W. (1990) *Nucleic Acids Res.*, **18**, 3881–3885.
12 Spitzner,J. and Muller,M. (1988) *Nucleic Acids Res.*, **16,** 5533–5556.
13 Sander,M. and Hsieh,T. (1985) *Nucleic Acids Res.*, **13**, 1057–1067.
14 Prestridge,D. (1991) *Comput. Applic. Biosci.*, **7**, 203–206.
15 Prestridge,D. (1995) *J. Mol. Biol.*, **249**, 923–932.
16 Dong,S. and Searls,D. (1994) *Genomics*, **23**, 540—551.
17 Mural,R., Einstein,J., Guan,X., Mann,R. and Uberbacher,E. (1992) *Trends Biotech.*, **10**, 66–69.
18 Kramer,J., Singh,G. and Krawetz,S. (1996) *Genomics*, **33**, 305–308.
19 Kramer,J. and Krawetz,S. *Mammalian Genome*, **6**, 677–679.
20 Staden,R. (1988) *Comput. Applic. Biosci.*, **5**, 89–96.
21 Rowen,L., Koop,B. and Hood,L. (1996) *Science*, **272**, 1755–1762.