SITECON: A QUALITY TOOL FOR PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES NOW HANDLES THOSE FOR SF-1.

EXPERIMENTAL VERIFICATION AND ANALYSIS OF REGULATORY REGIONS OF ORTHOLOGOUS GENES

Ignatieva E.V.*, Oshchepkov D.Yu., Klimova N.V., Vasiliev G.V., Merkulova T.I.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: eignat@bionet.nsc.ru

Key words: transcription factor binding site prediction, endocrine system, SF-1, conformational and physicochemical DNA properties

SUMMARY

Motivation: Methods that accurately predict transcription factor binding sites (TFBS) have always been important tools in studying the regulatory regions of eukaryotic genes. It is therefore important that more new high-performance methods for TFBS prediction be developed and their accuracy, assessed using experimental data.

Results: Using a new technique, SITECON, potential binding sites for the transcription factor SF-1 have been predicted in the 5′-flanking regions of a range of vertebrate steroidogenesis genes, for which it was unknown whether or not SF-1 participate in the regulation of their expression. A high predictive capacity of SITECON was proved by experimental verification: the predicted sites were all shown to be able to bind to SF-1 in vitro. Most of them are found at positions, which are similar to those at which known SF-1 binding sites with an experimentally proven functionality are located in the genes of other species. The new genes that we have thus detected are in fact potential targets for SF-1 and are perceived to be promising candidates for further experimental verification.


INTRODUCTION

Computer-based methods that predict binding sites for transcription factors are some of the most promising approaches, which it is believed can unravel the regulatory code of DNA. Statistical analysis of sample transcription factor binding sites allows their common contextual and context-dependent properties used for prediction of potential binding sites to be revealed. We have recently described SITECON (Oshchepkov, 2004a) (http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon/), our new development for determining the conservative context-dependent conformational and physicochemical properties of in transcription factor binding sites alignments. The properties so determined can be efficiently used for enhancement of binding site prediction accuracy. We have previously demonstrated how the method works on the binding sites for heterodimeric complex E2F/DP (Oshchepkov, 2004b). The discovered specific conservative properties for a set of these binding sites reflect the molecular mechanism of the heterodimer-DNA interaction.
We herein demonstrate SITECON performance on transcription factor SF-1 binding sites. The transcription factor SF-1 belongs to the family of nuclear receptors and binds to DNA as a monomer (Val et al., 2003). This factor plays a key role in the transcriptional regulation of steroidogenesis genes and is required for normal development of the hypothalamic-pituitary-adrenal and gonadal complexes (Busygina et al., 2003; Val et al., 2003). Experimental verification of SITECON predictions was performed and the location of the predicted sites was compared to the location of the functional SF-1 sites in the orthologous genes; the descriptions to these functional sites were taken from the literature.

METHODS AND ALGORITHMS

Nucleotide sequences of SF-1 binding sites and 5′-flanking regions. The training sample comprised the nucleotide sequences of 54 experimentally identified SF-1 binding sites retrieved from the TRRD database (Kolchanov et al., 2002). We were searching 5′-flanking regions of genes in two groups: a) 33 steroidogenesis genes with no experimental evidence for SF-1 binding sites in their regulatory regions; b) genes orthologous to those in first group that, according to TRRD, contain experimentally identified binding sites for SF-1.

SITECON. As the detection threshold, SITECON employs conformational similarity (Oshchepkov et al., 2004a), which was 94% for SF-1. The sensitivity to type I errors was assessed using the jack-knife method: sequences were removed from the training sample one by one in a series of iterations and served as controls. Type II errors were assessed based on the number of binding sites predicted to be present in a negative sequence 500,000 bp in length. That negative sequence was generated by random shuffling of the nucleotides of the sequences in the training sample; thus, the nucleotide compositions of both the positive and negative samples were identical and the search was made in both directions. Evaluation of type I and II is shown in Table 1.

Table 1. Errors in SF-1 binding site prediction by SITECON calculated for various conformational similarities

<table>
<thead>
<tr>
<th>Similarity (%)</th>
<th>Type I error</th>
<th>Type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td>92.00 %</td>
<td>0.30 (1/1368)</td>
<td>7.31E-04 (1/1368)</td>
</tr>
<tr>
<td>93.00 %</td>
<td>0.39 (1/1915)</td>
<td>5.22E-04 (1/1915)</td>
</tr>
<tr>
<td>94.00 %</td>
<td>0.56 (1/4484)</td>
<td>2.23E-04 (1/4484)</td>
</tr>
<tr>
<td>95.00 %</td>
<td>0.70 (1/14347)</td>
<td>6.97E-05 (1/14347)</td>
</tr>
</tbody>
</table>

Experimental verification of the potential binding sites for SF-1. For verification purposes, a gel retardation assay of labeled 32-bp double-stranded oligonucleotide probes corresponding to the predicted binding sites was performed. The source of SF-1 was testicle cell nuclear extracts from Wistar rats aged 14 days. If the corresponding retardation bands disappeared after adding antibodies to SF-1 (Upstate), the presence of SF-1 in DNA-protein complexes was assumed.

IMPLEMENTATION AND RESULTS

Detection of new potential binding sites for SF-1 in steroidogenesis gene promoter regions with SITECON and experimental ascertainment. SITECON detected 15 new SF-1 binding sites in the promoter regions of 33 steroidogenesis genes (Table 2A). These promoter sequences had previously not been tested for binding with SF-1. Additionally, we tested three more new potential binding sites predicted in 5′-flanking gene regions, which, according to TRRD, contained experimentally identified SF-1 binding sites (Table 2B). Two of these potential binding sites were predicted to be located in the human and rat Cyp17 genes (at positions –44 and –309, respectively).
SITECON suggests that the conformational similarity between the third potential binding site, which is at position –54 in the pig LHbeta gene, and the sequences of the known SF-1 binding sites is below the accepted threshold value. Because that binding site was located similarly to the known SF-1 binding sites in the orthologous (bovine, horse and rat LHbeta) genes, it was tested, too. All the predicted sites were tested by a gel retardation assay with antibodies. The ability to interact with SF-1 was confirmed for all the 18 binding sites (Table 2).

Table 2. Potential binding sites for SF-1 predicted with SITECON in steroidogenesis gene and experimentally ascertained

<table>
<thead>
<tr>
<th>Gene</th>
<th>SF-1 binding site position*</th>
<th>p**</th>
<th>Confirmed experimentally</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Cyp17 (Mouse)</td>
<td>-283</td>
<td>0.944</td>
<td>+</td>
</tr>
<tr>
<td>2 Cyp17 (Mouse)</td>
<td>-49</td>
<td>0.949</td>
<td>+</td>
</tr>
<tr>
<td>3 Ad (Bovine)</td>
<td>-428</td>
<td>0.962</td>
<td>+</td>
</tr>
<tr>
<td>4 Cyp11B1 (Guinea pig)</td>
<td>-126</td>
<td>0.945</td>
<td>+</td>
</tr>
<tr>
<td>5 Cyp11B3 (Rat)</td>
<td>-309</td>
<td>0.945</td>
<td>+</td>
</tr>
<tr>
<td>6 Cyp11B1 (Sheep)</td>
<td>-337</td>
<td>0.947</td>
<td>+</td>
</tr>
<tr>
<td>7 Oxt (Mouse)</td>
<td>-164</td>
<td>0.966</td>
<td>+</td>
</tr>
<tr>
<td>8 Oxt (Rat)</td>
<td>-167</td>
<td>0.962</td>
<td>+</td>
</tr>
<tr>
<td>9 Oxt (Human)</td>
<td>-159</td>
<td>0.961</td>
<td>+</td>
</tr>
<tr>
<td>10 Cyp11B2 (Rat)</td>
<td>-324</td>
<td>0.951</td>
<td>+</td>
</tr>
<tr>
<td>11 HSD3b (Mouse)</td>
<td>-113</td>
<td>0.942</td>
<td>+</td>
</tr>
<tr>
<td>12 Ad4BP/SF-1 (Mouse)</td>
<td>-224</td>
<td>0.952</td>
<td>+</td>
</tr>
<tr>
<td>13 CYP17 (Porcine)</td>
<td>-51</td>
<td>0.946</td>
<td>+</td>
</tr>
<tr>
<td>14 HSD17B1 (Rat)</td>
<td>-84</td>
<td>0.941</td>
<td>+</td>
</tr>
<tr>
<td>15 LH beta (Porcine)</td>
<td>-114</td>
<td>0.959</td>
<td>+</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 CYP17 (Human)</td>
<td>-44</td>
<td>0.944</td>
<td>+</td>
</tr>
<tr>
<td>17 CYP17 (Rat)</td>
<td>-309</td>
<td>0.944</td>
<td>+</td>
</tr>
<tr>
<td>18 LHbeta (Ss)</td>
<td>-58</td>
<td>0.928</td>
<td>+</td>
</tr>
</tbody>
</table>

* Position is given relative the transcription start site. **Conformational similarity to the known SF-1 binding sites as assessed by SITECON.

Analysis of transcription factor SF-1 binding site localization in the regulatory regions of orthologous genes. We compared the locations of the SF-1 binding sites in the regulatory regions of the orthologous genes using TRRD, experimental data published in the literature, and our results. The regulatory regions of five orthologous groups are presented in Fig. 1. In case of the Cyp17, LHeta, Cyp11B1 and Oxt genes (Fig. 1a, b, c, d), the predicted binding sites are at the similar positions as in the orthologous genes.

DISCUSSION

It was experimentally confirmed that all the binding sites predicted with SITECON (15 at the first stage (Table 2A) and two at the second stage (Table 2B)) are able to interact with SF-1. Additionally, the –60/–50 region of the pig LHbeta gene with the conformational similarity to the known SF-1 binding sites below the threshold (0.94), too, was found to be able to interact with that transcription factor. Analysis of the regulatory regions of steroidogenesis genes suggests that the positions of most of the predicted SF-1 binding sites are similar to those of the known, experimentally identified SF-1 binding sites with proven functionality. These and our experimental data on SF-1 binding provide strong evidence that the predicted sites must be functional. The new genes that we have revealed as potential targets to SF-1 seem to be worthy of experimental verification for functionality.
Figure 1. The regulatory regions in groups of orthologous genes for steroidogenesis. The curved arrow indicates the transcription start site. To the left above the designation of each sequence, the EMBL acc. number is indicated. Species designation: Hs, Homo sapiens; Mm, Mus musculus; Rn, Ratus norvegicus; Bt, Bos taurus; Ss, Sus scrofa; Ec, Equus caballus; Oa, Ovis aries; Cp, Cavia porcellus. 0.928 is the SITECON-based conformational similarity to the known SF-1 binding sites in the training sample.

ACKNOWLEDGEMENTS

The work was supported by Integration Project no. 119 from SB of RAS, State Contract with the Federal Agency for Science and Technology “Identification of potential targets for novel medicinal drugs based on reconstructed gene networks”, Frontiers in Genetics “Living Systems”, Innovation Project It-CP.5/001 “Development of software for computer modeling and design in postgenomic systems biology (in silico systems biology)” from the Federal Agency of Science and Innovation. The authors are grateful to V. Filonenko for translating this paper from Russian into English and to A.V. Osadchuk and T.V. Busygina for fruitful discussion.

REFERENCES


