Journal of Cellular Biochemistry 84:484–496 (2002)

Multiple Collagen I Gene Regulatory Elements Have Sites of Stress-Induced DNA Duplex Destabilization and Nuclear Scaffold/Matrix Association Potential

Christian Mielke,[1] Morten O. Christensen,[1] Ole Westergaard,[1] Jürgen Bode,[2] Craig J. Benham,[3] and Michael Breindl[4]*

[1]Department of Structural and Molecular Biology, University of Aarhus, Denmark
[2]GBF, National Center for Biotechnological Research, Braunschweig, Germany
[3]Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, New York
[4]Department of Biology and Molecular Biology Institute, San Diego State University, San Diego, California

Abstract The availability of the complete nucleotide sequences of numerous prokaryotic and eukaryotic organisms should stimulate the development and application of computer-based approaches for studying genome organization and function. Earlier work has shown that distinct regulatory DNA elements can be identified by computational analysis as sites of stress-induced DNA duplex destabilization (SIDD). Here we report the results of computational and experimental analyses of previously identified regulatory elements in the murine α1(I) collagen (*Col1a1*) gene domain. We found that several distal 5′ DNase I-hypersensitive sites (HSs) which function in the chromatin loop organization of the *Col1a1* gene are characterized by strongly destabilized SIDD profiles. Elements in the proximal 5′ promoter and first intron which differentially regulate *Col1a1* promoter activity in different collagen-producing cell types also contain SIDD sites. All 5′ elements associated with destabilized sites are shown to have nuclear matrix binding activity in an in vitro binding assay. Other putative regulatory elements in the transcribed and 3′-flanking regions of the *Col1a1* gene, including both of its polyadenylation sites, are also associated with SIDD peaks. The human *COL1A1* gene has periodic SIDD peaks within the transcribed region, suggesting that abundantly expressed genes may require SIDDs acting as topological sinks during transcription. The 5′ ends of the murine *Col1a1* and the homologous human gene revealed similar SIDD profiles, but limited DNA sequence similarity, indicating that some DNA functions are evolutionarily conserved by preserving higher order DNA structural properties rather than nucleotide sequence. Our results show that destabilized SIDD profiles are a common feature of eukaryotic regulatory DNA elements with such diverse functions as chromatin organization, cell-specific transcriptional enhancement, and initiation and termination of transcription. They demonstrate the usefulness of computational analyses that predict SIDD properties in reliably identifying DNA elements involved in the structural organization of the eukaryotic genome and the regulation of its expression. J. Cell. Biochem. 84: 484–496, 2002.

Key words: α1(I) collagen gene domain; DNase I-hypersensitive sites; regulatory DNA elements; stress-induced DNA duplex destabilization; nuclear matrix binding sites

Eukaryotic genomes are thought to be organized into independently regulated chromatin loop domains which contain coding and non-coding sequences as well as *cis*-regulatory DNA elements important for the regulation of DNA replication and transcription. Some of these elements are located in close proximity to the coding sequences, whereas others can be tens of thousands of base pairs (bp) away. Relatively little is known about the molecular mechanisms by which proximal and distal regulatory elements cooperatively regulate stage-and tissue-specific gene activity. One reason for this is that

Grant sponsor: Danish Cancer Society; Grant numbers: 97-100-32, 97-143-09-9132; Grant sponsor: NIH; Grant numbers: RO1-GM47012, AR41909; Grant sponsor: NSF; Grant number: DBI-99-04549; Grant sponsor: Deutsche Forschungsgemeinschaft; Grant number: Bo 419/6-2.

Christian Mielke's present address is Department of Clinical Chemistry, Medizinische Poliklinik, University of Würzburg, Germany.

Craig J. Benham's present address is Genome Center, University of California, Davis, USA.

*Correspondence to: Michael Breindl, Department of Biology and Molecular Biology Institute, San Diego State University, San Diego, CA 92182.
E-mail: mbreindl@sunstroke.sdsu.edu

Received 17 July 2001; Accepted 21 September 2001

DOI 10.1002/jcb.10034

the identification and functional analysis of distal regulatory elements is tedious and time-consuming. Conventionally, regulatory elements are identified in laborious analyses as DNase I-hypersensitive sites (HSs) or other sites of unusual chromatin structure before they can be analyzed in reporter gene constructs in transfection experiments or transgenic animals. The availability of the complete nucleotide sequences of numerous prokaryotic and eukaryotic organisms, including the human genome (International Human Genome Sequencing Consortium, 2001), should now allow the development of fast, computer-based approaches to identify different kinds of regulatory elements and facilitate their functional analysis.

The various processes involved in nucleic acid metabolism must occur in a spatially ordered manner within the eukaryotic nucleus. There is evidence that the nucleus has a distinct substructure, that interphase chromosomes occupy discrete territories and that DNA transcription and replication occur in defined "speckles" or "factories" [for recent reviews see Berezney and Wei, 1998; Lamond and Earnshaw, 1998; Cook, 1999; Stein et al., 1999]. However, the exact nature of the intranuclear architecture necessary for these processes to occur is at present not well understood. A large body of evidence supports the notion that eukaryotic chromatin interacts with replication, transcription, and processing machineries through special DNA sequences called nuclear scaffold or matrix attachment regions (S/MARs). S/MARs are operationally defined as DNA sequences that can bind to the nuclear scaffold or matrix, the residual structure remaining after high salt extraction of nuclei [Berezney and Coffey, 1974] or after nuclear extraction with the mild detergent lithium 3,5-diiodosalicylate (LIS) at physiological ionic strength [Mirkovitch et al., 1984]. There appear to be at least two types of S/MARs: constitutive S/MARs are usually extended sequences which are located at the ends of DNase I-sensitive chromatin domains and can function both as boundary elements and to protect genes from position effects [Stief et al., 1989; Phi-van et al., 1990; reviewed in Bode et al., 1998]. Shorter S/MARs occur within transcribed regions, for example in introns of the human immunoglobulin (Ig) k and μ genes where they are found in close proximity to transcriptional enhancer elements and affect DNA methylation and transcription [Kirillov et al., 1996; Jenuwein et al., 1997; Oancea et al., 1997; Forrester et al., 1999; Dang et al., 2000].

Several attempts have been made to identify the DNA sequence requirements of S/MARs. While no readily identifiable consensus sequences have been found, several motifs have been biochemically characterized, each of which occurs in some, but not in all S/MARs [reviewed in Boulikas, 1995]. The observation that several of the S/MAR-associated motifs, such as AT-rich regions and topoisomerase II binding sites, could affect higher order DNA structure, together with the fact that specific consensus sequences are not found, suggests that topological or structural properties could determine S/MAR activity. In particular, S/MARs have been demonstrated to contain base-unpairing regions (BURs) [Kohwi-Shigematsu and Kohwi, 1990; Bode et al., 1992; Paul and Ferl, 1993]. Previous studies have shown that such topological properties of DNA are amenable to computational analysis and that S/MARs coincide precisely with sites of predicted extensive stress-induced DNA duplex destabilization (SIDD) [Benham, 1993, 1996; Benham et al., 1997].

We are using the murine α1 type I collagen (*Col1a1*) gene as a model system to investigate molecular mechanisms of gene regulation. The type I collagen genes are large genes with > 50 exons and introns and are probably the most abundantly expressed genes in vertebrates. Although numerous studies have identified both *cis*-regulatory elements and *trans*-acting factors involved in the regulation of type I collagen genes in various species, many details of the regulation of these important genes remain elusive [reviewed in Vuorio and de Crombrugghe, 1990; Slack et al., 1993; Brenner et al., 1994 Bornstein, 1996]. In addition to proximal *cis*-regulatory elements located in the promoters and first introns, other elements have recently been identified that are located both in the distal 5′ and 3′-flanking regions of the murine α1 and α2 type I collagen (*Col1a1* and *Col1a2*) genes [Bou-Gharios et al., 1996; Rippe et al., 1997; Salimi-Tari et al., 1997; Krempen et al., 1999]. In the work presented here, we have tested several different regulatory elements, most of them with well defined functions in the stage and tissue-specific regulation of *Col1a1* gene expression, for the presence of SIDD sites by computational analyses and for the ability to

bind to the nuclear matrix in an in vitro binding assay. We found that a common feature of all known or putative *Col1a1* regulatory elements analyzed are destabilized sites in their SIDD profiles. This includes regulatory elements with such diverse functions as chromatin organization, cell-specific transcriptional enhancement, and initiation and termination of transcription. We found furthermore that all distal and proximal 5′ elements analyzed so far show nuclear matrix binding potential, and that the calculated SIDD properties of a DNA fragment predict its matrix binding activity and vice versa. Our results highlight the importance of DNA duplex destabilization properties, which are structural attributes not directly tied to the presence of consensus sequences or motifs, in the physiological functioning of DNA. And they further document the usefulness of computational methods in the analysis of structure and function of eukaryotic genomes.

MATERIALS AND METHODS

Cell Lines

Murine F9 embryonal carcinoma cells, Balb/c 3T3 fibroblasts and Os50K8 osteoblasts (a kind gift from Jörg Schmidt) were all grown at 37°C in a humidified atmosphere of 5% CO_2 in DMEM with Glutamax-I (GibcoBRL) supplemented with 10% fetal calf serum, 100 U penicillin/ml, and 100 μg streptomycin/ml. F9 cells were cultivated on gelatinized culture dishes.

DNA Sequences

The DNA sequences of the murine *Col1a1* gene 5′ and 3′ ends and of the human *COL1A1* gene are accessible in the GenBank data base (accession numbers X54876, U38307, U38544, U50767, and X98705). The sequences of the *Col1a1* upstream HSs have not yet been determined with sufficient confidence to deposit them in the GenBank data base but are available upon request.

SIDD Calculations

Previously developed theoretical techniques to analyze stress-induced destabilization of the DNA double helix [Benham, 1992] were used to calculate the equilibrium statistical mechanical distribution of a population of identical superhelical molecules among its available states. Briefly, a molecule of N base pairs has 2^N possible patterns of denaturation. The energy associated to each pattern (i.e., state) depends on the precise base pairs that denature, and the amount of residual stress that remains after the change of helicity consequent on the denaturation. The fractional occupancy of any individual state at equilibrium decreases exponentially as the state energy increases. This enables one to calculate the equilibrium distribution, once the free energy of each state has been specified. From this distribution one can calculate the equilibrium ensemble average value of any parameter of interest. In particular, the probability $P(x)$ of denaturation of the base pair at each position x along the DNA sequence can be computed this way. However, a more informative measure of destabilization is given by the incremental free energy $G(x)$ needed to separate the base pair at each position x [Benham, 1993, 1996]. A value of $G(x)$ near or below zero indicates an essentially completely destabilized base pair, which is predicted to denature with high probability at equilibrium. Positive values of $G(x)$ occur for base pairs where incremental free energy is needed to assure separation. SIDD profiles, plots of $G(x)$ versus x, show regions of the sequence where superhelical stresses destabilize the duplex. The calculations whose results are reported here have been performed at linking differences that correspond to a superhelix density of -0.055, which corresponds to a moderate physiological value. (The superhelix density is a measure of the torsional strain on a region of DNA. If the molecule is planar, the superhelix density is the fractional change of its total helical twist away from its unstressed B-form value. Negative superhelix densities correspond to undertwisting of the DNA, which can drive denaturation.)

Nuclear Matrix Preparation and DNA-Binding Assays

Nuclei were subjected to the LIS-extraction procedure of Mirkovitch et al. [1984] following our previously published protocol [Kay and Bode, 1995]. Specifically, 5×10^7 cells of each cell line were washed once in ice-cold isolation buffer (3.75 mM Tris-HCl, pH 7.4, 0.05 mM Spermine, 0.125 mM Spermidine, 0.5 mM EDTA/KOH, pH 7.4, 1% Thiodiglycol, 0.1 mM PMSF), scraped off the plate in ice-cold isolation buffer containing 0.1% digitonin (Sigma), and homogenized using a B-type pestle in a Dounce-homogenizer. Nuclei were washed once in the same buffer and then stabilized for

20 min at 42°C in 500 μl nuclear buffer (5 mM Tris-HCl, pH 7.4, 0.05 mM Spermine, 0.125 mM Spermidine, 20 mM KCl, 1% Thiodiglycol, 0.1 mM PMSF, 1% Aprotinin). Non-scaffold proteins were then extracted using 25 mM lithium 3,5-diiodosalicylate (LIS) (Sigma) in an ice-cold buffer containing 100 mM lithium acetate, 20 mM Hepes-NaOH, pH 7.4, 1 mM EDTA, and 0.1% digitonin. The mixture was carefully homogenized in a Dounce-homogenizer applying four strokes of a loosely fitting A-type pestle. Resulting nuclear halos were centrifuged (1,000*g*, 5 min, 4°C) and the resulting soft pellets were carefully washed three times in restriction buffer (20 mM Tris-HCl, pH 7.4, 0.05 mM Spermine, 0.125 mM Spermidine, 20 mM KCl, 70 mM NaCl, 10 mM MgCl2). The halo DNA was then digested with 1,000 U Eco RI in 1 ml for 1 h at 37°C, 600 μg sonified *E. coli* genomic DNA was added as a non-specific competitor, and the mixture was then filled up with restriction buffer to a final volume of 1.5 ml. A 150 μl aliquots were then provided with a probe mixture (~10 ng of ^{35}S-end-labeled restriction fragments) and incubated over night at 37°C under gentle agitation. Samples were then separated into a pellet (P) and supernatant (S) fraction by centrifugation (14,000*g*, 10 min), DNA was purified by Proteinase K-digestion and analyzed by agarose gel electrophoresis, and subsequent blotting onto a Nylon membrane and autoradiography. Quantification of autoradiographs was achieved with a phosphoimaging system (Molecular Dynamics). Under these experimental conditions the relative matrix affinity of specific fragments can be expressed as the per cent band intensity in the bound (P) fraction as compared to total input DNA (band intensities in P and S lanes).

RESULTS

Distal 5′ *Col1a1* DNase-HSs Contain Regions of SIDD

Previous work has shown that several types of functional DNA elements contain SIDD regions [Benham, 1993, 1996; Benham et al., 1997; Sheridan et al., 1998; He et al., 2000; Leblanc et al., 2000]. Here we have applied a computational method to predict SIDD sites in superhelical DNA sequences [Benham, 1992] to assess whether regulatory elements in the murine *Col1a1* gene are associated with SIDD sites. We have analyzed several different types of regulatory elements; their localization within the *Col1a1* domain are shown in Figure 1 [Breindl et al., 1984, Salimi-Tari et al., 1997]. The first type of elements was a cluster of DNase I-hypersensitive sites (HSs 6-9) in the distal 5′-flanking sequence 15–20 kb upstream of the start site of transcription (region I in Fig. 1) which have properties of a locus control region (LCR) [Krempen et al., 1999]. For uniformity, the computational analyses of plasmids containing the *Col1a1* upstream HSs were performed under standard conditions. Each sequence was computationally inserted into the pBluescript plasmid, and analyzed at a standard superhelical density of −0.055. Figure 2 shows the SIDD profiles that were found. The vector portions of the plasmids show three characteristic destabilized sites coinciding with the promoter and terminator of the β-lactamase gene, and with the f1 origin of DNA replication [Benham, 1996]. When a plasmid containing HS 8 and HS 9 was analyzed, a sequence adjacent to HS 9 showed a very pronounced SIDD peak (Fig. 2A) coinciding with a 120 bp very AT-rich (>90%) sequence. The destabilization at this site was so strong that other sites, both in HS 8 and in the vector, could not compete and their destabilization was barely detectable (Fig. 2A). When HS 8 was analyzed without HS 9, two destabilization peaks occurred in the insert and all vector sites were destabilized (Fig. 2B). The main SIDD peak in HS 8 is associated with a shorter AT-rich sequence (60 bp, 81% AT) and an in vivo cleavage site for topoisomerase II [unpublished observation]. HS 7 showed several SIDD peaks (Fig. 2C). The strongest is associated with a 50 bp, 84% AT-rich sequence whereas the other peaks do not coincide with any apparent sequence motifs. HS 6 also showed several SIDD peaks (Fig. 2D). HS 6

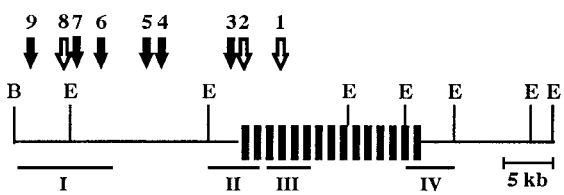
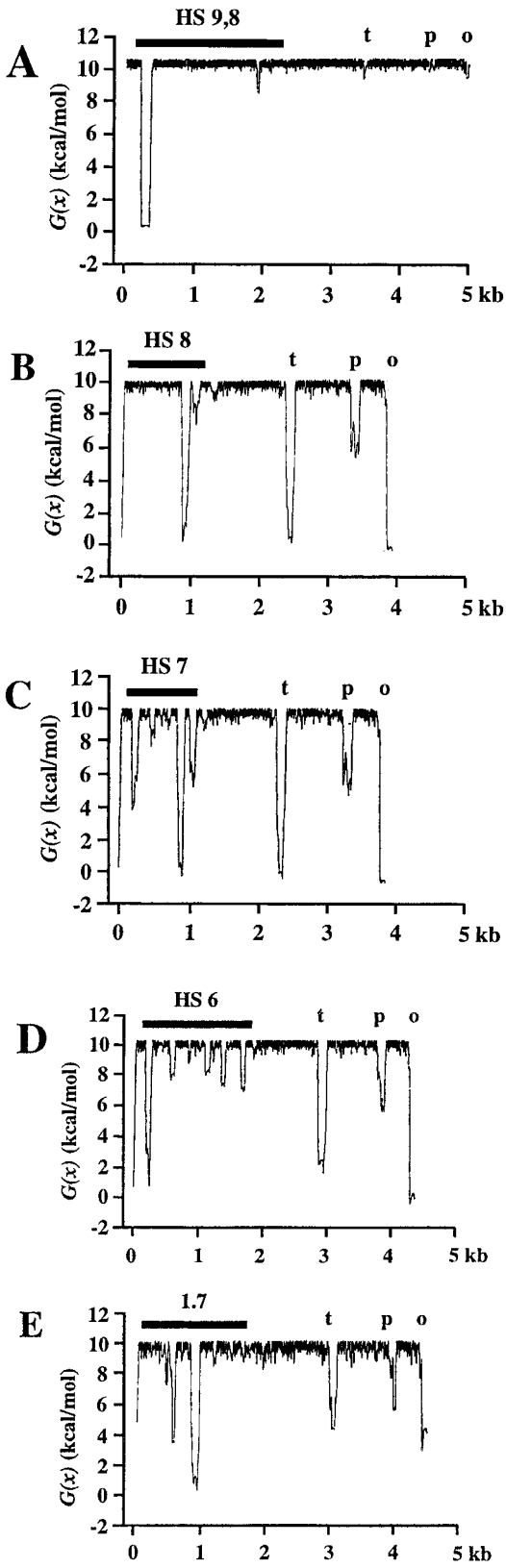


Fig. 1. Schematic representation of 55 kb of the mouse *Col1a1* gene domain showing the location of the coding region (striped box), of upstream and intragenic DNase I-HSs (open arrows: constitutive sites, closed arrows: transcription associated sites), of restriction sites (E, Eco RI; B, Bam HI), and of the various regions analyzed in this report (I–IV).



lacks AT-rich sequences but shows several unusual DNA motifs, predominantly microsatellite sequences such as several $(GA)_n$ dinucleotide repeats and consecutive $(GAAA)_n$, $(GGAA)_n$, and $(GGGA)_n$ tetranucleotide repeats. The plasmid designated 1.7 was initially found to have matrix binding activity in an in vitro binding assay (see the next paragraph) and subsequently sequenced and found to contain two SIDD peaks (Fig. 2E), one of which coincides with a 50 bp 94% AT-rich sequence.

Col1a1 Upstream HSs 6–9 Bind to Nuclear Matrices In Vitro

Because the *Col1a1* upstream HSs 6–9 showed strong destabilization in their SIDD profiles, we have used plasmids containing these sites in biochemical nuclear matrix binding assays to determine whether they indeed have affinities to lithium 3,5-diiodosalicate (LIS)-extracted nuclear matrices as described [Kay and Bode, 1995]. We used nuclear matrix preparations from collagen-producing fibroblasts and osteoblasts and non-collagen-producing F9 embryonal carcinoma cells. Northern analysis showed an about 50-fold higher steady state levels of *Col1a1* mRNA in the osteoblast cell line than in fibroblasts (Fig. 3). In preliminary experiments we had observed a size-correlated decrease in the affinity of HS-containing DNA fragments for the nuclear matrix, i.e., under the conditions used large fragments containing too much non-destabilized DNA adjacent to destabilized sites bound less well than smaller fragments. We therefore designed for these assays restriction digests that created two or more similarly sized restriction fragments from each plasmid which, when possible, contained centrally located unwinding elements (UEs) represented by SIDD peaks. The results of the matrix binding assays for the upstream HSs 6–9 are shown in Figure 4. A strongly binding 800 bp S/MAR fragment of the human β-interferon gene [fragment "IV" in

Fig. 2. Distal 5′ *Col1a1* regulatory elements contain regions of SIDD. DNA fragments containing the indicated regulatory elements (horizontal bars) were cloned into the circular pBluescript plasmid and their SIDD profiles calculated as described in Materials and Methods. p and t indicate the promoter and terminator of the ampicillin resistance gene and o the origin of DNA replication in the pBluescript vector, which are associated with characteristic SIDD peaks. 1.7 is a plasmid containing a region with S/MAR-binding activity and SIDD peak but no DNase-hypersensitive site. For details see the text.

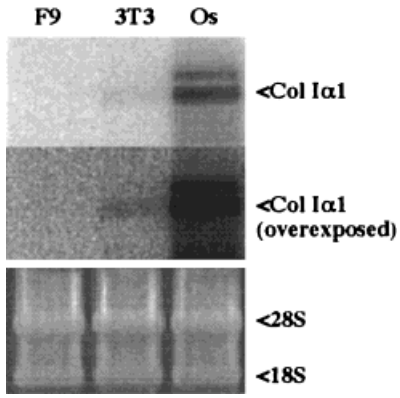


Fig. 3. Northern blot analysis of total RNA from embryonal carcinoma cells (F9), fibroblasts (3T3), and osteoblasts (Os). The ethidium bromide stained gel in the lower panel served as loading control.

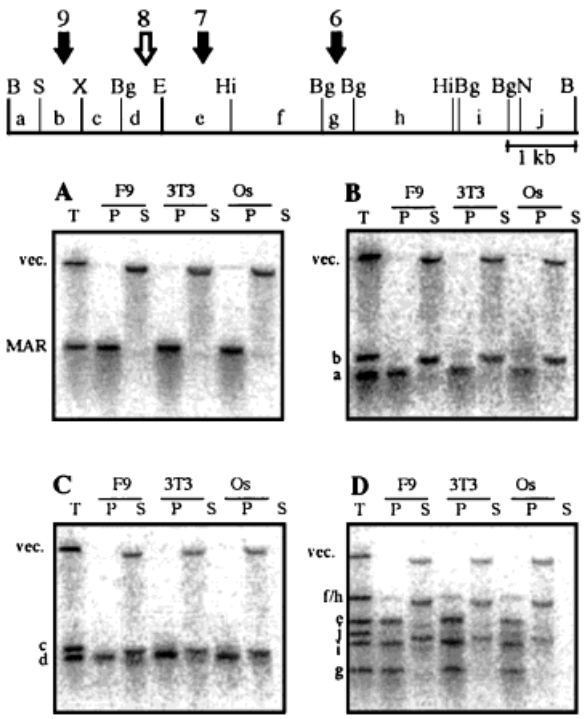


Fig. 4. Sequences associated with *Col1a1* upstream HSs 6–9 bind to nuclear matrices in vitro. Plasmids containing the human β-interferon S/MAR (**A**), HS 9 (**B**), HS 8 (**C**), and HSs 6 and 7 (**D**) were analyzed for their affinities to LIS-extracted nuclear matrices as described in Materials and Methods. Nuclear matrix preparations from collagen-producing fibroblasts (3T3) and osteoblasts (Os), and non-collagen-producing embryonal carcinoma cells (F9) were used. T, total input mixture; P, pellet fraction containing matrix-bound fragments; S, supernatant fraction containing unbound fragments. The location of used restriction fragments and HSs (comp. region I in Fig. 1) is indicated in the top panel. B, *Bam* HI; Bg, *Bgl* II; E, *Eco* RI; Hi, *Hind* III; S, *Stu* I; X, *Xba* I.

Mielke et al., 1990] served as positive control and bound to matrices from all three cell types with equally high affinity (Fig. 4A). The HS 9-containing plasmid was digested into vector sequences and two insert fragments of 460 and 590 bp, respectively. The smaller fragment containing the AT-rich sequence and a very strong SIDD peak (Fig. 2A) bound to the matrix preparations from all cell types, whereas the larger HS 9-containing fragment and the vector did not show any binding (Fig. 4B). Similarly, the HS 8-containing plasmid insert was digested into fragments of 674 and 536 bp, respectively. The smaller fragment containing HS 8 including the AT-rich sequence and SIDD peak (Fig. 2B) bound to all matrices, whereas the larger fragment did not (Fig. 4C). To test matrix binding potential of HSs 6 and 7, we used a plasmid digested into vector and six insert fragments with sizes between 467 and ~1500 bp (Fig. 4D). A 1 kb fragment containing HS 7 and the 467 bp fragment containing HS 6 bound to the matrices (fragments e and g in Fig. 4D). In addition, a 727 bp fragment located downstream of HS 6 which is not associated with a HS also bound to the matrices (fragment i in Fig. 4D). This fragment was subsequently sequenced and found to contain two strong SIDD peaks (plasmid 1.7, Fig. 2E). Our results show that all investigated *Col1a1* upstream HSs either have strong affinity to nuclear matrices or are located adjacent to DNA fragments that do. Furthermore, the calculated SIDD properties of a DNA fragment predict its biochemical matrix binding activity, and vice versa.

SIDD Profiles and Nuclear Matrix Binding Potential of the Region Surrounding the Start Site of *Col1a1* Gene Transcription

The second type of regulatory sequences analyzed were derived from the region surrounding the start site of transcription of the murine *Cola1* gene spanning 2.5 kb of proximal promoter sequence and exons and introns 1–5 (region II in Fig. 1). As shown in Figure 5A, the entire region has an extensively destabilized SIDD profile with several peaks in the promoter as well as the transcribed region. Because this region of the gene lacks extended AT-rich sequences, other sequence properties must be responsible for the destabilization potential.

The results of an analysis of the nuclear matrix binding potential of the area surrounding

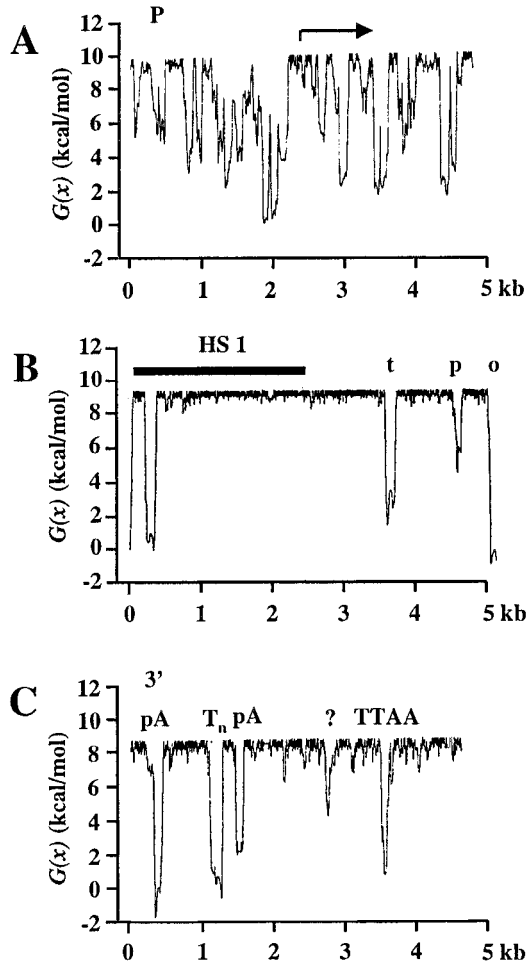


Fig. 5. The promoter and other potential regulatory elements in the murine *Col1a1* gene show extensive SIDD potential. SIDD profiles of DNA fragments containing the promoter and exons and introns 1–5 (**A**), HS 1 (**B**), and the 3′-flanking region (**C**) of the *Col1a1* gene. The 5′ end and 3′-flanking region were analyzed in the absence of plasmid sequences. The HS 1 insert is indicated by the horizontal bar. The start site of transcription in A is indicated by the arrow. The locations of the first five exons and introns are shown in Figure 8. For details see the text.

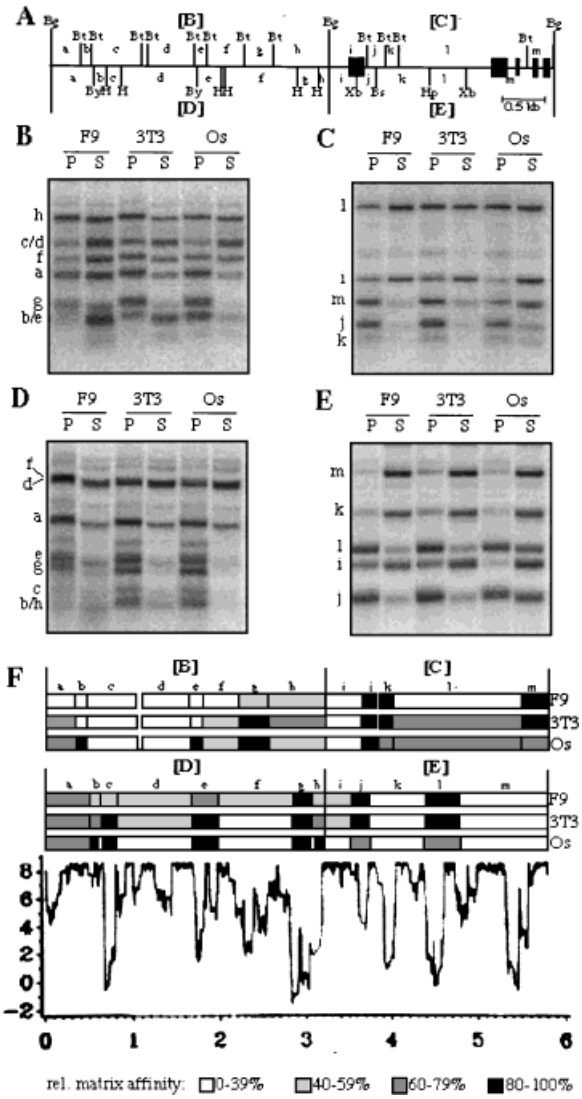


Fig. 6. The region surrounding the start site of *Col1a1* gene transcription shows complex interactions with the nuclear matrix. **A**: Gel-purified Bgl II-fragments representing *Col1a1* 5′-flanking sequences from −219 to −3435 or the proximal promoter and exons (black boxes) and introns 1–5 from −219 to +2362 were further cut with the indicated restriction enzymes (Bg, *Bgl* II; Bs, *Bst* EII; Bt, *Bst* SFI; By, *Bst* YI; H, *Hinf* I; H, *Hpa* I; Xb, *Xba* I), and analyzed for their affinities to LIS-extracted nuclear matrices as in Figure 3 (**B–D**). **F**: Shows a summary of relative matrix affinities for these fragments, as determined in multiple binding experiments, juxtaposed with the SIDD profile of the region. The regions analyzed in B–D are indicated in parenthesis in A and F, and subfragments are labeled in alphabetical order.

the *Col1a1* start site of transcription is shown in Figure 6. The relative matrix binding strengths of DNA fragments from this region of the gene were determined by averaging results from several independent binding experiments. We found a very good correlation between biochemical matrix affinity and computer-predicted destabilization potential (Fig. 6F). In most cases, when a SIDD peak was located in the center of a small fragment, a strong potential to bind to the nuclear matrix was seen (e.g., fragments c, e, g, j, and l in Figure 6D,E), while similar sized fragments lacking SIDD features did not bind to the matrix preparations (e.g., fragments i in Figure 6C,E). Importantly, matrix binding was detected only when relatively small DNA fragments with centrally located SIDD peaks were used in the binding assay but was lost when too much non-destabilized sequence was present. For example, fragments k in Figure 6C or l in Figure 6E

both were marked by a prominent SIDD peak and bound to the matrix, while corresponding larger fragments (l in Fig. 6C and k in Fig. 6E) bound less well. This shows that in the in vitro binding assays applied here, the matrix binding of regulatory elements is easily masked by neighboring sequences. This is in striking contrast to the observation that "classical" S/MARs require a minimum length for high-affinity binding [Mielke et al., 1990].

Other Potential Regulatory Elements in the *Col1a1* Gene Are Associated With SIDD Peaks

Previous studies have shown that several relatively short eukaryotic genes display tripartite patterns in their SIDD profile: destabilization in the promoter and terminator but no destabilization in the coding regions [Benham, 1993, 1996; Benham et al., 1997]. To determine whether the *Col1a1* gene shows a similar pattern, we analyzed several *Col1a1* sequences in addition to the promoter/first intron region shown in Figure 5A. Plasmid HS 1 (region III in Fig. 1) contains exons and introns 6–10, including a HS of unknown function in intron 5 of the *Col1a1* gene [HS 1; Breindl et al., 1984]. A strong SIDD peak was found at the 5′ end of the insert next to, but not coinciding with, HS 1 (Fig. 5B). Figure 5C shows an analysis of the 3′-untranslated and flanking region of the *Col1a1* gene (region IV in Fig. 1). Strong SIDD peaks were associated with the two polyadenylation signals [pA; Mooslehner and Harbers, 1988]. One additional strong SIDD peak was located between the two polyadenylation sites and coincides with a very T rich stretch of DNA (T_n, Fig. 5C). Another strong peak was located further downstream and co-localizes with a $(TTAA)_6$ tetranucleotide repeat next to a distal 3′ regulatory element that stimulates *Col1a1* gene expression in transfection experiments [Rippe et al., 1997]. Our results so far indicate that DNA destabilization appears to be a general property associated with promoter and terminator regions of eukaryotic genes and with many other known or potential regulatory elements. The matrix binding potential of the fragments showing SIDD peaks in Figure 5B,C has not yet been analyzed.

Human *COL1A1* Gene Shows Periodic SIDD Peaks Throughout the Transcribed Region

One potential function of SIDD sites and/or S/MARs is that they may act as topological sinks to facilitate the release of superhelical stress introduced into DNA during transcription or replication [Benham, 1996; Mielke et al., 1996; Bode et al., 1999]. If that were the case one would expect SIDD regions to be present at periodic intervals in long transcribed regions. To test this hypothesis we analyzed the SIDD profile of 17 kb of DNA sequence containing the promoter and exons and introns 1–43 of the human *COL1A1* gene. We used the human gene because much less DNA sequence information is available for the murine gene. As shown in Figure 7, strong SIDD peaks occurred in the promoter region, similar to the ones seen in the murine gene (Fig. 5A). In addition, SIDD peaks appeared throughout the transcribed region at intervals of approximately 5 kb. The peak at sequence position 5 kb (Fig. 7A) is located at the

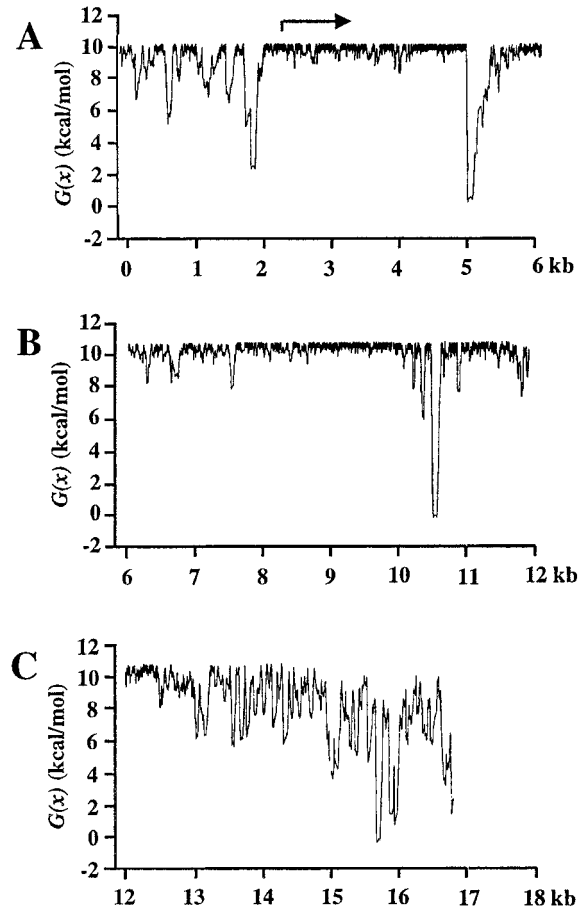


Fig. 7. The human *COL1A1* gene shows periodic SIDD peaks throughout the transcribed region. **Panel A** shows the SIDD profile of 2 kb of the promoter and the first 4 kb of the gene; the start site of transcription is indicated by the arrow. **Panels B** and **C** show the SIDD profiles of additional 11 kb of transcribed *COL1A1* sequences for which sequence information is available (exons 1–44). For details see the text.

same position as the SIDD peak next to HS 1 in intron 5 of the murine gene (Fig. 5B), and the SIDD peak at sequence position 10.5 kb (Fig. 7B) is located at a similar position as a previously mapped but not further studied HS in the murine gene [Breindl et al., 1984]. This suggests an evolutionary conservation of the position of regulatory elements in homologous genes, as has been suggested before [Salimi-Tari et al., 1997]. Another region with very high destabilization potential is located at sequence position 15–17 kb (Fig. 7C), a region with no known function. These results indicate that genes as large and abundantly expressed as the type I collagen genes may require periodic SIDDs to act as topological sinks to allow high rates of transcription.

5′ Ends of the Human and Mouse *Col1a1* Genes Show Similar SIDD Profiles but Only Limited Nucleotide Sequence Homology

The SIDD profiles of the promoters of the human and murine *Col1a1* genes shown in Figures 5A and 7A showed peaks at similar positions. In a more detailed analysis we compared the SIDD profiles of identical portions of the 5′ ends (2.5 kb of promoter sequence and exons and introns 1–5) of the murine and human genes. As shown in Figure 8, the murine sequence was more destabilized than the human sequence when analyzed at a superhelix density of −0.06 (compare Fig. 8A,C). However, both genes showed an unusually high destabilization potential, with the main SIDD peaks located at similar positions. The similarity became even more striking when the human profile at a superhelix density of −0.055 was compared to the murine profile at −0.06 (compare Fig. 8B,C). Intriguingly, this similarity between the SIDD profiles is not reflected in the nucleotide sequence. As indicated by the horizontal bars in Figure 8E there are only a few segments of nucleotide sequence similarity, most of which coincide with transcription factor binding sites [Rhodes et al., 1994; Bedalov et al., 1995; Slack et al., 1995; Rossert et al., 1996]. Our results strongly suggest that some DNA functions may be evolutionary conserved by preservation of higher order DNA structural properties rather than primary nucleotide sequence.

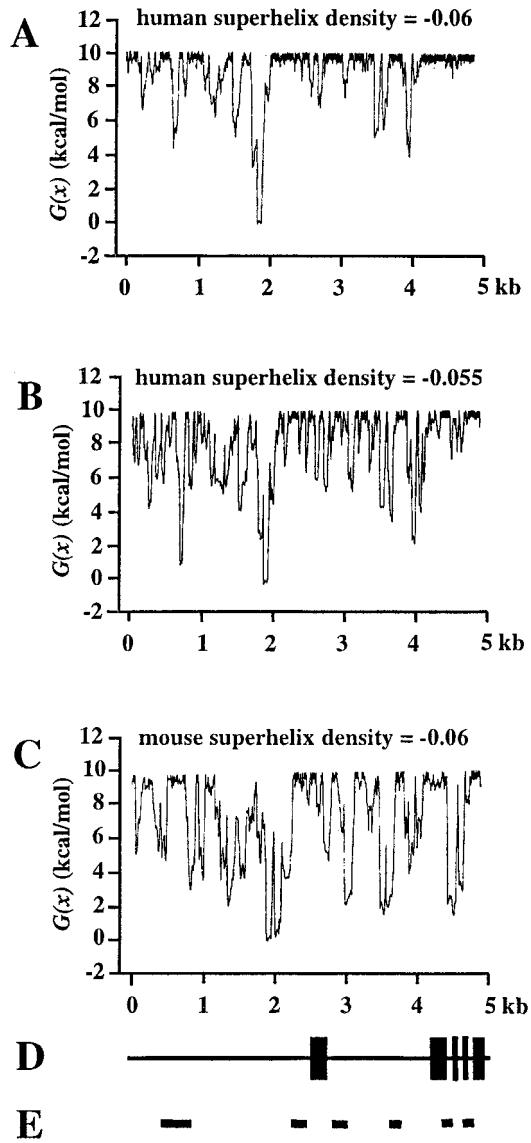


Fig. 8. The 5′ ends of the human and mouse *Col1a1* genes show similar SIDD profiles but only limited nucleotide sequence homology. The SIDD profiles of the regions surrounding the start 5′ ends of the human (**A**, **B**) and murine (**C**) *Col1a1* genes were analyzed at the indicated superhelix densities. **D**: The positions of exons 1–5 of the genes are indicated in the schematic drawing at the bottom by the vertical bars. **E**: Areas of sequence homology as determined by BLAST sequence comparison are shown as horizontal bars.

DISCUSSION

We have performed computational and biochemical analyses of previously identified and characterized regulatory elements within the domains of the murine *Col1a1* and the homologous human *COL1A1* genes. The sequences we analyzed contain regulatory DNA elements with a variety of different functions. First, we analyzed a cluster of distal 5′ HSs (region I in Fig. 1) which presumably functions in the chromatin loop organization of the *Col1a1* domain. This assumption is supported by the

following observations: these HSs enhance the position-independent expression of a reporter gene in transgenic mice [Krempen et al., 1999]. One of the sites (HS8) contains an in vivo topoisomerase II cleavage site [unpublished observation]. A sequence within HS 8 was found to specifically interact in a yeast one-hybrid assay with Smarce1-related protein [Wu, 2001]. Smarce1, a matrix-associated regulator of chromatin, is a unique addition to higher eukaryotic SWI/SNF chromatin remodeling complexes [Wattler et al., 1999]. A cluster of HSs located at a similar position in the murine *Col1a2* gene has strong transcriptional enhancer activity and confers position-independent expression in transgenic mice [Bou-Gharios et al., 1996]. The second type of regulatory sequences analyzed here are derived from the region surrounding the start site of transcription of the murine *Cola1* gene. This region spans 2.5 kb of proximal promoter sequence and exons and introns 1–5 (region II in Fig. 1). While the function of the first intron sequences remains controversial [Bornstein, 1996; Hormuzdi et al., 1998], the proximal promoter sequences have been shown in numerous experiments using transfections and transgenic animals to contain a modular arrangement of elements regulating *Col1a1* promoter activity in different collagen-producing cell types including skin, fascia and tendon fibroblasts, osteoblasts and odontoblasts [Bedalov et al., 1995; Rossert et al., 1995, 1996; Krempen et al., 1999]. Another region analyzed was an intragenic region (exons and intron 6–10; region III in Fig. 1) containing a HS of unknown function. Finally, we analyzed sequences derived from the 3′ end of the *Col1a1* gene (region IV in Fig. 1), including the 3′-untranslated region, the two polyadenylation sites [Mooslehner and Harbers, 1988], and a distal 3′ stimulatory element [Rippe et al., 1997]. All known or putative regulatory elements were found to exhibit regions of DNA duplex destabilization which could be localized in SIDD profiles (Figs. 2 and 5). The distal as well as proximal 5′ regulatory elements that were associated with destabilized sites also had experimental nuclear matrix binding activity in an in vitro binding assay (Fig. 4 and 6). Taken together with other results [Benham, 1993, 1996; Benham et al., 1997; Salimi-Tari et al., 1997; Sheridan et al., 1998; Krempen et al., 1999; He et al., 2000; Leblanc et al., 2000] the experiments described here show that destabilized sites in SIDD profiles are a common and may be universal feature of regulatory DNA elements with such diverse functions as chromatin structure organization, cell-specific transcriptional enhancement, and initiation and termination of transcription. Furthermore, one DNA fragment that had matrix-binding activity but was not associated with a HS (Fig. 4D) was subsequently found to have destabilized sites (Fig. 2E). This shows that the analyses used here detect sites that may or may not be associated with DNase I-hypersensitivity, and that the calculated SIDD properties of the sequences we analyzed reliably predicted their matrix binding activity, and vice versa. While further experimentation will be required to determine the matrix-binding properties of the destabilized sites in the transcribed region and the 3′-flanking region, it appears that computational analyses that predict SIDD properties are reliable and powerful tools for identifying regulatory DNA elements with a wide variety of functions. As more and more complete nucleotide sequences of prokaryotic and eukaryotic organisms become available, including the human genome (International Human Genome Sequencing Consortium, 2001], this structure-based approach should greatly facilitate the identification of regulatory DNA elements and their functional analysis.

A comparison of the SIDD profiles of the promoters of the human and murine *Col1a1* genes showed that they share unusually high destabilization potentials with the main SIDD peaks localized at similar positions. Perhaps surprisingly, the nucleotide sequences of these regions do not have extensive similarity. The regions of sequence similarity that are present occur mainly at positions that interact with transcription factors (Fig. 8). This suggests that two distinct requirements may influence the evolution of non-coding DNA sequences. The nucleotide sequence of some sites must be conserved to preserve their abilities to participate in specific DNA-protein interactions, while at other sites only the ability to adopt certain higher order structures must be retained. This latter requirement does not depend on a strict conservation of primary nucleotide sequence. Our study demonstrates that a comparison of the SIDD properties of homologous eukaryotic genes from different species provides a powerful technique to test this concept.

The results presented here show that multiple regulatory elements within a single chromatin domain have S/MAR functions. This is in support of a chromatin loop model in which proximal and distal regulatory elements can be juxtaposed by their affinity to the nuclear scaffold or matrix, allowing their cooperation in regulating promoter activity. It should be emphasized, though, that the matrix binding properties of the regulatory elements analyzed here differ, in certain aspects, from those we have previously assigned to "prototype" attachment regions such as the 800 bp standard in Figure 4A. The S/MARs in HS 9, HS 8, and HS 7 are "classical" as far as they coincide with AT-rich sequences, although relatively short ones. However, S/MAR sequences in HS 6, the promoter and the first intron are not AT-rich at all, underlining the complexity of S/MAR functions. Furthermore, while the S/MAR fragments from the human β-interferon gene and the *Col1a1* upstream HSs bound equally well to nuclear matrix preparations from collagen-producing and non-producing cells (Fig. 4), a somewhat different behavior was observed for some of the fragments derived from the promoter. For example, fragments b, e, and g in Figure 6B had much higher affinities to matrices from collagen-producing 3T3 fibroblasts and osteoblasts than to matrices from non-producing F9 cells. Fragment e is particularly interesting because it only binds to matrices from osteoblasts and contains a regulatory element necessary for *Col1a1* gene expression in osseous tissues that binds a protein selectively present in osteoblasts [Rossert et al., 1996]. The larger fragment e in Figure 5D, which binds to matrices from all cell types, contains additional sequences that are bound by ubiquitous transcription factors [Rossert et al., 1996]. This fragment contains a binding site for the nuclear matrix architectural transcription factor NMP4 [Alvarez et al., 1997]. Thus, our results reveal, at least in some instances, a correlation between in vitro S/MAR-binding activity and cell-specific transcriptional control elements. The exact nature and composition of the nuclear matrix is still controversial [Pederson, 1998; Hancock, 2000], and a future task will be to characterize more precisely the structures different S/MARs attach to in vivo and the intranuclear architecture that mediates their respective functions. However, the in vitro matrix binding activity revealed by the assays employed in this and other studies are a unique, reproducible, and mathematically treatable biochemical property of a specific subset of DNA sequences found in close association with diverse regulatory elements and therefore seem to faithfully reflect S/MAR function in vivo. The strong correlation between this biochemical property and destabilized sites in SIDD profiles found in this, as well as our previous studies [Benham et al., 1997], corroborates the usefulness of this type of computational analyses in reliably identifying DNA elements involved in the structural organization of the eukaryotic genome and the regulation of its expression.

ACKNOWLEDGMENTS

This work was supported in part by grants 97-100-32 and 97-143-09-9132 from the Danish Cancer Society to CM, MOC, and OW, grants RO1-GM47012 from the NIH and DBI-99-04549 from the NSF to CJB, grant Bo 419/6-2 from Deutsche Forschungsgemeinschaft to JBO and grant AR41909 from the NIH to MB.

REFERENCES

Alvarez M, Long H, Onyia J, Xu W, Bidwell J. 1997. Rat osteoblast and osteosarcoma nuclear matrix proteins bind with sequence specificity to the rat type I collagen promoter. Endocrinology 138:482–489.

Bedalov A, Salvatori R, Dodig M, Kronenberg MS, Kapural B, Bogdanovic Z, Kream BE, Woody CO, Clark SH, Mack K, Rowe DW, Lichtler AL. 1995. Regulation of *COL1A1* gene expression in type I collagen producing tissues: identification of a 49 base pair region which is required for transgene expression in bone of transgenic mice. J Bone Miner Res 10:1443–1452.

Benham CJ. 1992. The energetics of the strand separation transition in superhelical DNA. J Mol Biol 225: 835–847.

Benham CJ. 1993. Sites of predicted stress-induced DNA duplex destabilization occur preferentially at regulatory loci. Proc Natl Acad Sci USA 90:2999–3003.

Benham CJ. 1996. Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions. J Mol Biol 255:425–434.

Benham CJ, Kohwi-Shigematsu T, Bode J. 1997. Stress-induced duplex DNA destabilization in scaffold/matrix attachment regions. J Mol Biol 274:181–196.

Berezney R, Coffey DS. 1974. Identification of a nuclear protein matrix. Biochem Biophys Res Comm 60:1410–1417.

Berezney R, Wei X. 1998. The new paradigm: integrating genomic function and nuclear architecture. J Cell Biochem (Suppl) 30-31:238–242.

Bode J, Kohwi Y, Dickinson L, Joh T, Klehr D, Mielke C, Kohwi-Shigematsu T. 1992. Biological significance of unwinding capability of nuclear matrix-associated DNAs. Science 255:195–197.

Bode J, Bartsch J, Boulikas T, Iber M, Mielke C, Schübeler D, Seibler J, Benham C. 1998. Transcription-promoting genomic sites in mammalia: their elucidation and architectural principles. Gene Ther Mol Biol 1:551–580.

Bode J, Benham C, Knopp A, Mielke C. 1999. Transcriptional augmentation: modulation of gene expression by scaffold/matrix attached regions (S/MAR elements). Crit Rev Eukaryot Gene Expr 10:73–90.

Bornstein P. 1996. Regulation of expression of the α1(I) collagen gene: a critical appraisal of the role of the first intron. Matrix Biol 15:3–10.

Bou-Gharios G, Garrett LA, Rossert J, Niederreiter K, Eberspacher H, Smith C, Black C, deCrombrugghe B. 1996. A potent far-upstream enhancer in the mouse pro α1(I) collagen gene regulates expression of reporter genes in transgenic mice. J Cell Biol 134:1333–11344.

Boulikas T. 1995. Chromatin domains and predictions of MAR sequences. Int Rev Cytol 162A:279–388.

Breindl M, Harbers K, Jaenisch R. 1984. Retrovirus-induced lethal mutation in collagen I gene of mice is associated with an altered chromatin structure. Cell 38:9–16.

Brenner DA, Rippe RA, Rhodes K, Trotter JF, Breindl M. 1994. Fibrogenesis and type I collagen gene regulation. J Lab Clin Med 124:755–760.

Cook PR. 1999. The organization of replication and transcription. Science 284:1790–1795.

Dang Q, Auten J, Plavec I. 2000. Human beta interferon scaffold attachment region inhibits de novo methylation and confers long-term, copy number-dependent expression to a retroviral vector. J Virol 74:2671–2678.

Forrester WC, Fernandez LA, Grosschedl R. 1999. Nuclear matrix attachment regions antagonize methylation-dependent repression of long-range enhancer-promoter interactions. Genes Dev 13:3003–3014.

Hancock R. 2000. A new look at the nuclear matrix. Chromosoma 109:219–225.

He L, Liu J, Collins I, Sanford S, O'Connell B, Benham CJ, Levens D. 2000. Loss of FBP function arrests cellular proliferation and extinguishes c-myc expression, EMBO J 19:1034–1044.

Hormuzdi SG, Penttinen R, Jaenisch R, Bornstein P. 1998. A gene-targeting approach identifies a function for the first intron in expression of the α1(I) collagen gene. Mol Cell Biol 18:3368–3375.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Jenuwein T, Forrester WC, Fernàndes-Herrero LA, Laible G, Dull M, Grosschedl R. 1997. Extension of chromatin accessibility by nuclear matrix attachment regions. Nature 385:269–273.

Kay V, Bode J. 1995. Detection of scaffold-attached regions (SARs) by in vitro techniques: activities of these elements in vivo. Methods Mol Cell Biol 5:186–194.

Kirillov A, Kistler B, Mostoslavsky R, Cedar H, Wirth T, Bergman Y. 1996. A role for nuclear NF-κB in B-cell-specific demethylation of the *Igk* locus. Nat Genet 13:435–438.

Kohwi-Shigematsu T, Kohwi Y. 1990. Torsional stress stabilizes extensive base unpairing in DNA flanking the immunoglobulin heavy chain enhancer. Biochemistry 29:9551–9560.

Krempen K, Grotkopp D, Hall K, Bache A, Gillan A, Rippe RA, Brenner DA, Breindl M. 1999. Far upstream regulatory elements enhance position-independent and uterus-specific expression of the murine α1(I) collagen promoter in transgenic mice. Gene Expr 8:151–163.

Lamond AI, Earnshaw WC. 1998. Structure and function in the nucleus. Science 280:547–553.

Leblanc B, Benham CJ, Clark DJ. 2000. An initiation element in the yeast CUP1 promoter is recognized by RNA polymerase II in the absence of TATA box-binding protein if the DNA is negatively supercoiled. Proc Natl Acad Sci USA 97:10745–10750.

Mielke C, Kohwi Y, Kohwi-Shigematsu T, Bode J. 1990. Hierarchial binding of DNA fragments derived from scaffold-attached regions: correlation of properties in vitro and function in vivo. Biochemistry 29:7475–7485.

Mielke C, Maass K, Tümmler M, Bode J. 1996. Anatomy of highly expressing chromosomal sites targeted by retroviral vectors. Biochemistry 35:2239–2252.

Mirkovitch J, Mirault ME, Laemmli UK. 1984. Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold. Cell 39:223–232.

Mooslehner K, Harbers K. 1988. Two mRNAs of mouse proα1(I) collagen gene differ in the size of the 3′ untranslated region. Nucleic Acids Res 16:773.

Oancea AE, Berru M, Shulman MJ. 1997. Expression of the (recombinant) endogenous immunoglobulin heavy-chain locus requires the intronic matrix attachment region. Mol Cell Biol 17:2658–2668.

Paul A-L, Ferl R. 1993. Osmium Tetroxide Footprinting of a scaffold attachment region in the maize ADH1 promoter. Plant Mol Biol 22:1145–1151.

Pederson T. 1998. Thinking about a nuclear matrix. J Mol Biol 277:147–159.

Phi-van L, van Kries JP, Ostertag W, Strätling WH. 1990. The chicken lysozyme 5′ matrix attachment region increases transcription from a heterologous promoter in heterologous cells and dampens position effects on the expression of transfected genes. Mol Cell Biol 10:2302–2307.

Rhodes K, Rippe RA, Umezawa A, Nehls M, Brenner DA, Breindl M. 1994. DNA methylation represses the murine α1(I) collagen promoter by an indirect mechanism. Mol Cell Biol 14:5950–5960.

Rippe RA, Umezawa A, Kimball JP, Breindl M, Brenner DA. 1997. Binding of upstream stimulatory factor to an E-box in the 3′-flanking sequence stimulates α1(I) collagen gene transcription. J Biol Chem 272:1753–1760.

Rossert J, Eberspaecher H, de Crombrugghe B. 1995. Separate *cis*-acting DNA elements of the mouse pro-α1(I) collagen promoter direct expression of reporter genes to different type I collagen-producing cells in transgenic mice. J Cell Biol 129:1421–1432.

Rossert JA, Chen SS, Eberspaecher H, Smith CD, deCrombrugghe B. 1996. Identification of a minimal sequence of the mouse α1(I) collagen promoter that confers high-level osteoblast expression in transgenic mice and that binds a protein selectively present in osteoblasts. Proc Natl Acad Sci USA 93:1027–1031.

Salimi-Tari P, Cheung M, Safar CA, Tracy JT, Tran I, Harbers K, Breindl M. 1997. Molecular cloning and chromatin structure analysis of the murine α1(I) collagen gene domain. Gene 198:61–72.

Sheridan SD, Benham CJ, Hatfield GW. 1998. Activation of gene expression by a novel DNA structural transmission mechanism that requires supercoiling-induced DNA duplex destabilization in an upstream activating sequence. J Biol Chem 273:21298–21308.

Slack JD, Liska DAJ, Bornstein P. 1993. Regulation of expression of the type I collagen genes. Am J Med Genet 45:140–151.

Slack JL, Parker MI, Bornstein P. 1995. Transcriptional repression of the α1(I) collagen gene by *ras* is mediated in part by an intronic AP1 site. J Cell Biochem 58:380–392.

Stein GS, van Wijnen AJ, Stein JL, Lian JB. 1999. Interrelationships of transcriptional machinery with nuclear architecture. Crit Rev Eukaryot Gene Expr 9:183–190.

Stief A, Winter DM, Strätling WH, Sippel AE. 1989. A nuclear DNA attachment element mediates elevated and position-independent gene activity. Nature 341:343–345.

Vuorio E, de Crombrugghe B. 1990. The family of collagen genes. Annu Rev Biochem 59:837–872.

Wattler F, Wattler S, Kelly M, Skinner HB, Nehls M. 1999. Cloning, chromosomal location, and expression analysis of murine Smarce1-related, a new member of the high-mobility 365 group gene family. Genomics 60:172–178.

Wu D. 2001. Analysis of the murine α1(I) collagen gene hypersensitive site 8 and identification of DNA binding proteins. Mater ' s Thesis, San Diego State University.