Introduction

During recent years, high throughput genome sequencing, its annotation, gene array analyses and proteomics have significantly enhanced the pace of functional genomics (1-4). Defining the molecular circuitry that coordinates the temporal and spatial patterns of gene expression is a key requirement for the successful exploitation of this information. In eukaryotes, an array of cis-regulatory DNA elements (CRDEs) located in the proximal and distal regions of the mRNA encoding (Pol II responsive) genes are targeted by their cognate transcription factors and regulate the temporal and spatial pattern of gene expression (5, 6). Painstaking research by numerous laboratories has led to the identification of hundreds of such proteins, their target DNA sequences and the cognate promoters (7-9). Nonetheless, while ~35,000 predicted human genes are regulated by ~1850 transcription factors (10), only 1871 human promoters have yet been reported and only a fraction of those have been extensively characterized (http://cmgm.stanford.edu/help/manual/databases/epd.html). Conventionally, identification of gene regulatory proteins involves experimental delineation of the regulatory sequences in the target promoter followed by molecular cloning of the cognate factors. However, alternative approaches for the identification of novel gene expression pathways have lately been emerging (11). A number of laboratories have developed algorithms for (i) inter species sequence comparison (12-14) (ii) identification of clusters of CRDEs in gene promoters (15, 16); (iii) identification of common sequences in the promoters of co-regulated genes (17-19); and (iv) finding repetitive sequences in the entire genome (20-22). Nonetheless, CRDEs targeted by transcription factors are small, of variable lengths (four base pairs and above) and often show limited degeneracy at various nucleotide positions. Therefore often it is difficult to identify them in the background of the entire genome, especially in metazoan organisms (23-26). It is thus expected that further development of
methodologies for efficient identification of novel CRDEs and their cognate protein factors will contribute to the comprehensive understanding of cellular transcriptomes.

Recent years have witnessed the emergence of the developing heart as a paradigm of cell differentiation and organogenesis. (27-29) Cardiac development is a dynamic and highly orchestrated process that begins with a small population of cells in the lateral plate mesoderm and ends with the formation of a highly structured organ with distinct gene expression programs even within different sub-domains of each chamber (30). It is thus anticipated that cardiac development and function involve a highly intricate and coordinated gene expression apparatus (31).

In this study we have developed a highly efficient strategy for the construction of a library of CRDEs from developing chick heart. A number of parameters were used for establishing the authenticity of the candidate binding sites in the library. Finally, in silico determination of the exact protein binding sequences led to the detection of a large number of novel regulatory elements. Potential use of this library for identifying novel gene expression pathways in heart and in other tissues is discussed.
Experimental Procedures

Materials: Oligonucleotides used in this study were synthesized from Genosys, USA. Polymerase chain reaction was done using KlenTaq DNA polymerase (). All chemicals were purchased from Sigma Aldrich unless mentioned otherwise.

Construction of a Pre-library: Synthetic oligonucleotides (Table 1) used for the construction of a pre-library of random decanucleotide sequences was as described by Nallur et al. (32). Single-stranded oligonucleotide was amplified by PCR using primers encompassing the terminal sequences (M1 and M2, Table 1). The PCR product was then digested with HincII and SpeI and cloned into pBluescript plasmid vector (+/-) digested with the same enzymes. The frequency of recombination was checked by PCR amplification of randomly picked up bacterial colonies using the same primer. More than eighty percent of transformants contained the recombinant plasmids and were pooled in batches containing 1-1.2 X 10^4 colony-forming units. Thus we had a pre-library of 1.5 X 10^5 recombinants segregated in fifteen independent pools.

Selection of binding sites: The cognate binding sites were then selected from the pre-library by four consecutive steps of gel-mobility shift followed by PCR. Plasmid DNA isolated from each pool of recombinants was amplified by T3 and T7 primers (Table 1) and 1 ng of amplified DNA was used for binding with 40 μg nuclear protein prepared from 13 day old chick embryonic heart. The reaction was then resolved on 8% acrylamide gel and the entire lane starting from 1 cm above the unbound oligonucleotide (identified by ethidium bromide staining) to 0.5 cm below the well was excised, crushed in to small pieces and eluted overnight in a buffer containing 500mM Ammonium acetate, 10mM MgCl2, 1mM EDTA, 0.1% SDS. The eluted DNA was then precipitated using 100 ng of poly dI-dC as a carrier and amplified by a second set
of primer pairs (P3 and P4, Table 1), which are essentially same as T3 and T7 primers except that of having two more nucleotides at the 3' end added from the multiple cloning sites of the pBluescript vector. The selection process was then repeated followed by the amplification with a third set primer pairs (P1 and P2, Table 1). Use of sequentially internal primer pairs during selection-amplification cycles helped in obtaining cleaner PCR products after each rounds of selection. After three rounds of selection, the eluted material was finally amplified by the M1 and M2 primer pairs and cloned in to the pBluescript vector at the Hinc II and SpeI sites as before (for the construction of the pre-library). The recombinants thus obtained were then pooled and saved as the “library of cis-regulatory DNA elements”.

**Nuclear Extract Preparation:** Tissues (heart, brain, liver and skeletal muscle) from 13day old chick embryos were homogenized in 1 ml of Buffer A (20mM HEPES pH7.9, 20% glycerol, 10mM NaCl, 1.5mM MgCl₂, 0.2mM EDTA, 1mM DTT, 0.1% Triton X100 and protease inhibitors viz., PMSF, Leupeptin, aprotinin and Pepstatin) in a glass hand-homogenizer. Homogenates were centrifuged at 2000rpm at 4°C for 15 minutes and the nuclear pellet was resuspended in 500 μl of Buffer B that was essentially same as buffer A except containing 500 mM NaCl. The nuclei were lysed by incubation for one hour on ice with intermittent tapping. Homogenates were then centrifuged at 10,000 rpm at 4°C for 15 minutes and the supernatants were aliquoted and snap frozen at −80°C till used.

**Gel mobility shift assay:** DNA protein binding reaction was carried out in 40μl binding buffer (20mM HEPES pH 7.9, 5% glycerol, 60mM NaCl, 1.5mM MgCl₂, 1.0 mM EDTA, 1.0mM DTT) containing 20 μg nuclear extract, 1μg poly dI dC and ³²P- labeled probe (40,000 cpm) on ice for 40 minutes. Competition reaction was carried out using 100-fold molar excess of unlabeled
DNA. Protein-DNA complex were fractionated on an 8% polyacrylamide gel in 0.5X TBE for 3 hours at 200V, 4°C.

**Foot printing analysis:** Plasmid DNA isolated from each clone was amplified with T7 and Fp (internal to pBluescript M13 primer, Table 1) primers and then digested by the restriction enzyme BssHII. Digested fragment was end-labeled by 32-P-alpha-dCTP and Klenow fragment. One ng of probe was then used for a binding reaction containing 40-50 μg of nuclear extracts in a reaction volume of 40 μl. The binding reaction was essentially as described under “gel mobility shift assay”. The reaction was incubated at 4°C for 60 minutes and the DNase I digestion was initiated by the adjusting the concentration of MgCl2 and CaCl2 to 10mM and 5mM respectively. DNase I solution was then added to a final concentration of 0.01 to 0.02 unit/ml and incubated at 37°C for the required period of time. The reaction was then stopped by adding stop buffer (1% SDS, 200mM NaCl, 20mM EDTA final concentration). Samples were then extracted by phenol-chlorophorm, precipitated and then resuspended in loading dye containing 95% formamide, 0.1mM EDTA and tracking dyes. The DNA samples were then heated at 95°C for 4 min and electrophoresed in 6% acrylamide gel containing 5M urea.

**DNA sequencing:** The nucleotide sequences of selected binding sites were done manually using sequencing kit from Epicenter Biotechnologies, USA.
Results

Construction of library of CRDEs from embryonic chick heart: Methodologies for the isolation of DNA sequences targeted by recombinant transcription factors or total nuclear proteins have previously been described (32, 33). Such selections are done from a pool of synthetic double stranded oligonucleotides harboring all possible sequence combinations. However, such oligonucleotide pools would have heterogeneity in the order of $1 \times 10^6$ for decanucleotides (and more for larger sequences). Selection of protein binding sites from such a large number of sequences using total nuclear proteins would be inefficient especially for less abundant or highly specific transcription factors. Such limitations would be further accentuated for larger target sequences, which will be of even lesser frequency. Therefore, in order to increase the selection efficiency, we segregated a pool of decanucleotide sequences for batch wise selection. We first cloned random decanucleotide sequences in pBluescript plasmid vector and made a pre-library (see Materials and Methods for details). Batches of recombinants containing 1-1.2 $\times 10^4$ independent sequences were then screened for candidate CRDEs by four consecutive rounds of gel mobility shift and PCR using nuclear extracts from 12day old embryonic chick heart. The selected sites were again cloned into pBluescript vector. To assess the efficacy of the selection process thus adopted, groups of ten recombinants (containing the selected candidate CRDEs) were randomly picked up, radiolabeled in a pool by PCR using vector specific primers (M1 and M2, Table 1) and tested by gel mobility shift assay. Binding profiles of two such independent groups (G1 and G2) were then compared. We argued that if the selection protocols were effective, each group of recombinants containing the CRDEs would generate multiple complexes, the profile of which would differ from each other, at least partially. We also tested their specificity by competition with hundred fold molar excess of another
unrelated candidate CRDE randomly picked up from the library of binding sequences. Such analysis showed that a number of complexes present in each group were not diminished by non-specific competitors and thus were sequence specific (Fig 1, identified by arrowheads). We also observed that in the group G1, there was a fast migrating complex (identified by #) formed only with the liver extract. However, close examination of the profile generated with cardiac extract showed the presence of the same complex in a barely detectable level (identified by *). Therefore the methods we had adopted were sensitive enough in selecting sequences for which the cognate binding activities was very low in cardiac cells (while they were incidentally high in liver cells). Thus, the library of CRDEs we constructed could be considered to be comprehensive.

Compilation of the library of CRDEs: We subsequently determined the nucleotide sequences of 164 candidate target sites of which 123 were unique sequences and 41 were in duplicates arising due to repeated PCR adopted during their selection. Those unique sequences were then compared with the transcription factor binding sites available in the TRANSFAC database (http://transfac.gbf.de/cgi-bin/matSearch/matsearch.pl). Forty-seven sequences turned out to be previously characterized in the contexts of various promoters expressing in both cardiac and non-cardiac tissues. The remaining 86 sequences were novel. A summary of this analysis is given in Table 2. cis-regulatory sequences are known to have different levels of degeneracy at different sequence positions (34). We also observed that at least some of the known sequences we had isolated were different from their canonical counterparts reported in the literature. We thus verified the authenticity of a number of such sequences viz., NF-1, AP-1 and AP-2. Radiolabeled probes were prepared from commercially available (Stratagene) binding sites for NF-1, AP-1 and AP-2 and the specific complexes were challenged by hundred-fold molar excess of non-radioactive competitors generated from the corresponding binding sequences (as identified by by
As shown in Fig. 2, such analyses demonstrated that out of four NF-1 like sequences (consensus: TTGGC(A/G)_6 CC), only H2.49 (GCATATGGCAGga) substantially diminished the complex formed by the canonical NF-1 binding site, while both H5.9 (TGGCAGTAC) and H2.78 (TGTTGGCTCC) were moderate competitors and the fourth sequence i.e., H4.12 (CTGGCTTCAG) did not compete at all. Since the consensus NF-1 sequence is thirteen nucleotide long and the random pool used for the selection were only ten-nucleotides in length, all the candidate NF-1 sequences we had isolated had either the 5'-end (TTGGC) or other parts of the NF-1 sequence, thereby representing only partial NF-1 binding activity. We also performed similar analysis with the AP-1 sequence (GTCATCACGC) and confirmed its identity (data not shown). Such analyses therefore established the authenticity of the known sequences and prepared the basis for the characterization of the novel sequences.

**Characterization of novel CRDEs:** For characterizing the novel CRDEs, we randomly took about twenty candidates (cloned in plasmid vector) and tested their protein binding activities or otherwise using 12-day-old embryonic chick heart nuclear extract. Probes were generated from each sequence by PCR (using M1 and M2 primer pair) and the specificity of binding was tested by competition with hundred fold molar excess of unlabeled and unrelated (i.e., with no sequence similarity with the test sequence) sequences picked up from the same group. We observed that at least seventeen of them had specific binding activities, albeit with differences in the intensity of complexes, while remaining three either had non-specific or no binding activities. We also simultaneously tested the tissue distributions of a number of those sequences using nuclear extracts from brain, skeletal muscle, heart and liver of 12-day-old chick embryo. Although none of them had absolute tissue specificity (cardiac or otherwise), some had binding activities enriched in one tissue as compared to the other. One such activity that was enriched in
cardiac and skeletal tissues (as compared to brain and liver) is shown in Fig 3A. For a number of sequences, the binding profile also varied from one tissue to another indicating the presence of tissue specific isoforms of the cognate proteins. Incidentally we came across one binding site that had shown significant difference in the binding profiles between liver and brain extracts. However, this particular sequence showed very weak complex formation (at a position corresponding to the liver specific complex) with the cardiac extract (Fig 3B). It thus appears that this particular binding site was isolated by the cardiac extract although the cognate binding activity was very low, thereby exemplifying the efficiency of selection protocol.

We also had tested the binding activities of a number of target sites by Southwestern analysis and in agreement with the gel mobility shift assays, most of the target sites thus tested had distinct protein binding activities (data not shown).

Since the length of cis-regulatory DNA sequences can be four and above, it was expected that the exact protein binding sequences in each of the selected sites would be ten base pairs (the length of the randomized pool used for selection) or less. We thereafter performed DNAse I footprinting analysis for identifying the exact protein binding sites for a number of novel sequences. As shown in Figure 4, DNAse-I protected regions ranged from penta-nucleotides (Panel C), hepta-nucleotides (Panel A), nona-nucleotides (Panel C), entire decanucleotide core sequence along with part of the vector sequences (P and E) or a bipartite tetra-nucleotide sequence with a two-nucleotide spacer (Panel D). Table 3 summarizes the repertoire of target sequences we thereby identified by foot-printing analysis. In agreement with the gel mobility shift analysis as described above, foot-printing analysis also showed that only one out of ten sequences tested did not have any binding activity.
We further argued that in the pre-selection pool of sequences, smaller binding sequences (four or five nucleotides) would occur more frequently than the larger sequences and as a consequence, smaller binding sites would also occur more frequently in a group of selected sequences. We tested this possibility by comparing the binding sites identified by foot printing with the remaining of the sequences. As expected, smaller binding sites (as identified by foot printing) had number of related sequences in the entire group (summarized in Table 4). We also experimentally proved their functional relatedness by gel mobility shift-competition assay. As shown in Fig 5, clone H2.43 (TGAACCTTTG) showed a bipartite protected sequence TGAA and CTTG in foot-printing analysis (Fig 4, panel D). Upon sequence comparison, we found that at least one other isolate harbored similar sequence pairs (#31: GGTCAAGTCTgg) and three others contained at least one of them (#62: ATCTTGGGCT, #76: GTCTTGTCGG, #78: TgTTGGCTCC). As expected, in gel mobility shift-competition assay using H2.43 (TGAACCTTTG) as probe and hundred fold molar excess of each of those sequences as competitors, sequence #31 effectively competed, #76 and #62 moderately competed and #78 poorly (or not at all) competed the H2.43 specific complex (Figure 5). Similar analysis was also done with two other sequences identified and the data is summarized Table 4.

Following experimental identification of a number of exact binding sites and their related sequences, we independently did a computational analysis to segregate the entire set of sequences (total 123) in to groups of related sequences. Accordingly, we used the program Pratt and derived fuzzy sequence patterns, which included at least five sequences. Pratt identified twenty such patterns (Table 5) and as expected, some of them were either known transcription factor binding sites or some of the sequences we had already identified by foot printing analysis. In order to avoid any mutual bias in experimental (by foot-printing) and in silico identification of
those sequences, the foot-printing analysis was done prior to the computational analysis. These data therefore established that the pool of novel target sites could be successfully exploited for both in vivo and in silico identification of novel protein (transcription factor) binding motifs. We also attempted to search the available database of about 2500 eukaryotic promoters for the occurrences of those binding sites or the patterns as identified above and as expected, a very large number of hits were found for each of them (data not shown) which is likely to be due to their smaller lengths as well as to the high level of degeneracy encountered within them. It is realized that establishing the statistical significance to their occurrences in the genome(s) must take into account contexts dependent information such as their position with respect to other such binding sites as well as to the coding sequences. Such studies are currently underway.
Flow of information from gene to mRNA to protein is a well-choreographed process and is specified by the cellular and biological contexts. Understanding the molecular mechanisms by which cells decipher its genetic information requires identification of the gene promoters, their cis-regulatory codes, and the cognate transcription factors. Although comparative genomics have facilitated such understanding, comprehensive knowledge of cis-regulatory codes remains a major challenge due to their diminutive nature in the background of an enormously large and complex genome. A complete catalogue of eukaryotic transcriptional regulators and their cognate sites (with positional weight matrices) is a goal yet to be achieved. In the present study, we have isolated for the first time a large repertoire of cis-regulatory DNA elements functional in developing chick heart. During the recent years developing heart has drawn considerable attention for studying cell differentiation and organ development. Furthermore, genetic cascades that regulate cardiac development are highly conserved from *Drosophila* to human and therefore are of wider significance. We have adopted an experimental technique previously used by others (32) for the isolation of novel DNA sequences targeted by transcription factors. Nonetheless, by substantial modifications of the methodology, i.e., construction of a pre-library followed by batch-wise extraction of target sites we have significantly enhanced the efficiency and sensitivity of the selection process and thereby isolated a large number of sequences with putative binding activities. Notably, out of 123 sequences analyzed yet, 47 represented binding sites for known transcription factors. Simple numerical extrapolation therefore indicates presence of thousands of novel sites (but not necessarily thousands of independent activities) in this library. Since our objective is to create a comprehensive catalogue of DNA sequences involved in heart development, we tested a number of criteria to assess the quality of the selected sequences. As
desired, majority of the tested sequences showed specific binding activities. Since the selection was done from a pool of decanucleotide sequences, the protein binding sequences, however, restricted to 5-10 bases (as identified by foot printing). Also, the present study was done in a limited scale to characterize a small number (total 123) of candidate sequences and in future, high throughput sequencing will further enrich the database.

The promoter contexts for each of the novel sequences we had isolated are not known yet. Nonetheless, we argue that these sequences might belong to two groups. In the first category there are sequences for which the prototype cognate promoters are yet to be identified. While in the second category, there are sequences for which the cognate promoters are known but those binding sites have not yet been detected due to experimental limitations. In a recent study a set of genes over-expressed in heart were identified from the EST database. Upon analysis of their upstream regions, clusters of conserved sequence elements that includes a number of known transcription factor binding sites such as AP-1, AP-2, and NFAT as well as a substantial number of novel sequences were identified (35). Significantly, the regulatory regions of at least two of those genes i.e., myosin light chain-2 and myosin heavy chain have previously been characterized by a number of laboratories but those conserved sequences were not identified as functional regulatory elements (36, 37).

The library of cis-regulatory sequences we have constructed is presumably too large for the experimental identification of each of the target sites. Neither it is feasible to use any parameter for identifying the relative importance of each of the sequences. To address these issues we identified a number of frequently occurring patterns and matched them with the experimentally identified binding sites (Table 4). The robustness of this approach and additional
parameters such as occurrence with other known sequences in clusters are currently being tested for the determination of the promoter context for each of these sequences.

In conclusion, we envisage a number of possible applications of this database of CRDEs. (i) some of the novel sequences can be directly used for cloning the cognate transcription factors by the expression screening of appropriate libraries (we have tried this approach and have isolated two novel cDNAs), (ii) the binding sites can also be used as a probe for comparing the cognate activities and thereby identifying differentially active transcription factors/DNA binding proteins, (iii) the cognate proteins can be affinity purified and characterized and finally, (iii) the sequence information thus created can be used for identifying novel gene expression pathways by genome analysis.
Acknowledgements

Sindhu KV and Vibha Rani were the recipients of fellowships from the University Grants Commission and Council of Scientific and Industrial Research, India. The research work was supported by the grants awarded to S K Goswami by the Council of Scientific and Industrial research, Govt. of India (37/1015/99-EMR II and). We also acknowledge generous help received from Prof. D P Sarkar, Department of Biochemistry, Delhi University South Campus, New Delhi
Reference lists:


Figure Legends:

Fig 1. **Selected target sites have sequence specific protein binding activities:** Radiolabeled probes were generated from two independent pools (G1 and G2) of ten candidate CRDEs and gel mobility shift assay was done using nuclear extract from 12-day-old chick embryonic heart and liver. Specificity of the complexes was tested by competition (+) with 100 fold molar excess of a randomly picked up unrelated sequence. Specific complexes are identified by arrows. One liver enriched complex is identified as # and its cardiac counter part is identified as *.

Fig 2. **Confirmation of NF-1 binding sites as identified by TRANSFAC.** Radiolabeled probes were prepared from commercially available (Stratagene) consensus NF-1 binding site. Complexes formed by 12-day-old chick cardiac nuclear extracts were competed with hundred fold molar excess of each of the candidate sequences. Left panel: NF-1 probe, S: competition with unlabeled-self DNA. Other competitors are identified on top of each lane. Ns: competition with an unrelated (non-homologous) candidate-binding site.

Fig 3. **Binding profile of two novel binding sites:** Nuclear extracts from skeletal muscle (Sk), brain (B), liver (L) and heart (C) of 12-day-old chick embryo was used for gel-mobility shift assay. Left panel: binding profile of H4.19 sequence (GGATGGGGT). S: competition with hundred-fold molar excess of the unlabeled self-DNA. Ns: competition with hundred molar fold excess of an unrelated sequence. Thin arrow identifies the specific complex. Thick arrow marks non-specific complexes generated by the adjoining vector sequences. Right panel: binding profile of H1.3 sequence (ACGGCGGGCT). Thin arrows identify the specific complexes in brain and liver tissues. Non-specific complexes are identified by *. S: competition with hundred-fold molar excess of the unlabeled self-DNA. N: competition with hundred molar fold excess of an unrelated sequence.
Fig 4. **DNase I foot printing analysis of a number of novel sequences.** In each panel, left lane is the control digestion (without any nuclear protein), middle lane is digestion with cardiac extract and the right lane is digestion with liver extract. Protected region is shown by a bracket and the cognate sequence is marked by blue letters. Protein binding sequences from the vector is given in small letters and the unprotected sequences are shown in black letters.

Fig 5: **Comparison of the DNA binding properties of homologous novel sequences:** Radiolabeled probe was prepared from the sequence TGAACCTTTG (Table 5). Competition analysis was done with four independent sequences having complete or partial homology to TGAA and CTTG. Each competing sequence is identified on the top of each lane: #62: ATCTTTGGGCT, # 76: GTCTTGTTCG, #78: TGTTGGCTCC, # 31: GGTCAAGTCTgg. Megy et al independently identified homologous sequences (CTTG and TGTT) in the promoter regions of a number of genes highly expressed in heart (35).
Fig 3

H.4.19

H.1.3
Table 1. Oligonucleotides and primers used in the construction and characterization of the library.

<table>
<thead>
<tr>
<th>Primers</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Oligonucleotide Used for constructing the pre-library</td>
<td><code>5’ CGAGGTCGACGGTATCGNNNNNNN 3’</code></td>
</tr>
<tr>
<td>T3 Primer</td>
<td><code>5’ AATTAACCCTCACTAAAA 3’</code></td>
</tr>
<tr>
<td>T7 Primer</td>
<td><code>5’ TAATACGACTCACTATA 3’</code></td>
</tr>
<tr>
<td>P3 Primer</td>
<td><code>5’ TAACCCTCACTAAAGGG 3’</code></td>
</tr>
<tr>
<td>P4 Primer</td>
<td><code>5’ TACGACTCACTATAGGG 3’</code></td>
</tr>
<tr>
<td>P1 Primer</td>
<td><code>5’ TCGAGGTCGACGGTAT 3’</code></td>
</tr>
<tr>
<td>P2 Primer</td>
<td><code>5’ CGCTCTAGAATAGTGG 3’</code></td>
</tr>
<tr>
<td>M1 Primer</td>
<td><code>5’ CGAGGTCGACGGTATCG 3’</code></td>
</tr>
<tr>
<td>M2 Primer</td>
<td><code>5’ GCTCTAGAATAGTGGATC 3’</code></td>
</tr>
</tbody>
</table>
Table 2: Primary characterization of selected sequences:

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of sequences</td>
<td>164</td>
</tr>
<tr>
<td>Number of sequences occurring in duplicate</td>
<td>041</td>
</tr>
<tr>
<td>Number of unique sequences</td>
<td>123</td>
</tr>
<tr>
<td>Number of known sequences</td>
<td>047*</td>
</tr>
<tr>
<td>Number of unknown sequences</td>
<td>086</td>
</tr>
</tbody>
</table>

Table 3: Compilation of target sites as determined by DNase-I foot printing analysis:

<table>
<thead>
<tr>
<th>Sequence No.</th>
<th>Sequence (Binding site is given in bold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H2.2</td>
<td>ATGTGTTGGG</td>
</tr>
<tr>
<td>H2.3</td>
<td>AGTTCTCAT</td>
</tr>
<tr>
<td>H2.5</td>
<td>tcgCGGAACTCGgga</td>
</tr>
<tr>
<td>H2.9</td>
<td>ATTAAGGGGT</td>
</tr>
<tr>
<td>H2.41</td>
<td>TAAGAGTCG</td>
</tr>
<tr>
<td>H2.43</td>
<td>TGAACTCTTTG</td>
</tr>
<tr>
<td>H2.50</td>
<td>ATAGTGCTTT</td>
</tr>
<tr>
<td>H2.54</td>
<td>CACTAGGTCG</td>
</tr>
<tr>
<td>H2.55</td>
<td>ATACTTTTT</td>
</tr>
<tr>
<td>H2.62</td>
<td>ATCTTGGGCT</td>
</tr>
<tr>
<td>H2.74</td>
<td>GTGTCAGG</td>
</tr>
<tr>
<td>H2.80</td>
<td>CCTCTGGTGG</td>
</tr>
<tr>
<td>H4.30</td>
<td>tcgTACTTGGTCC</td>
</tr>
<tr>
<td>H1.7</td>
<td>GTGGTTTCC</td>
</tr>
</tbody>
</table>
Table 4: Similar target sites as independently identified by sequence comparison: and Foot printing.

<table>
<thead>
<tr>
<th>Binding sequence</th>
<th>Similar Sequence</th>
<th>Competition Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATAGTGTT (H2.50)</td>
<td>ATGTGTTGGG</td>
<td>+ + +</td>
</tr>
<tr>
<td></td>
<td>GTACAGTGTT</td>
<td>+ + +</td>
</tr>
<tr>
<td></td>
<td>TAAGaGTcG</td>
<td>+ +</td>
</tr>
<tr>
<td></td>
<td>TTAGaGTgTC</td>
<td>-</td>
</tr>
<tr>
<td>TGAACTCTTG (H2.43)</td>
<td>ATCTTGGGCT</td>
<td>+ +</td>
</tr>
<tr>
<td></td>
<td>GTCTTGTGGG</td>
<td>+ +</td>
</tr>
<tr>
<td></td>
<td>TgTTGGCTCC</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>GGTCAaGTCTgg</td>
<td>+ + +</td>
</tr>
<tr>
<td>ATACTTTTT (H2.55)</td>
<td>gTACTTTTTT</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>gTACTTTTTg</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>TACTTTgTcc</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 5. Compilation of similar sequences by *in silico* analysis.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Pattern (score)</th>
<th>Status*</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGTTGGctcc</td>
<td>T- [CG]-T-T-G-G (24.1729)</td>
<td>FP (CTTG)</td>
</tr>
<tr>
<td>aTCTTGgct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>atgTGTGGGg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aTCTTGgct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>atgTGTGGGg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aTAGTGTgtt</td>
<td>T-A-G-T-[CG]-T (24.1729)</td>
<td>ND</td>
</tr>
<tr>
<td>aTAGTGTgtt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TACTCtact</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tTACTCTgc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tTAGTGTgtc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>atAGTGTGtt</td>
<td>A-G-T-[CG]-T-G (24.1729)</td>
<td>ND</td>
</tr>
<tr>
<td>atAGTGTGtt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ctaAGTCTGt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ttAGTCTGc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ttTAGTGTgtc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATGGGCgtt</td>
<td>A-T-G-G-[CG] (24.1729)</td>
<td>ND</td>
</tr>
<tr>
<td>ATGGGCgtt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cATGGGGccg</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ggATGGGGt  
ggATGGGGtg  
GAACCCatta  
cgGAACTCg  
cgGAACTCg  
GAACCCaagt  
GAACCCaagt  
tGAACCTCttg  

aTACTTTtt  
aTACTTTtt  
gTACTTTtgc  
gTACTTTtt  
TACTCTcact  

atagTGTGTT  
aTGTGTTggg  
atagTGTGTT  
aTGTGTTggg  
ttagTGTGTC  

atAGTGTGtt  
ND  

T-A-C-T-[CT]-T 24.0745  
ND  

T-G-T-G-T-[CT] (24.0745)  
FP (GTGTT)  

A-G-T-G-[GT] (24.0506)  
FP (GTGTT)  

gtagAGTGTT  
gtagAGTGTT
ttAGTGTGtc

atACTTTTtt A-C-T-T-[GT]-T (24.0506)  ND
atACTTTTtt
gtACTTTTtg
gtACTTTTtt
tACTTGTtcc

aTACTTTTtt T-A-C-T-T-[GT] (24.0506)  ND
aTACTTTTtt
aTACTTTTtt
gTACTTTTtg
gTACTTTTtt
TACTTGttcc

GTGTGTatc G- [AT]-G-T-G-T (24.0483)  ND
ataGTGTGTt
ataGTGTGTt
gtaGAGTGTt
tgaGAGTGTc
ttaGAGTGTc

ttTTCTTTTc T-[AT]-C-T-T-T (24.0483)  NFAT-1
tTTCTTTTca

aTACTTTTtt
aTACTTTTtt
gTACTTTTtg
gTACTTTtt

attAAGGGGt A-[AT]-G-G-G-G (24.0483) FP (AAGGGGT)
attAAGGGGt
CATGGGGccg
ggATGGGGt
ggATGGGGGt

atTAAGGGgt
TAAGAGctag
TAAGAGcttg
TAAGAGtcg

ATAGTGttgtt A-[CGT]-A-G-T-G 23.5388 ND
ATAGTGttgtt
ctACAGTGtgc
gtACAGTGtt
gtAGAGTGtt
tgAGAGTGtc

tGTGTTGCTcc G-T-T-G-G-[CGT] (23.5388) NF-1
atgtGTTGGG
atctGTTGTT
atgtGTTGGG
atctGTTGTT
atgtGTTGGG  G-[ACT]-T-G-G-G (23.5351)  ND
atgtGTTGGG
 gGATGGGgt
 gGATGGGgtg
tGCTGGGca

tgtTAGGGT  T-A-G-[AGT]-G-T (23.4591)  ND
aTAGTGTgtt
 aTAGTGTgtt
 gTAGAGTgtt
tTAGTGTgtc

atctTGGGCT  T-G-G-G-C-[AGT] (23.4591)  FP (TGGGCT)
aTGGGCtgt
tctTGGGCT
atctTGGGCT
aTGGGCtgt
tgcTGGGCA

tcccCTATTT  C-T-[AGT]-T-T-T (23.4591)  ND
ataCTTTTT
 ataCTTTTT
 CTGTTTcggc
* FP: Binding site was independently confirmed by foot print analysis. ND: Foot printing analysis has not been done.