VERSITA

CELLULAR & MOLECULAR BIOLOGY LETTERS
Volume 11, (2006) pp 191 – 213
http://www.cmbl.org.pl

DOI: 10.2478/s11658-006-0016-4

Received: 15 December 2005
Revised form accepted: 26 February 2006

In-silico PREDICTION AND OBSERVATIONS OF NUCLEAR MATRIX ATTACHMENT [#]

ADRIAN E. PLATTS[1], AMELIA K. QUAYLE[2] and STEPHEN A. KRAWETZ[1,2,3]*
[1]Department of Obstetrics and Gynecology and [2]The Center for Molecular Medicine and Genetics, [3]Institute for Scientific Computing Wayne State University School of Medicine, 253 C.S. Mott Center, 275 E Hancock, Detroit, MI 48201, USA

Abstract: The nuclear matrix is a functionally adaptive structural framework interior to the nuclear envelope. The nature and function of this nuclear organizer remains the subject of widespread discussion in the epigenetic literature. To draw this discussion together with a view to suggest a way forward we summarize the biochemical evidence for the modalities of DNA-matrix binding alongside the *in-silico* predictions. Concordance is exhibited at various, but not all levels. On the one hand, both the reiteration and sequence similarity of some elements of Matrix Attachment Regions suggest conservation. On the other hand, *in-silico* predictions suggest additional unique components. In bringing together biological and sequence evidence we conclude that binding

---

[#] Invited paper

* Corresponding author, e-mail: steve@compbio.med.wayne.edu; tel: (313)-577-6770, fax: (313)-577-8554

Abbreviations used: ChrClass - a linear discriminant analysis approach to MAR prediction; CS - chromosomal scaffold; CT - chromosome territory; IUPAC - International Union of Pure and Applied Chemists; LDA - linear discriminant analysis; MAR - matrix attachment region, MARFinder - a cumulative probability MAR prediction tool; MARSCAN - a MAR prediction tool to detect the MRS; MRS - the bipartite MAR recognition signature; MARWIZ - a commercial implementation of marfinder; MHC - major histocompatibility complex; mRNP - messenger ribonucleic acid protein; NM - nuclear matrix; PWM - position weight matrices; SIDD - stress induced duplex destabilization; S/MAR - scaffold/matrix attachment regions (synonymous with MAR); SMARTest - a MAR prediction tool developed commercially by Genomatix; Tw - number of helical turns in a constrained DNA loop; Wr - wumber of superhelical turns in a constrained loop.

All referenced websites were verified. URLs and their content are subject to change. The content referenced in this paper can be accessed using internet archive tools such as www.archive.org with the query restricted to November 2005.

may be hierarchical in nature, reflective of a biological role in replicating, transcribing and potentiating chromatin. Nuclear matrix binding may well be more complex than the widely accepted simple loop model.

Key words: Nuclear matrix, Matrix attachment regions, *In-silico*, Prediction, MARSCAN, MarFinder, MARWIZ, ChrClass, SMARTest, SIDD

INTRODUCTION

It has become widely accepted that sets of reiterated nucleotide motifs as well as distinct structural conformations of eukaryotic DNA can bind a functionally active nuclear organizing framework [1, 2]. Termed the nuclear matrix (NM), the nuclear scaffold and by some the nucleoskeleton, this framework has now been imaged through both optical [3-5] and electron microscopy [1, 2]. As summarized in Fig. 1, imaging has revealed a structure with several distinct and likely functionally differentiated components. The lamina, also termed the peripheral or type I nuclear matrix, is a two dimensional structure, just internal of the inner nuclear membrane. It is tethered at this position by a set of membrane-spanning lamin-binding proteins [6]. The type II inner nuclear matrix forms a three dimensional network throughout the nucleus.

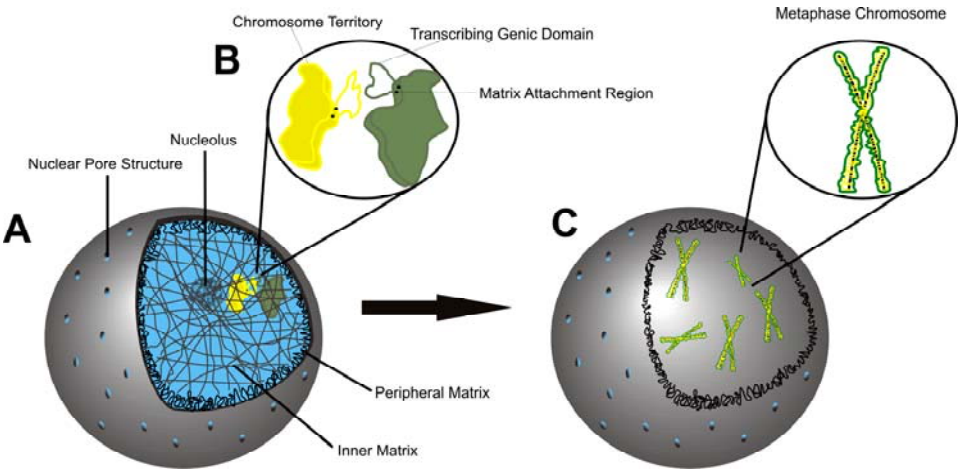

Fig. 1. The nuclear matrix. A - Interphase nucleus. The outer nuclear matrix is represented in black and the inner nuclear matrix is shown in gray. B - Chromatin territories (CTs). Nuclear matrix attachment regions (black dots) decrease the torsional strain on genic domains. These domains then loop into trans-factor reservoirs between CTs to participate in active transcription. C - Metaphase nucleus showing condensed sister chromatids. The outer nuclear matrix will disperse with the nuclear membrane while the inner nuclear matrix partitions into two parts: a central chromosomal scaffold (black core along the length of chromatids) and a perichromosomal layer (green), the majority of which is composed of mRNPs.

The complex nature of the NM is evidenced by its protein composition. In total, over 500 NM binding proteins are now recorded in the NM protein database [7]. Approximately 200 can be detected at any given time by 2D gel electrophoresis [8]. Laboratory protocols have been developed to isolate both the specific sequences binding to the nuclear matrix and the proteins located on the NM. Some of the most thoroughly investigated NM proteins include SAFB [9], Topoisomerase II [10], SATB1 [11], NFμNR [12] and CTCF [13].

Chromatin staining has demonstrated that the nuclear matrix is broadly enriched in the A and C, as well as the more acidic B1 and B2 lamins that range in size from 60 to 75 kDa [14]. Intermediate fibers are likely formed through the binding of lamins with filamentous actin and actin binding proteins [8, 15]. These fibers are found dispersed throughout the nuclear core where they generate a mesh-like network interlinked by finer 10 nm filaments rich in globular mRNPs [16-19]. Recent studies show a widespread intermingling of otherwise segregated DNAs from different chromosomes following RNAse treatment [18]. Accordingly, the matrix binding mRNPs may be considered as serving a structural role in the inner NM. The dense lamins that predominantly define the peripheral NM are resistant to solubilization in high salt solutions while the sparse inner nuclear matrix lamins are not. Nonetheless, the inner and peripheral networks appear physically continuous, suggesting that they are physically linked but functionally distinct [20].

Several sub-nuclear structures have been recognized as being NM associated. These include:

1. the central 0.5-1 μm topoisomerase II rich chromosomal scaffold (CS) [21] to which condensing metaphase chromatin attaches;
2. the rosette-like structures at the core of the CS [22-24];
3. the synaptonemal complex, a prophase structure similar to the CS formed during meiotic division [25];
4. the perichromosomal mRNP rich layer that covers metaphase chromosomes [26, 27];
5. the internal fibrillar and external granular structures observed within nucleoli [28].

Consensus is emerging on the role of the nuclear matrix as a binding framework that is both structurally adaptive and also capable of bringing about a synchronous nuclear coordination. Chromosomes are organized into their segregated chromosome territories (CTs) through binding to this framework. With the exception of rDNA, this organization is maintained even throughout interphase where the chromosomes are relatively dispersed [5]. Time-lapse microscopy reveals that CTs are repositioned slowly relative to each other in contrast to the more dynamic thermally induced intra-CT motion [29]. Without the coordinating influence of a semi-rigid matrix, this lack of relative motion would be perplexing.

Actively expressed sequences within CTs are organized by their interactions at discrete matrix binding sites into conformationally open domains that may be

both facultatively and constitutively potentiated by NM binding [30]. Examples of such domains include the human β-globin [31], β-interferon [32] and protamine [33] gene clusters. These domains appear to be actively positioned towards the borders of the territories encompassing highly expressed genes or gene dense clusters [34, 35]. Their repositioning occurs within a 10 second timeframe that has been related to active enzymatic processes [36]. At the perichromatic CT borders, DNA from different chromosomes may overlap, placing actively transcribing segments in juxtaposition with multiple reservoirs of splicing factors [37-39]. The reduced chromatin density along the perichromatic borders permits a more rapid diffusion of transcripts between CTs, throughout the nucleus and ultimately to the nuclear pores.

FUNCTIONS OF THE MATRIX AND BINDING MODALITIES

Image analysis of binding to the NM suggests that in mammalian genomes, there may be as few as 30,000 or as many as 100,000 sites of DNA attachment to the nuclear matrix [40, 41]. The function of these attachment sites has been widely investigated and linked to:

1. demarcating the ends of genic domains, between which loops of chromatin can be coordinated for transcriptional potentiation [42];
2. constraining long range enhancing and silencing elements to their requisite domains [43];
3. spatially coordinating transcription and replication [44, 45];
4. maintaining chromosome structure and morphology, e.g., though telomeric and centromeric regions [46];
5. countering mechanical disruption through structural support [47].

Matrix binding sequences can, to an extent, be qualitatively predicted. They occur frequently, but not exclusively, within AT-rich regions, where the weak AT bond permits the ready binding of duplex-splitting proteins. Serially reiterated simple repeat sequences with distinct higher-order structures such as the telomeric $(TTAAGGG)_n$ repeats [48, 49] and the pericentromeric alpha satellite [50] repeats also exhibit a strong nuclear matrix association. Progress in defining other interactors and sites of contact has been hampered by the time-consuming *in-vivo* analyses for each well defined target. Of the many thousands of anticipated binding sequences, currently, only 559 have been experimentally verified, each to a various extent and recorded in the scaffold/matrix attachment region transaction database (http://smartdb.bioinf.med.uni-goettingen.de/SMARt DB/browse/index.html) [51]. This is similar to the number of distinct proteins located in the matrix with either a direct or indirect potential to bind DNA.

Computational approaches

The development of *in-silico* tools has been presented as a means to eliminate the Matrix Attachment Region profiling bottleneck. These approaches have largely arisen alongside techniques used for transcription factor and regulatory

Tab. 1. Sequence motifs employed by the MAR identification algorithms as either protein binding sequences or approximations of topological features of DNA. The Genomatix PWMs are unpublished.

Sequence	MAR Type	Used by
AWWRTAANNWWGNNNC & AATAAYAA	MAR Recognition Signature	MarScan ChrClass SMARTest
$(GC)_n$	Z-form DNA	ChrClass n>5
$(A+T)_{rich}$ or $(AT)_n$	Z-form/Cruciform DNA, Origin of replication sites, Duplex unpairing sites	ChrClass n>5 MarFinder
ATTA, ATTTA, ATTTTA	Origin of replication sites	MarFinder
$TGN_{2\text{-}4/9\text{-}12}TG$ $CAN_{2\text{-}4/9\text{-}12}CA$ $TAN_{2\text{-}4/9\text{-}12}TA$	Kinked DNA	MarFinder ChrClass
$A_n/G_n/T_n/C_n$	Homopolytracts, Lamin binding	ChrClass n>4
$R_n/Y_n/S_n/K_n/M_n$	Short Repeats	ChrClass, n>6
AATATATTT	base unpairing sequence	ChrClass SMARTest
A4N7A4N7A4 or T3A3	Curved DNA	ChrClass MarFinder
$(RY)_n$ Genomatix PWM	TopoIsomerase II binding	ChrClass SMARTest
$(TTAGGG)_n$	Vertebrate Telomeric Repeats	ChrClass
$(TG)_{rich}$	3' UTR MARs	MarFinder ChrClass
TCTTTAATTTCT AATATATTTAGAA Genomatix PWM	SATB1 Motif	MarFinder SMARTest
H_n	Rice Motif associated with Helix destabilization	MarFinder (n>20)
Genomatix PWM	NFmμNR	SMARTest
Genomatix PWM	Bright	SMARTest

element binding-sequence detection typified by a weighted pattern matching to identify sets of representative core motifs. The core patterns currently used for MAR detection are summarized in Tab. 1. While sites of regulatory control tend to be precisely located and well matched to their consensus sequences, a matrix binding potential can be dispersed, serially reiterated and sometimes

poorly matched to proposed motifs. This difference could reflect the process of potentiation, i.e., opening a chromatin domain. Transcription factor binding sites are utilized after potentiation, but those long range *cis*-elements that mediate potentiation must be of sufficient flexibility to permit functional binding, irrespective of whether an optimal or transitional chromatin folding has been achieved. Thus, a protracted binding sequence that is available to the matrix even where the site is sterically hindered becomes favored.

On the one hand, elements that potentiate chromatin, even while benefiting from reiteration over multiple nucleosomes, may not be strictly conserved. Indeed, their very reiteration may well anti-correlate with conservation, since the biological function of a highly reiterated sequence would likely be robust relative to the effects of a base change or insertion-deletion event. On the other hand, MAR sequences that are linked to transcriptional initiation or stabilization would likely be more similar to transcription factor binding elements, exhibiting both higher conservation and perhaps nucleosomal periodicity. Accordingly, we may anticipate that MAR sequences will vary in format from dispersed and reiterated, at the level of the domain, to a single focused element at the gene level.

Several MAR sequence detection strategies have been developed. Most have relied on pattern-matching strategies. These tools adopt sequence pattern approximations to structural models rather than explicitly modeling the structural conformation of the query sequence. Explicit DNA modeling has also been employed which has proven a computationally intensive task. In an attempt to increase specificity and/or selectivity others have adopted meta-approaches where both structural approximations and sequence algorithms are simultaneously applied. Beginning with MarFinder [52], the software tools to detect MARs including SMARTest, SIDD, ChrClass and MARSCAN [53] have evolved a set of strategies that in some ways now differentiate them from other regulatory sequence detection schemes.

SMARTest

SMARTest [51, 54] has been developed as an element within the Genomatix suite of sequence analysis tools (http://www.genomatix.de) to predict nuclear matrix attachment sites. Detection is determined relative to a set of Position Weight Matrices (PWMs) derived from an unsupervised mining of MAR binding sequences compiled from both the literature, EMBL and S/MARtDB [51]. The sequence alignment tool DiAlign was used to generate representations of the conserved motifs between 34 training sequences from 16 plant (7 in Arabidopsis) and 18 metazoan matrix binding sequences.

DiAlign [55] employs a novel approach to sequence alignment relative to the more widely implemented Needlman-Wunsch [56] approach. DiAlign, and later variants specifically target sequences across multiple training sets that share consistently ordered but not necessarily contiguous subsequences. This approach allows once proximate sequences that have moved apart between species or

across segmental duplications to be re-identified. As such, the DiAlign strategy is well suited for searching multipartite or highly divergent sequences.
Having been transformed into weight matrices [57], conserved sequences from DiAlign were screened using a variant of the Genomatix CoreSearch tool [58]. A total of 97 PWMs were initially identified that ranged in size from 10 to 21 bp. They were present at a frequency of ≤ 4 per 10 kb in random genomic DNA. This represents a substantial number of matrices relative to a training set of 34 sequences. When evaluated *post-hoc* against its training set, SMARTest correctly predicted 27 of its 34 training MARs, suggesting that either more subtle PWMs are required or binding modalities that could not be identified using this PWM approach exist.
A-priori testing of SMARTest using sequences containing known MARs revealed that relative to 37 experimentally verified MARs in 310 kb of DNA, SMARTest correctly predicted 14 and mispredicted a further 9 sequences. If this is representative of genomic sequences, then estimates are subject to an over-prediction or type I error rate of 40%. While the type II error rate of not identifying a biologically known MAR is 62%. Analysis of the *Arabidopsis* genome with SMARTest [54] predicted a MAR approximately every 5.5 kb, a spacing similar to the intergenic distance in this species. The limited sequence diversity of the initial PWM training set as well as the selection of only AT-rich sequences, places a likely lower limit on the type II error possible with the first iteration of this approach. It is noteworthy that this software is under continuous development since its original description in the literature. The predictions made with the tool may be subtly altered over time reflecting the continued refinement of the weight matrices as experimental evidence accumulates.

MARSCAN

MARSCAN is part of the Emboss suite (http://emboss.sourceforge.net/). An online interactive version of the tool has become available at http://bioweb.pasteur.fr/seqanal/interfaces/marscan.html, while a windows executable can be found at http://perso.wanadoo.fr/ablavier/embosswin/embosswin.html. MARSCAN uses a similar strategy to the transcription factor module search to identify the bipartite MAR Recognition Signal (MRS) [53]. The MRS is comprised of two AT-rich motifs. They are represented by IUPAC sequence strings AWWRTAANNWWGNNNC of 16 bp and AATAAYAA of 8 bp. The two motifs must be present within 200 bp of each other in order for this region to be classified as containing a MAR. One would expect by chance alone that each pattern would occur once every 128 kb and once every 32 kb respectively in a sequence neutral strand of DNA. An alternate 15 bp IUPAC representation AWARTAANNAWGWNN, with somewhat greater sequence specificity can be derived from the original sequence data. Position weight matrices that provide a confidence weighted representation of the information content of sequences can also be generated from this data.

The Emboss implementation of MARSCAN relies on generic DNA sequence search algorithms. A string search based on a Boyer Moore strategy [59] with multiple possible mismatches can be undertaken to find the 16 bp sequence. This relatively CPU intensive approach is necessitated by both the sequence mismatch and ambiguity requirements. The shorter 8 base sequence can be searched through a more direct shift approach [60]. Even with the requirement for a relatively fuzzy search, the Emboss tool is well able to accommodate long sequences in a reasonable time (Tab. 2).

Tab. 2. MAR detection algorithms explored against the 4.7 Mb human Chromosome 6 MHC locus. Indicated are the times taken to complete the analysis (top row) and mean number of base pairs assigned to be matrix associated per MAR together with the number of distinct MARs (first column). In the body of the table are the total number of base pairs overlapping between algorithms and the number of overlapping MARs between algorithms. Also shown are the anticipated base pair overlaps from a random base selection of equal size relative to which is derived an enrichment metric.

	MarFinder (Time: 20m)	MarScan (Time: 2m**)	SMARTest (Time: 5m)	SIDD (Time: 20h)	ChrClass (Time: 18h**)
MarFinder (590 bp: 27 MARs)					
MarScan (65 bp: 447 MARs)	146 bp/4MARs [97 bp - 1.5]				
SMARTest (503 bp: 282 MARs)	4kb/9MARs [0.5 kb - 8]	4 kb/61MARs [866bp - 4.6]			
SIDD (44 bp: 1398 MARs)	511bp/7MARs [208 bp - 2.5]	527 bp/17MARs [378 bp - 1.4]	3.4 kb/75MARs [1.8 kb - 1.8]		
ChrClass* (860 bp:1521 MARs)	10 kb/23MARs [4.4 kb - 2.3]	11.6 kb/186MARs [7.9 kb - 1.5]	89 kb/220MARs [38 kb - 2.3]	19.3 kb/452MARs [17 kb - 1.1]	

*all predictions **time taken on a standard 3GHz windows desktop machine

In the absence of extended reiteration, the relative nucleosomal location of the two motifs should be of significance reflecting the formation of an *in vivo* protein complex. As shown in Fig. 2, when the frequency of the distance between the components of the MRS motif [53] is plotted, the distribution reflects the periodicity of the nucleosomally bound DNA. If the motifs are recognized by a single protein complex that binds ~15 bp as in the case of hnRNPa1 [38], then elements offset by ~60 - 80 bp could be on the same side of the nucleosome and aligned end to end. This periodicity of ~86 bp is denoted by the MRS signal immediately upstream of TAP1 indicated by the red arrows in Fig. 2 [61]. With a periodicity of ~155 - 200 bp the motifs could colocate with a slightly different orientation on the linker strand crossover. If colocation on

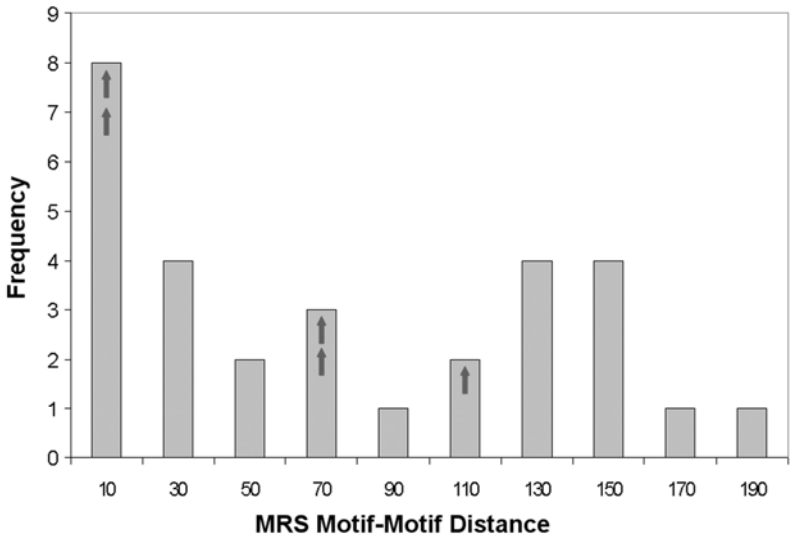

Fig. 2. Spacing between the MRS bipartite AWWRTAANNWWGNNNC and AATAAYAA motifs. The points represent the number of times the MRS bipartite signature with that spacing was identified within non-overlapping 20 base pair windows [52]. The spacing of the major histocompatibility complex MRS sites upstream of TAP1 corresponding to experimentally validated MARs are shown as red arrows [61]. The distribution is indicative of a preferential nucleosomal orientation for the MRS motifs on the 85 nucleotide turn around the nucleosome.

one side of the nucleosome is favored, then an offset around 0 bp would be preferable, as larger offsets would incrementally be more likely to be located on either a linker or neighboring nucleosome depending on the location of the first binding site. Indeed the colocation of several elements at the dyad center of the SV40 nucleosome and at entry and exit locations has been noted [53]. The two components of the motif may share similar nucleosomal positioning as above. MARSCAN could potentially be extended to better predict the *in-vivo* binding potential if a compound probability density function derived from the inter-motif distance is incorporated as part of the algorithm. The median MRS-MRS distance predicted with MARSCAN in the mammalian genome matches the median periodicity of genes within gene-clusters but not the much longer average intergenic distance outside of these clusters.

STRESS INDUCED DUPLEX DESTABILIZATION (SIDD)

SIDD has been continuously honed since its initial introduction in 1979 [62-65]. The approach models the distribution of torsional stress along a sequence, dividing this between a non-optimally wound helical structure and base unpaired regions. Identifying regions with a propensity for base unpairing can identify functionally specific sites of protein binding. For example, these base unpaired

regions are often coincident with sites of transfactor contact that require base unpairing for strand interaction. One of the most energetically favorable regions for strand destabilization is AATATATTT [66]. Depending on context AATATATTT could serve as an AT rich origin of replication. Accordingly, early replicating sequences have been associated with nuclear matrix binding by nascent strand excision [67, 68].

Applying or removing stress always changes the form of superhelical coiling in a manner that is partitioned between its two orthogonal curvature elements, the writhe Wr (number of superhelical turns) and the twist Tw (number of helical turns). In a continuous loop these sum to the structurally constant linking number Lk:

$$Lk = Wr + Tw \quad (1)$$

In a relaxed state, β-DNA forms a 10.4 base right handed slightly negatively supercoiled helix that repeats every 3.5 nm. This is a state optimally configured to wrap the strand of DNA around the nucleosome. The energy minima of the relaxed state Lk_0 is achieved with 70:30 Wr:Tw [69]. Any change in Tw that is not compensated by Wr or vice-versa, leads to an accumulation of torsional energy α typically applying a more negative superhelical stress:

$$\alpha = Lk - Lk_0 \quad (2)$$

Local sequence characteristics can be used to determine how this stress is then distributed along the strand when modeled under physiological conditions in which α/LK_0 can be taken as approximately -0.055 [70]. The process of distributing torsional energy and relieving stress takes place as a competition between sites. Exceeding a sequence-specific threshold initiates the process of destabilizing a duplex at a core unwinding element that is energetically predisposed to initiate base-pair dissociation. Once the energy has been released to split this unbound base pair, the region can readily undergo further destabilization. For a given temperature and salt concentration, the stiffness of both unpaired (freely coiled) and helical structures are experimentally known. Hence the free energy for all possible combinations of destabilized DNA can be calculated. The algorithm reports the propensity to occupy a denatured state for each nucleotide position along the sequence.

SIDD approaches

The initial implementation of SIDD [71] considered that any linking number other than Lk_0 develops additional helical stress that will be distributed between the torsional Tw component D_t and an experimentally determined residual Wr bending. The torsional difference D_t can either distort without destabilizing the helical structure T_{helix} or be rapidly dispersed in a denatured coiled structure T_{coil}:

$$D_t = \Delta T_{coil} + \Delta T_{helix} \quad (3)$$

It is evident that the energy dispersed through each structure is the product of the rate at which that structure releases excess torsional stress less the energy required to form that structure for a given number of bases. For a coiled state of

n unpaired bases accommodating a twist rate of R_c above that of the unstressed duplex, the released torsional deformation is:

$$\Delta T_{coil} = \frac{nR_c}{2\pi} - \frac{n}{10.4} \quad (4)$$

As SIDD was refined, the twist rate R_c was permitted to vary with base (k) along the base unpaired region as $R_{c(k)}$. The free energy of this denatured conformation has three components:

1. the energy required to melt n base pairs where the melting energy of the i'th base pair is B_i
2. the additional energy required to melt the first base pair in the sequence A_j,
3. the linearly elastic free energy required to maintain a limit thin molecule of coiled torsional stiffness S_c at an excess twist relative to the R_c over n bases.

Since A_j and B_i are known to trace complex relationships with respect to variables such as temperature, ionic concentration and nearest neighbor, a range of values derived under varying experimental conditions are used. In alternate approaches a single biologically based approximation replaces these terms. With A_j and B_i mapped, the free energy of the denatured DNA can be represented as:

$$E(n)_{coil} = \frac{nS_cR_c^2}{2} + A_j + \sum_{i=1}^{n} B_i \quad (5)$$

For non-destabilized helical DNA having a non-optimal twist rate R_h over n turns, twist is released at a rate of:

$$\Delta T_{helix} = \frac{nR_h}{2\pi} - \frac{n}{10.4} \quad (6)$$

For a retained helical conformation the free energy term does not require a melting energy component. Hence the free energy of a helix of excess twist rate R_h and a torsional stiffness S_h is represented as:

$$E(n)_{helix} = \frac{nS_hR_h^2}{2} \quad (7)$$

The energy states of a sequence of N bases in which n are in a coiled state and (N-n) remain in a helical state can now be explicitly calculated. Since the residual superhelicity within the helical region is the difference between the total excess twist and that which is not released in the coiled state, the residual excess helical twist can itself be expressed in terms of the length of the denatured region and the total twist to be relieved is:

$$R_h = \frac{(2\pi D_t + \frac{2\pi n}{A_i} - nR_c)}{(N-n)} \quad (8)$$

Substituting the R_h expression into equation (7) and introducing the free energy for n of N bases starting at base j being in a coiled state as derived from [5] yields:

$$E(R_c, n_j, N) = \frac{nS_cR_c^2}{2} + A_i + \sum_{i=j}^{n+j} B_i + \frac{S_h}{2(N-n)}\left[2\pi(D_t + \frac{n}{A_i}) - nR_c\right]^2 \quad (9)$$

The approach solves for R_c by partial differentiation since for a given n, an energy minima is found where $\frac{dE}{dR_c} = 0$. This permits substitution of R_c with an expression dependent solely on n_j, N, the imposed stress and the high torsional stiffness S_h (8.5×10^{-12} erg/radian^2) of the helix and more flexible S_c (3.6×10^{-13} erg/radian^2) of the coil:

$$E(n_j, N) = \frac{nS_c\left[\frac{2\pi S_h(D_t + \frac{n}{A_j})}{S_c(N-n) + S_h n}\right]^2}{2} + A_j + \sum_{i=j}^{n+j} B_i + \frac{S_h}{2(N-n)}\left[2\pi(D_t + \frac{n}{A_j}) - n\left[\frac{2\pi S_h(D_t + \frac{n}{A_j})}{S_c(N-n) + S_h n}\right]\right]^2 \quad (10)$$

Alternatively, the torsional stiffness can be replaced by residual superhelicity due to its relatively small energy contribution and the need to simplify the calculation of the partition function. In either case, three potential sequence states can be identified for a given set of bases and these are largely dependent upon the melting energies B and to a lesser extent A. When B is zero, a helical structure does not form. However, when B is excessive, the helical structure is never destabilized under biological conditions. Only when B lies between these values, as in physiological DNA, can a partially destabilized duplex form. The propensity of a sequence to base-unpair can then be determined by the distribution of possible energy states corresponding to varying values of A (~11 kcal/mol) and B (G≡C: 1.31 kcal/mol) B (A=T: 0.26 kcal/mol) along the sequence [72].

The distribution between the possible energy states $E(n_j,N)$ is modeled as a Boltzmann distribution. The extent to which physiologically constrained and actively maintained DNA will be free to adopt the states of an ideal Boltzmann distribution is interesting. It is then a computationally intensive task to determine the normalizing partition function for the Boltzmann distribution due to the unconstrained number of low occupancy higher energy levels. SIDD employs various simplifying summation schemes over the higher order elements to determine the partition function Z [70]. Once computed, this yields the probability p of any given state as a function of the energy of that state, temperature T and the Boltzmann constant k:

$$P(E(n_j, N)) = \frac{1}{Z} e^{\frac{-E(n_j, N)}{kT}} \quad (11)$$

Biologically normal salt and temperature conditions are introduced and the distribution integrated to determine the relative proportion of states lying above the dissociation threshold for each base in a queried sequence.

SIDD has been refined to include a more rigorous analysis of the partition function [72] as well as corrections for local stacking [73] energies. It is clear that a SIDD strategy will identify functional binding sites as regions with a propensity for base unpairing. These will include a multitude of elements, e.g., MARs and transcription factors. Given the ongoing development of SIDD, it is valuable to date each profile produced with tools such as WebSIDD [63]. While SIDD identifies DNA conformations that likely approach a base unpaired state, other DNA conformations such as curved and kinked DNA have also been shown by statistical analysis to be bound by matrix proteins. The tools to model these and other conformations could likely be enhanced through sequence affinity and destabilization strategies.

ChrClass

ChrClass [74, 75] (ftp://ftp.bionet.nsc.ru/pub/biology/chrclass/chrclass.zip) uses motif based supervised machine learning to develop a MAR classification function. IUPAC representations of potentially classifying motifs are then assessed relative to known S/MAR binding sites through a linear discriminant approach. Many of the motifs used are similar to those detected by MarFinder (Tab. 1). To form the initial training set, 27 animal and plant sequences were collected from the literature. In addition, 16 sequences bound to the protein cores of central rosette-like features of interphase chromosomes, along with 35 bound to the synaptonemal complex, 25 associated with the nuclear lamina and 24 from 1.5 kb 5' promoter regions of facultatively expressed genes were included. The length of each sequence was matched using 116 randomly generated base neutral sequences. The Linear Discriminant Analysis (LDA) based classification was then derived on the basis of the partitioning of the set of published consensus MAR motifs relative to these training sequences.

To generate a classification function with maximal separation, the published classification motifs were evaluated and where predictive power relative to the training sequences was found, used as the independent variables of the classification function with correlation weights b. These independent variables were combined in the ChrClass software to predict the dependent group classifications h for each of a set of tested sequences (equation 12). The training step of an LDA can take various forms and several classification functions can be chosen. However, the assumed prior distribution in the standard Bayesian approach is a variance standardized normal or normal-transformed distribution. The independent variables are weighted to optimize the ratio of the between class variance to within class variance with a view to approaching complete

linear separation. The approach is particularly suited to MAR analysis, as the classifying sequences are not required to be orthogonal. This approach has the potential to recognize some of the mutual information between classifiers. The canonical root function resembles the integration of the vector input subspaces into the most significant compound (principal) axis generated by:

$$h_k = b_{k0} + b_{k1}X_1 + \ldots + b_{kn}X_n \quad (12)$$

Where $X_{1\ldots n}$ are the frequencies of the training motifs in the tested sequences and b_{kn} are the weights for sequence k with respect to sub-sequences 1 to N. Given that the training functions (Tab. 1) are generally AT-rich and hence not orthogonal, the parameters that represent mutual sequence entropy and variance would generally not be simple Euclidian distances, i.e., the estimated covariance matrix elements a_{ij} would not be diagonal:

$$b_{ki} = (N - K)\sum_{j=1}^{n} a_{ij}X_{jk} \quad (13)$$

ChrClass uniquely predicts the nuclear sub-localization of the MAR on the basis of the differences noted in the posterior distributions between several training sequences. For example, rosette-structure (CS) DNA was most readily classified while lamin binding DNA was least readily classified. This is consistent with observations that lamin binding is driven by higher order DNA structure and hence not well represented by simple sub-sequences. The prediction quality of ChrClass is also returned as either a confidence measure for the classification or an indicator of potential binding affinity.

While providing a more powerful approach than a simple presence absence test, ChrClass remains limited in the extent that it can model the subtle matrix of cross interactions between candidate MAR sequences that a Markov or ANN approach could incorporate [76]. Given the rapid pace at which the literature describing the modes of matrix binding is growing, the potential for further development of MAR detection using linear, quadratic and flexible discriminant approaches is evident.

MarFinder

MarFinder [77] (http://www.futuresoft.org/MarFinder/), commercialized as MarWiz, uses a semi-supervised approach to MAR detection. The original motifs that define the MAR sequences as well as the higher order DNA structures were derived from the manual curation of the literature. The motifs are represented as IUPAC subsequences that are combined in Boolean sets to form rules. However, the exact combination of the motifs contributing to a MAR is not defined, making the search approach partially unsupervised. MarFinder is essentially a rule density scanner with a correction introduced for local genomic nucleotide bias based on the probability of randomly detecting one of the target sequences. Tab. 1 compares the core motifs employed by MarFinder alongside those of the other strategies. There are 19 core MarFinder motifs that are combined to form 6 core MAR detection rules. A further 20 motifs are under

development for inclusion in MarWiz. These are represented by 11 augmented rules [78]. Implementation has met with tempered enthusiasm by the lack of understanding their interdependence.

The probability of the background occurrence of each motif relative to the local base sequence bias is first calculated. Accordingly, in an unbiased random sequence the motif AATT would have a 1/256 (0.004) chance of being present, while in a sequence where 40% of bases are A and 20% T, the probability would be 0.4^2x0.2^2 (0.006). Each rule is then defined as a probabilistic model created by combining the background probabilities of discovering the matching motifs.

A 1000 bp sliding window with a 100 bp step was adopted as the default, since MAR binding regions are anticipated to range from 100 – 500 bp. The user is able to set various parameters including the rules used, the p value threshold for a MAR (default: 0.7), as well as the number of consecutively detected windows required to indicate the presence of a MAR (default: 3). For each window of length W the background sequence bias is calculated and hence for each rule i the p_i of its random occurrence is calculated. The observed rule detection probabilities relative to this background are combined as independent Poisson processes with parameter λ_i given by p_iW. Hence the probability P of observing a set of $f_{1...k}$ matches to k rules each with a bias corrected probability of occurring over the W bases of $\lambda_{1...k}$ is:

$$P = \prod_{i=1}^{k} \frac{e^{-\lambda_i} \lambda_i^{f_i}}{f_i!} \tag{14}$$

The MAR potential is defined as $\log(1/\alpha)$ where α is the probability of rejecting the null hypothesis given x_i instances of each rule is:

$$\alpha = \sum_{x_1=f_1}^{\infty} \frac{e^{-\lambda_1} \lambda_1^{x_1}}{x_1!} \cdot \sum_{x_2=f_2}^{\infty} \frac{e^{-\lambda_2} \lambda_2^{x_2}}{x_2!} \cdot \cdots \cdot \sum_{x_m=f_m}^{\infty} \frac{e^{-\lambda_m} \lambda_m^{x_m}}{x_m!} \tag{15}$$

Several elements require consideration. It may be anticipated that sequences that fall below the MAR detection threshold in ChrClass will be reported as candidates by MarFinder, since support will be derived from multiple rules matching the same underlying sequence. For example, multiple ORI (origin of replication initiation) sequences will also be identified by the AT rich rule. A model of independent Poisson processes requires a high degree of rule independence. In its absence, an interdependence or mixing function is required. Clearly this criteria is breached for a set of short largely AT-rich motifs. Hence, a meta or compound Poisson process that introduces mixing density functions may be useful.

DISCUSSION

The *in-silico* strategies described above include many common elements in part due to the literature they share. As summarized in Tab. 1, SMARTests' PWMs will likely reflect to a certain extent the IUPAC sequences used by both ChrClass and MARFinder. Similarly, SIDD's core base unpairing regions can be

approximated by AATATT and ATATTT, corresponding to MARFinder's origin of replication initiation motif. The DNAse I sensitive purine-pyrimidine-tract $(RY)_n$ found in DNA triplex structures is used by ChrClass to represent the core topoisomerase II binding signal that is represented by a consensus binding motif in MARFinder. Like MARFinder, SMARTest uses a pre-determined threshold density of classifier matches to the sequence. The incorporation of a statistical distance correction term and a cross-interaction matrix based on observed coincidence or anti-coincidence of MAR sequences may extend these approaches.

Several studies have been pursued to biologically evaluate nuclear matrix binding relative to *in-silico* predictions [31, 79]. One of the first studies compared the ability of MARFinder, MARSCAN and SMARTest to accurately identify the MARs encompassing the human beta-globin domain [31]. All of the programs tested, over-predicted the sites of attachment. The significance of this discordance is not immediately obvious since the sites of nuclear matrix attachment physically changed depending on cell type and expression status. This exemplifies the complex nature of the nuclear matrix detection problem. Using these and other *in-silico* approaches and again tolerating a substantial over-prediction, Purbowasito aligned 95% of the experimentally validated MARs [79] with at least one of the predictions. These and other studies have consistently revealed a residual variance even after the substantial over-prediction between the *in-silico* predictions and the *in-vivo* evidence.

Tab. 2, summarizes the extent of inherent concordance between algorithms employed with their default parameters. All were tested on a 4.7 Mb segment of human chromosome 6 that incorporates the major histocompatibility complex (Unigene Build 35.1 Chr6:28,804,582-33,559,407). The left hand column describes the average MAR length per algorithm and total number of MARs discovered. The body of the table records the total length of overlapping predictions and number of mutually distinct overlapping predictions. Given the number of base pairs predicted by each pair of algorithms, the enrichment in overlap between predictions relative to a completely random base selection of equal length is noted. The column header row records the time taken to complete the analysis from which it is clear that only a subset of the approaches are currently suitable for rapid chromosome or genome-level analysis. Of the 282 SMARTest predictions, 220 overlapped at least in part by the 1521 predictions made by ChrClass. Equally, 4 kb of the 16 kb predictions made by MARFinder (MarWiz) overlap the 142 kb of predictions made by SMARTest. In contrast, the predictions made by SIDD relative to those of ChrClass and MARSCAN overlap only slightly. This occurs more often than would be expected by chance alone. That the extent of concordance reflects the approaches to selecting training sequences is beyond doubt. However, since all the approaches are ultimately derived from an evidential base, it raises the issue as to whether multiple and in some cases relatively well differentiated classes of MARs are being predicted. For example, SIDD and MARSCAN predictions are derived from essentially unrelated assumptions. Their oligonucleotide length sequence predictions

markedly contrast the requirements made by MarFinder for the extensive classifier reiteration necessary to define MAR sequences. The extent of both over-prediction and algorithm disconcordance supports the conclusion that further algorithmic development may prove beneficial.

Several promising routes are already being explored. To address over-prediction, further iterations of SIDD are anticipated that will better differentiate among the different binding features through their iteration profiles. The SIDD approach will nonetheless remain limited by the standard modeling factors. It is inherently a simplification relative to regional topological constraints outside of the modeling parameters. Some of these exogenous factors as well as the subtle electrostatics of thermally perturbed DNA are however already being incorporated in other work [80].

There may well be evidence to suggest that not only are MAR domains differentiated, but that the forms of MAR binding predicted by the algorithms is multi-level. This would extend the notion of biological differentiation around facultative and constitutive MARs and open up the potential for algorithms to target features located at different levels of a hierarchy. This is a modality that has been proposed elsewhere from diverse evidence [32, 81]. Indeed some tools, notably ChrClass, implicitly define a hierarchy by noting different values of n, i.e., the motif reiteration required to demonstrate a MAR. The MRS would lie at the base of such hierarchy, resembling the transcription factor binding site. Major constitutive chromosome organizing sites that coordinate the chromatin domains on a broad scale would require much longer clusters of matrix attachment regions in order to be readily targeted in decondensing chromatin. Such stratified functional differentiation would contribute to the long standing debate over the validity of simple loop models. If a stratified model is accepted, then the compound *in-silico* approaches to MAR identification will require further development to associate the form of MAR binding to the extent of reiteration, fuzziness and ultimately to broader genomic neighborhood features.

Acknowledgements. This work was supported by NICHD grant HD-36512 to S.A.K. Predoctoral fellowship support to A.K.Q. from the American Heart Association supported by grant 0515441Z is gratefully acknowledged.

REFERENCES

1. Boulikas, T. Nature of DNA sequences at the attachment regions of genes to the nuclear matrix. **J. Cell. Biochem.** 52 (1993) 14-22.
2. Fawcett, D.W. On the occurrence of a fibrous lamina on the inner aspect of the nuclear envelope in certain cells of vertebrates. **Am. J. Anat.** 119 (1966) 129-145.
3. He, D., Zeng, C. and Brinkley, B.R. Nuclear matrix proteins as structural and functional components of the mitotic apparatus. **Int. Rev. Cytol.** 162B (1995) 1-74.

4. Stadler, S., Schnapp, V., Mayer, R., Stein, S., Cremer, C., Bonifer, C., Cremer, T and Dietzel, S. The architecture of chicken chromosome territories changes during differentiation. **BMC Cell Biol.** 5 (2004) 44.
5. Cremer, T. and Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. **Nat. Rev. Genet.** 2 (2001) 292-301.
6. Gotzmann, J. and Foisner, R. Lamins and lamin-binding proteins in functional chromatin organization. **Crit. Rev. Eukaryot. Gene Expr.** 9 (1999) 257-265.
7. Mika, S. *NMP-db* Available from: http://cubic.bioc.columbia.edu/db/nmpdb/.
8. Capco, D.G., Wan, K.M and Penman, S. The nuclear matrix: three-dimensional architecture and protein composition. **Cell** 29 (1982) 847-858.
9. Renz, A. and Fackelmayer, F.O. Purification and molecular cloning of the scaffold attachment factor B (SAF-B), a novel human nuclear protein that specifically binds to S/MAR-DNA. **Nucleic Acids Res.** 24 (1996) 843-849.
10. Kas, E. and Laemmli, U.K. In vivo topoisomerase II cleavage of the Drosophila histone and satellite III repeats: DNA sequence and structural characteristics. **Embo. J.** 11 (1992) 705-716.
11. Dickinson, L.A., Joh, T., Kohwi, Y. and Kohwi-Shigematsu, T. A tissue-specific MAR/SAR DNA-binding protein with unusual binding site recognition. **Cell** 70 (1992) 631-645.
12. Zong, R.T. and Scheuermann, R.H. Mutually exclusive interaction of a novel matrix attachment region binding protein and the NF-muNR enhancer repressor. Implications for regulation of immunoglobulin heavy chain expression. **J. Biol. Chem.** 270 (1995) 24010-24018.
13. Yusufzai, T.M. and Felsenfeld, G. The 5'-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element. **Proc. Natl. Acad. Sci. U. S. A.** 101 (2004) 8620-8624.
14. Steinert, P.M. and Roop, D.R. Molecular and cellular biology of intermediate filaments. **Annu. Rev. Biochem.** 57 (1988) 593-625.
15. Rando, O.J., Zhao, K and Crabtree, G.R. Searching for a function for nuclear actin. **Trends Cell Biol.** 10 (2000) 92-97.
16. He, D.C., Nickerson, J.A and Penman, S. Core filaments of the nuclear matrix. **J. Cell Biol.** 110 (1990) 569-580.
17. Narayan, K.S., Steele, W.J., Smetana, K and Busch, H. Ultrastructural aspects of the ribonucleo-protein network in nuclei of Walker tumor and rat liver. **Exp. Cell Res.** 46 (1967) 65-77.
18. Ma, H., Siegel, A.J and Berezney, R. Association of chromosome territories with the nuclear matrix. Disruption of human chromosome territories correlates with the release of a subset of nuclear matrix proteins. **J. Cell Biol.** 146 (1999) 531-542.
19. Miralles, F., Ofverstedt, L.G., Sabri, N., Aissouni, Y., Hellman, U., Skoglund, U and Visa, N. Electron tomography reveals posttranscriptional binding of pre-mRNPs to specific fibers in the nucleoplasm. **J. Cell Biol.** 148 (2000) 271-282.

20. Jackson, D.A. and Cook, P.R. Visualization of a filamentous nucleoskeleton with a 23 nm axial repeat. **EMBO. J.** 7 (1988) 3667-3677.
21. Earnshaw, W.C. and Heck, M.M. Localization of topoisomerase II in mitotic chromosomes. **J. Cell Biol.** 100 (1985) 1716-1725.
22. Glazkov, M.V., Poltaraus, A.B. and Lebedeva, I.A. Nucleotide sequence of DNA isolated from protein cores of rosette-like structures (elementary chromomeres) of mouse interphase chromosomes. **Genetika** 30 (1994) 1146-1154.
23. Prusov, A.N., Poliakov, V., Zatsepina, O.V., Fais, D. and Chentsov Iu, S. Isolation of rosette-like structures from partially deproteinized chromatin in rat hepatocytes. **Tsitologiia** 27 (1985) 1026-1030.
24. van Driel, R. and Fransz, P. Nuclear architecture and genome functioning in plants and animals: what can we learn from both? **Exp. Cell Res.** 296 (2004) 86-90.
25. Ierardi, L.A., Moss, S.B. and Bellve, A.R. Synaptonemal complexes are integral components of the isolated mouse spermatocyte nuclear matrix. **J. Cell Biol.** 96 (1983) 1717-1726.
26. Gautier, T., Robert-Nicoud, M., Guilly, M.N. and Hernandez-Verdun, D. Relocation of nucleolar proteins around chromosomes at mitosis. A study by confocal laser scanning microscopy. **J. Cell Sci.** 102 (Pt 4) (1992) 729-737.
27. Hernandez-Verdun, D. and Gautier, T. The chromosome periphery during mitosis. **Bioessays** 16 (1994) 179-185.
28. Berezney, R., Mortillaro, M.J., Ma, H., Wei, X. and Samarabandu, J. The nuclear matrix: a structural milieu for genomic function. **Int. Rev. Cytol.** 162A (1995) 1-65.
29. Abney, J.R., Cutler, B., Fillbach, M.L., Axelrod, D. and Scalettar, B.A. Chromatin dynamics in interphase nuclei and its implications for nuclear structure. **J. Cell Biol.** 137 (1997) 1459-1468.
30. Vogelstein, B., Pardoll, D.M. and Coffey, D.S. Supercoiled loops and eucaryotic DNA replicaton. **Cell** 22 (1980) 79-85.
31. Ostermeier, G.C., Liu, Z., Martins, R.P., Bharadwaj, R.R., Ellis, J., Draghici, S. and Krawetz, S.A. Nuclear matrix association of the human beta-globin locus utilizing a novel approach to quantitative real-time PCR. **Nucleic Acids Res.** 31 (2003) 3257-3266.
32. Mielke, C., Kohwi, Y., Kohwi-Shigematsu, T. and Bode, J. Hierarchical binding of DNA fragments derived from scaffold-attached regions: correlation of properties in vitro and function in vivo. **Biochemistry** 29 (1990) 7475-7485.
33. Kramer, J.A., Adams, M.D., Singh, G.B., Doggett, N.A. and Krawetz, S.A. Extended analysis of the region encompassing the PRM1-->PRM2-->TNP2 domain: genomic organization, evolution and gene identification. **J. Exp. Zool.** 282 (1998) 245-253.
34. Labrador, M. and Corces, V.G. Setting the boundaries of chromatin domains and nuclear organization. **Cell** 111 (2002) 151-154.

35. Williams, R.R. Transcription and the territory: the ins and outs of gene positioning. **Trends Genet.** 19 (2003) 298-302.
36. Heun, P., Laroche, T., Shimada, K., Furrer, P and Gasser, S.M. Chromosome dynamics in the yeast interphase nucleus. **Science** 294 (2001) 2181-2186.
37. Melcak, I., Cermanova, S., Jirsova, K., Koberna, K., Malinsky, J. and Raska, I. Nuclear pre-mRNA compartmentalization: trafficking of released transcripts to splicing factor reservoirs. **Mol. Biol. Cell** 11 (2000) 497-510.
38. Donev, R.M., Doneva, T.A., Bowen, W.R. and Sheer, D. HnRNP-A1 binds directly to double-stranded DNA in vitro within a 36 bp sequence. **Mol. Cell. Biochem.** 233 (2002) 181-185.
39. Cremer, T., Kupper, K., Dietzel, S and Fakan, S. Higher order chromatin architecture in the cell nucleus: on the way from structure to function. **Biol. Cell** 96 (2004) 555-567.
40. Krawetz, S.A., Draghici, S., Goodrich, R., Liu, Z and Ostermeier, G.C., *In Silico* and wet-bench identification of nuclear matrix attachment regions. in: **Hypertension, Methods and Protocols** (Fennell, J.P., Baker, A.H., Eds.), Vol. 108, Humana Press, 2004, 439-458.
41. Bode, J., Stengert-Iber, M., Kay, V., Schlake, T and Dietz-Pfeilstetter, A. Scaffold/matrix-attached regions: topological switches with multiple regulatory functions. **Crit. Rev. Eukaryot. Gene Expr.** 6 (1996) 115-138.
42. Kramer, J.A., McCarrey, J.R., Djakiew, D and Krawetz, S.A. Differentiation: the selective potentiation of chromatin domains. **Development** 125 (1998) 4749-4755.
43. Gerasimova, T.I. and Corces, V.G. Boundary and insulator elements in chromosomes. **Curr. Opin. Genet. Dev.** 6 (1996) 185-192.
44. Cook, P.R. The organization of replication and transcription. **Science** 284 (1999) 1790-1795.
45. Leonhardt, H., Rahn, H.P., Weinzierl, P., Sporbert, A., Cremer, T., Zink, D and Cardoso, M.C. Dynamics of DNA replication factories in living cells. **J. Cell Biol.** 149 (2000) 271-280.
46. Strissel, P.L., Espinosa, R., III, Rowley, J.D and Swift, H. Scaffold attachment regions in centromere-associated DNA. **Chromosoma** 105 (1996) 122-133.
47. Lammerding, J., Schulze, P.C., Takahashi, T., Kozlov, S., Sullivan, T., Kamm, R.D., Stewart, C.L and Lee, R.T. Lamin A/C deficiency causes defective nuclear mechanics and mechanotransduction. **J. Clin. Invest.** 113 (2004) 370-378.
48. Vorlickova, M., Chladkova, J., Kejnovska, I., Fialova, M and Kypr, J. Guanine tetraplex topology of human telomere DNA is governed by the number of (TTAGGG) repeats. **Nucleic Acids Res.** 33 (2005) 5851-5860.
49. Moyzis, R.K., Buckingham, J.M., Cram, L.S., Dani, M., Deaven, L.L., Jones, M.D., Meyne, J., Ratliff, R.L and Wu, J.R. A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. **Proc. Natl. Acad. Sci. U.S.A.** 85 (1988) 6622-6626.

50. Lobov, I.B., Tsutsui, K., Mitchell, A.R and Podgornaya, O.I. Specificity of SAF-A and lamin B binding in vitro correlates with the satellite DNA bending state. **J. Cell. Biochem.** 83 (2001) 218-229.
51. Frisch, M., Frech, K., Klingenhoff, A., Cartharius, K., Liebich, I and Werner, T. In silico prediction of scaffold/matrix attachment regions in large genomic sequences. **Genome Res.** 12 (2002) 349-354.
52. Singh, G.B., Kramer, J.A and Krawetz, S.A. Mathematical model to predict regions of chromatin attachment to the nuclear matrix. **Nucleic Acids Res.** 25 (1997) 1419-1425.
53. van Drunen, C.M., Sewalt, R.G., Oosterling, R.W., Weisbeek, P.J., Smeekens, S.C and van Driel, R. A bipartite sequence element associated with matrix/scaffold attachment regions. **Nucleic Acids Res.** 27 (1999) 2924-2930.
54. Rudd, S., Frisch, M., Grote, K., Meyers, B.C., Mayer, K and Werner, T. Genome-wide in silico mapping of scaffold/matrix attachment regions in Arabidopsis suggests correlation of intragenic scaffold/matrix attachment regions with gene expression. **Plant Physiol.** 135 (2004) 715-722.
55. Morgenstern, B., Dress, A and Werner, T. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. **Proc. Natl. Acad. Sci. U.S.A.** 93 (1996) 12098-12103.
56. Needleman, S.B. and Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **J. Mol. Biol.** 48 (1970) 443-453.
57. Quandt, K., Frech, K., Karas, H., Wingender, E and Werner, T. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. **Nucleic Acids Res.** 23 (1995) 4878-4884.
58. Wolfertstetter, F., Frech, K., Herrmann, G and Werner, T. Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. **Comput. Appl. Biosci.** 12 (1996) 71-80.
59. Tarhio, J. and Ukkonen, E. *Approximate Boyer-Moore String Matching*. 2005; Available from: http://www.cs.hut.fi/~tarhio/papers/abm.ps.gz.
60. RA Baeza-Yates, G.G. Fast text searching for regular expressions or automaton searching on tries. **Journal of the A.C.M.** 43 (1996) 915-936.
61. Donev, R., Horton, R., Beck, S., Doneva, T., Vatcheva, R., Bowen, W.R and Sheer, D. Recruitment of heterogeneous nuclear ribonucleoprotein A1 in vivo to the LMP/TAP region of the major histocompatibility complex. **J. Biol. Chem.** 278 (2003) 5214-5226.
62. Wang, H., Noordewier, M and Benham, C.J. Stress-induced DNA duplex destabilization (SIDD) in the E. coli genome: SIDD sites are closely associated with promoters. **Genome Res.** 14 (2004) 1575-1584.
63. Bi, C. and Benham, C.J. WebSIDD: server for predicting stress-induced duplex destabilized (SIDD) sites in superhelical DNA. **Bioinformatics** 20 (2004) 1477-1479.

64. Benham, C.J. and Bi, C. The analysis of stress-induced duplex destabilization in long genomic DNA sequences. **J. Comput. Biol.** 11 (2004) 519-543.
65. Benham, C., Kohwi-Shigematsu, T and Bode, J. Stress-induced duplex DNA destabilization in scaffold/matrix attachment regions. **J. Mol. Biol.** 274 (1997) 181-196.
66. Bode, J., Kohwi, Y., Dickinson, L., Joh, T., Klehr, D., Mielke, C and Kohwi-Shigematsu, T. Biological significance of unwinding capability of nuclear matrix-associating DNAs. **Science** 255 (1992) 195-197.
67. Vassetzky, Y.S., Bogdanova, A.N and Razin, S.V. Analysis of the chicken DNA fragments that contain structural sites of attachment to the nuclear matrix: DNA-matrix interactions and replication. **J. Cell. Biochem.** 79 (2000) 1-14.
68. Girard-Reydet, C., Gregoire, D., Vassetzky, Y and Mechali, M. DNA replication initiates at domains overlapping with nuclear matrix attachment regions in the xenopus and mouse c-myc promoter. **Gene** 332 (2004) 129-138.
69. Beaudouin, J., Gerlich, D., Daigle, N., Eils, R and Ellenberg, J. Nuclear envelope breakdown proceeds by microtubule-induced tearing of the lamina. **Cell** 108 (2002) 83-96.
70. Benham, C.J. Stress-induced DNA duplex destabilization in transcriptional initiation. **Pac. Symp. Biocomput.** (2001) 103-114.
71. Benham, C.J. Torsional stress and local denaturation in supercoiled DNA. **Proc. Natl. Acad. Sci. U.S.A.** 76 (1979) 3870-3874.
72. Fye, R.M. and Benham, C.J. Exact method for numerically analyzing a model of local denaturation in superhelically stressed DNA. **Phys. Rev. E.** 59 (1999) 3408-3426.
73. Chengpeng Bi , C.J.B. The approximate algorithm for analysis of the strand separation transition in super helical DNA using nearest neighbor energetics. **Proceedings of the IEEE Computer Society Conference on Bioinformatics** (2003) 460.
74. Rogozin, I.B., Glazko, G.V and Glazkov, M.V. Computer prediction of sites associated with various elements of the nuclear matrix. **Brief. Bioinform.** 1 (2000) 33-44.
75. Glazko, G.V., Rogozin, I.B and Glazkov, M.V. Comparative study and prediction of DNA fragments associated with various elements of the nuclear matrix. **Biochim. Biophys. Acta** 1517 (2001) 351-364.
76. Baldi, P. and Brunak, S. **Bioinformatics: the machine learning approach: Adaptive computation and machine learning.** (Dietterich, T., Ed.), 2nd edition, MIT Press, Cambridge, Mass., 2001, 1-452.
77. Kramer, J.A., Adams, M.D., Singh, G.B., Doggett, N.A and Krawetz, S.A. A matrix associated region localizes the human SOCS-1 gene to chromosome 16p13.13. **Somat. Cell Mol. Genet.** 24 (1998) 131-133.

78. Singh, G.B. and Krawetz, S.A., Data Mining for Discovering Matrix Association Regions (MARs). in: **Data mining and knowledge discovery: theory, toolsand technology II.** (Dasarathy, B.V.E., Ed.), Proceedings of Spie, (2000) 330-341.
79. Purbowasito, W., Suda, C., Yokomine, T., Zubair, M., Sado, T., Tsutsui, K and Sasaki, H. Large-scale identification and mapping of nuclear matrix-attachment regions in the distal imprinted domain of mouse chromosome 7. **DNA Res.** 11 (2004) 391-407.
80. Ubbink, J. and Odijk, T. Electrostatic-undulatory theory of plectonemically supercoiled DNA. **Biophys. J.** 76 (1999) 2502-2519.
81. Belmont, A.S., Sedat, J.W and Agard, D.A. A three-dimensional approach to mitotic chromosome structure: evidence for a complex hierarchical organization. **J. Cell Biol.** 105 (1987) 77-92.