Identification and mapping of DNA binding proteins target sequences in long genomic regions by two-dimensional EMSA

Igor P. Chernov, Sergey B. Akopov, Lev G. Nikolaev, and Eugene D. Sverdlov

BioTechniques 41:90-96 (July 2006)
doi 10.2144/000112197

Specific binding of nuclear proteins, in particular transcription factors, to target DNA sequences is a major mechanism of genome functioning and gene expression regulation in eukaryotes. Therefore, identification and mapping specific protein target sites (PTS) is necessary for understanding genomic regulation. Here we used a novel two-dimensional electrophoretic mobility shift assay (2D-EMSA) procedure for identification and mapping of 52 PTS within a 563-kb human genome region located between the FXYD5 and TZFP genes. The PTS occurred with approximately equal frequency within unique and repetitive genomic regions. PTS belonging to unique sequences tended to group together within gene introns and close to their 5′ and 3′ ends, whereas PTS located within repeats were evenly distributed between transcribed and intragenic regions.

INTRODUCTION

The publication of the human genome sequence (1,2) and sequences of other metazoan genomes greatly facilitated positioning and analysis of various genomic functional elements and first of all coding sequences (3,4). At the same time, a complete functional annotation of sequenced eukaryotic genomes is supposed to include positions of all noncoding regulatory elements. Unfortunately, experimental data on genomic positions of a multitude of regulatory sequences, like enhancers, promoters, transcription terminators, and replication origins, are very limited, especially at the whole genome level. In general, most genomic regulatory elements (e.g., enhancers) are gene-, tissue-, or cell-specific, and prediction of these elements by computational methods is difficult and not always reliable. Therefore, the development of high-throughput experimental approaches to identification and mapping of genomic functional elements is highly desirable.

Specific binding of nuclear proteins to target DNA sequences is a major mechanism of genome functioning and regulation in eukaryotes (5), that makes identification and mapping of specific protein target sites (PTS) necessary for understanding genomic regulation. To date, several approaches to unbiased mapping of PTS have been proposed and used. The most widely used is a chromatin immunoprecipitation-on-a-chip (ChIP-on-chip) technique that allowed to map target sites for the NF-κB (6) and CREB (7) transcription factors across human chromosome 22 and the Sp1, c-Myc, and p53 factors across human chromosomes 21 and 22 (8). Another experimental approach named DamID (9) was recently used for mapping GAGA (10), Myc, Max, and Mad/Mnt target sites across the whole Drosophila genome (11). It should be noted that the both techniques are applicable only to mapping binding sites of well-characterized transcription factors.

Computational identification of PTS is, in turn, strongly limited by the lack of experimental data necessary for development of algorithms and validation of the results (12,13). A general approach should include identification of a whole set of specific PTS and grouping them according to their functional role and interactions with other regulatory units. The result should be a protein binding map of extended genomic regions or even whole genomes, ideally a dynamic map depending on cell origin, environmental conditions, and other factors.

Recently, we proposed experimental approaches for identification and mapping of nuclear matrix binding regions (S/MARs) (14) within a 1-Mb human chromosome 19 locus between the FXYD5 and COX7A1 markers (15). The locus contains 45 Reference Sequence (16) genes expressed with different tissue specificities and therefore could be a good model for the study of the mammalian genome regulatory network. Here we present an approach for high-throughput identification and mapping of a multitude of PTS within a given genomic region. Using this approach, we mapped 52 sequences capable of specifically binding Jurkat cell nuclear proteins within a 563-kb long FXYD5-TZFP human chromosome 19 region, a fragment of the FXYD5-COX7A1 locus mentioned above.
MATERIALS AND METHODS

Basic Protocols

Growth and transformation of *Escherichia coli* cells, preparation of plasmid DNA, agarose gel electrophoresis, electrophoretic mobility shift assay (EMSA), and other standard manipulations were performed as described (17).

Cells and Nuclear Extract

Jurkat cells (acute T cell leukemia, TIB-152; ATCC, Manassas, VA, USA) were grown in suspension at 37°C, supplemented with 10% fetal calf serum, up to approximately 2 × 10^6/mL. Nuclear extract was isolated as previously described (18) with modifications (19).

Preparation of a Short-Fragment Library

DNA of cosmids R30072, R28588, F19410, R30879, F24108, F16632, R26667, F12426, R28461, F14121, R31396, F25451, R31076, R28052, and P1-derived artificial chromosome (PAC) PC28130 (kindly provided by A. Olsen, Lawrence Livermore National Laboratory, Livermore, CA) was isolated and digested with restriction endonucleases *Sau*3A and *Csp*6I, ligated with the library primer 5'-ACT TGAAGCTGAGTATCCATGAACAGTATCCATGAACA-3', and PCR-amplified with the same primer as described previously (14,20).

Two-Dimensional EMSA

For two-dimensional EMSA (2D-EMSA), a pool of short DNA fragments of a 563-kb *FXYD5-TZFP* region of human chromosome 19 was radioactively labeled by PCR with the library primer and purified as described previously (21). The 2D-EMSA was performed generally as described (21), but instead of purified DNA binding protein, 2.5 μg Jurkat cell nuclear extract protein was added to the initial EMSA reaction.

The gel was then autoradiographed overnight, the area containing PTS (see Figure 1A) was excised, cut into small pieces, and DNA was eluted and precipitated as described (21). To increase the specificity of selection, the above 2D-EMSA procedure was repeated twice. Finally, 2 μL polyacrylamide gel eluate were PCR-amplified (20 cycles of 94°C for 20 s, 60°C for 60 s, and 72°C for 90 s) and cloned in a pGEM®-T vector (Promega, Madison, WI, USA) according to the manufacturer's recommendations. White colonies (184) were selected and arrayed on a 96-well microplate. The selected clones were checked by PCR, and those lacking inserts or producing more than one PCR product (double inserts) were discarded.

One-Dimensional EMSA

For EMSA, inserts of individual clones were labeled by PCR as described above and purified by electrophoresis in a 5% polyacrylamide gel. EMSA was done essentially as described above with a 50,000 counts per minute (cpm) probe, 1 μg nuclear extract protein, and 1 μg poly(dI-dC)*poly(dI-dC). For competition experiments, an excess of an unlabeled probe was added.

Sequencing, Computer Analysis, and Mapping

Sequencing was done with an ABI Prism® BigDye™ Terminator v. 3.1 kit using an ABI Prism 3100-Avant™ automated sequencer (all from Applied Biosystems, Foster City, CA, USA). The sequences obtained were mapped by comparison with those deposited in GenBank® using the BLAST (22) server at the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov/BLAST). The data were further analyzed using the University of California, Santa Cruz (UCSC) Human Genome Browser (genome.ucsc.edu) (23).

RESULTS

EMSA is one of the most widely used methods to explore interactions between DNA and nuclear proteins. The EMSA approach was initially proposed for quantifying interactions between DNA and purified proteins (24,25) and was later adapted for crude cellular or nuclear protein extracts (26). Recently, we proposed a two-dimensional variant of EMSA (2D-EMSA) that allowed us to identify and map binding sites of the CTCF transcription factor within a 1-Mb human genome region (21). Here we present a modification of 2D-EMSA that allows one to obtain and clone DNA fragments capable of binding nuclear extract proteins of given cells, with the pattern of the fragments probably being
characteristic of these cells. Using this approach, we identified and mapped several tens of potential nuclear protein target sequences across an approximately 600-kb human genome region.

Figure 1A presents the principle of 2D-EMSA. A pool of short DNA fragments covering the genome region of interest is prepared and labeled with $^{32}$P. This pool is mixed with nuclear extract prepared from a specific cell line and containing DNA binding proteins characteristic of this cell type. Then, the DNA-protein complexes are separated in a nondenaturing first-dimension gel, as in the conventional EMSA. The resulting gel strip that contains DNA-protein complexes and free DNA fragments is localized by autoradiography, excised from the gel, and incubated in sodium dodecyl sulfate (SDS)-containing buffer to disrupt DNA-protein complexes. The strip is then loaded onto a second-dimension gel (the same as for the first dimension but with 0.1% SDS), and the gel was run and autoradiographed. Figure 1B shows the results of the 2D-EMSA. The diagonal spots in the resulting gel represent the fragments with approximately the same electrophoretic mobility in both dimensions (i.e., the fragments that did not bind to proteins). The fragments bound to proteins in the first dimension are retarded, but their mobility is restored in the second dimension due to dissociation of DNA-protein complexes. Therefore, the spots corresponding to the fragments initially bound by nuclear proteins are located below the diagonal, and the area outlined by a dashed line (Figure 1A) is supposed to contain a majority of such fragments.

Construction and Properties of a Nuclear PTS Library

A short-fragment library of a human chromosome 19 region located between the $FXYD5$ and $TZFP$ markers that contained about 2000 fragments, with a mean length of approximately 400 bp, was prepared. To this end, DNA isolated from 14 overlapping cosmids and one PAC was digested with restriction endonucleases $SalI$ and $Csp6I$, and primers were ligated to the ends of the resulting fragments. The $FXYD5$-$TZFP$ region under study is part of the $COX7A1$ locus used in our previous work (14). It contains about 20 known genes with different tissue specificities.

After two rounds of 2D-EMSA, the DNA binding events were amplified, and the resulting library of potential target sites of nuclear DNA binding proteins was cloned and arrayed. Inserts of 120 clones were labeled with $^{32}$P, and their ability to specifically bind nuclear proteins was tested by one-dimensional EMSA, as described in the Materials and Methods section and exemplified in Figure 2A. Inserts (98 of 120 or approximately 80%) were found to bind the proteins, suggesting a high specificity of the 2D-EMSA selection procedure. In addition, six randomly selected clones were checked for specificity of binding by competition with an unlabeled probe (Figure 2B), and their specificity was confirmed.

Number of PTS in the Genome

The protein binding fragments were sequenced, and their sequences compared with the Human Genome Database. In total, 78 target sequences belonging to the $FXYD5$-$TZFP$ region were identified. The remaining 20 sequences were mapped to other human genome regions or belonged to the E. coli genome. A comparison of the 78 sequences with each other revealed 52 unique sequences. These sequences were mapped across the $FXYD5$-$TZFP$ region of human chromosome 19 (Figure 3), and their precise positions in the corresponding genome context are presented in Supplementary Table S1 (available online at www.BioTechniques.com).

To estimate the number of independent clones in the library, we assumed that the frequency of occurrence of cloned sequences in library samples (see Supplementary Table S1) fit the Poisson distribution. Accordingly, the Poisson curve was adjusted by the least-squares method to fit the data obtained and used to calculate the library size as (the number of selected clones)/$q$, where $q$ is a parameter of the Poisson distribution. The estimated number was found to be about 120. Therefore, 52 PTS found here may represent about a half of potential PTS of this region identifiable by 2D-EMSA in Jurkat cells.

The possible number of protein binding sites in a genomic region of a given length can be evaluated. It can be assumed that expression of each gene is controlled by several transcription factors, so this number should depend on the number of both active and silenced genes in the region. Recent
estimations for the complete human genome (3000 Mb) were 12,000 Sp1, 25,000 c-Myc, 1600 p53 (8), and about 12,000 NF-kB target sites (6). We have identified at least 52 nuclear PTS in the FXYD5-TZFP region of 563 kb in length (over 270,000 if extrapolated to the whole genome) in a single cell type, which should represent a considerable fraction of all PTS. Taking into account the calculations above, the actual number could be twice as large.

PTS Map

All identified PTS (Supplementary Table S1) were subdivided into two almost equal groups—unique and repeat-containing, the members of which contain over 25% repeated sequences. The repeats were most often represented by Alu or long-interspersed nuclear elements (LINEs) retroelements, as well as long terminal repeats (LTRs) of human endogenous retroviruses. A complete list and the positions of the repeats can be found in Supplementary Table S2. The finding that half of the protein binding sites were embedded in repeats suggests an important role of repeated elements, especially retroelements, in gene regulation.

All repeat-containing and unique PTS were placed on a physical map of the FXYD5-TZFP region (Figure 3). An interactive map can be obtained with the use of the Human Genome Browser (genome.ucsc.edu/cgi-bin/hgGateway?db = hg12), assembly of May 2004, and an annotation file in Supplementary Table S3. The distribution of PTS within the region is summarized in Table 1. About 50% of the FXYD5-TZFP region is occupied by genes and expressed sequence tag (ESTs), and the rest is occupied by intergenic sequences. The data presented here support the earlier observation (6,8) that PTS are not restricted to the 5′ regions of genes, but just as often are localized 3′ to or within genes.

All PTS were subdivided into intronic, 5′ and 3′ sites (located within 1 kb from the 5′ or 3′ ends of genes, respectively), and intragenic PTS depending on their positions relative to genes and ESTs. As seen from Table 1, unique PTS tend to be located inside or near genes—only 2 of 25 unique sites were found >1 kb distant from the corresponding gene. In contrast, over 50% (14 of 27) of the repeat-containing PTS were located within intragenic regions.

Although some target sequences overlap with exons, none was found within exons. However, as borders of binding sites within target sequences are not known, the existence of exonic binding sites cannot be completely ruled out.

DISCUSSION

Recently, we utilized the 2D-EMSA technique for identification of binding sites for a well-characterized transcription factor. This technique allowed us to find 10 new binding sites of the CTCF protein within the human chromosome 19 FXYD5-COX7A1 locus (21). In the present work, we applied the 2D-EMSA to a much more complex system that includes a multitude of DNA binding proteins present in the nuclear extract of Jurkat cells.

Table 1. Transcription Factor Binding Sites Statistics

<table>
<thead>
<tr>
<th>Position</th>
<th>Repeated Sites</th>
<th>Unique Sites</th>
<th>Repeated Plus Unique Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>3′ to gene/EST</td>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>5′ to gene/EST</td>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Intronic</td>
<td>8</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>Intergenic</td>
<td>14</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>25</td>
<td>52</td>
</tr>
</tbody>
</table>

EST, expressed sequence tag.
The modified 2D-EMSA approach allowed us to select, clone, and map DNA fragments (PTS) capable of specific interaction with nuclear proteins. The method is simple and allows identification of hundreds of protein binding fragments in one experiment. Moreover, the results of 2D-EMSA will strongly depend on the protein composition of the nuclear extracts used and thus might characterize the pool of nuclear proteins of given cells or tissues (i.e., tissue specificity of transcription factors binding). The 2D-EMSA technique is unbiased with respect to DNA binding factors and can therefore select PTS of yet unidentified proteins. At the same time, this approach has some limitations. First, it is based on in vitro DNA-protein interactions and, therefore, leaves aside possible in vivo modifications of this process due to chromatin structure variations, methylation of DNA, and the modification of proteins. In addition, it preferentially detects DNA sites with the highest affinity to target proteins, as low affinity sites will produce relatively small amounts of complexes. For these reasons, some existing in vivo contacts can be missed. However, due to its dynamic nature, chromatin structure can most probably create only transient barriers to the binding of transcription factors (discussed in Reference 27). Our data on the CTCF binding are in line with this hypothesis—all identified in vitro CTCF binding sites did bind CTCF in vivo as verified by ChiP. Interestingly, three of the CTCF binding sites coincided or overlapped with PTS 34, 37, and 43 found in this work. It means that at least some of the proteins that can bind specific DNA sequences in vitro will be able to bind them also within chromatin.

Only a rather small fraction (approximately 15%) of the identified PTS was found in promoter regions (5’ regions) of genes. Most (70%) of the PTS were located within intronic sequences or in intergenic regions. One may speculate that these PTS belong to regulatory elements not structurally linked to genes, like enhancers, insulators, or locus control regions.

In conclusion, we propose an approach for identification and cloning of a large number of DNA binding protein target sequences within long genomic regions in a single experiment. This approach can be a useful instrument of functional genomics, especially in combination with microarray technologies.

ACKNOWLEDGMENTS

The authors thank V.K. Potapov and N.V. Skaptsova for oligonucleotide synthesis and B.O. Glotov for critical reading of the manuscript. The authors would like to express special gratitude to A. Olsen (Lawrence Livermore National Laboratory) for providing cosmid DNAs. This work was supported by the Russian Foundation for Basic Research (project 05-04-48814), President of the Russian Federation grant NSh 2006.2003.4, and the Russian Academy of Sciences Physical and Chemical Biology Program. DNA sequencing was done at the Genome Center (www.genome-centre.narod.ru) under support of the Russian Foundation for Basic Research (grant no. 00-04-55000).

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES


Received 14 February 2006; accepted 12 April 2006.

Address correspondence to Lev G. Nikolaev, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, 16/10 Miklukho-Maklaya, 117997, Moscow, Russia. e-mail: lev@humgen.siobc.ras.ru

To purchase reprints of this article, contact: Reprints@BioTechniques.com