**SCPD: a promoter database of the yeast Saccharomyces cerevisiae**

Jian Zhu and Michael Q. Zhang

Cold Spring Harbor Laboratory, PO Box 100, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

Received on October 6, 1998; revised on November 25, 1998; accepted on December 3, 1998

**Abstract**

**Motivation:** In order to facilitate a systematic study of the promoters and transcriptionally regulatory cis-elements of the yeast Saccharomyces cerevisiae on a genomic scale, we have developed a comprehensive yeast-specific promoter database, SCPD.

**Results:** Currently SCPD contains 580 experimentally mapped transcription factor (TF) binding sites and 425 transcriptional start sites (TSS) as its primary data entries. It also contains relevant binding affinity and expression data where available. In addition to mechanisms for promoter information (including sequence) retrieval and a data submission form, SCPD also provides some simple but useful tools for promoter sequence analysis.

**Availability:** SCPD can be accessed from the URL http://cgsigma.cshl.org/jian. The database is continually updated.

**Contact:** zhuj@cshl.org or mzhang@cshl.org

**Introduction**

The complete genomic sequence of the yeast *Saccharomyces cerevisiae* reveals >6000 open reading frames (ORFs). About 3000 have been assigned functions (Goffeau et al., 1996). The transcription is largely controlled and regulated by their promoter region located upstream of the coding regions (e.g. Hampsey, 1998). There are ~200 known yeast transcription factors including activators/repressors and co-activators/co-repressors, in addition to basal factors.

The availability of the complete genome enables large-scale functional studies using DNA microarray and oligonucleotide chip technologies, in which the expression pattern of >6000 genes can be simultaneously monitored (DeRisi et al., 1997). To facilitate the gene regulation analysis of the large-scale expression data, it is imperative to build a database representing the current knowledge of yeast promoters. We constructed SCPD, a promoter database of the yeast *S.cerevisiae*. Since SCPD is based on published results of individual genes, it can be used either to complement or substantiate large-scale genomic expression data. Furthermore, its information on conserved sequence patterns of transcription factor (TF) binding sites can be used to map putative sites in uncharacterized promoter regions. SCPD provides more up-to-date information specific to yeast than other databases such as TRANSFAC (Wingender et al., 1996), TRRD (Heinemeyer et al., 1998), EPD (Cavin Perier et al., 1998) and TFD (Ghosh, 1998). All sequences in SCPD refer to the corresponding genomic records in SGD (Cherry et al., 1998) and GENBANK (Benson et al., 1998). Since for most yeast genes the transcriptional start site (TSS) is not mapped, the location of a DNA element in the promoter region is defined relative to the translational start site (A of ATG is at +1). SCPD incorporates differences in results of DNA footprinting studies, terminologies (such as alternative gene names) and reference sources. Mechanisms are also provided for submitting data directly into SCPD by users.

To date, SCPD contains the information on ~6223 open reading frames and 2921 experimentally characterized genes. It contains >1000 site records. Table 1 compares some of the SCPD primary data with the yeast-related portion of TRANSFAC (Release 3.4).

**Table 1. Comparison between SCPD and TRANSFAC (Release 3.4)**

<table>
<thead>
<tr>
<th></th>
<th>TRANSFAC</th>
<th>SCPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>139</td>
<td>205</td>
</tr>
<tr>
<td>Sites</td>
<td>312 (include artificial sites)</td>
<td>580</td>
</tr>
<tr>
<td>Factors</td>
<td>159 (include those with no mapped sites)</td>
<td>103</td>
</tr>
<tr>
<td>TSS</td>
<td>N/A</td>
<td>425</td>
</tr>
<tr>
<td>Coordinates</td>
<td>As published</td>
<td>First base of coding region at +1</td>
</tr>
<tr>
<td>Sequence</td>
<td>As published</td>
<td>Referred to SGD</td>
</tr>
<tr>
<td>Consensus sequences</td>
<td>21</td>
<td>48</td>
</tr>
<tr>
<td>Matrices</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>Affinity data</td>
<td>N/A</td>
<td>7</td>
</tr>
</tbody>
</table>
Conceptual design of SCPD

SCPD was created in two phases. The first was to build an entity-relationship (ER) model, which was used to explain the storage of the promoter data. The second was to construct an object model, which describes the dynamic features.

The ER model of SCPD is given in Figure 1a. It contains the following entities: gene, factor, matrix, consensus, affinity and putative_site, which are shown in boxes in Figure 1a. The ovals linked to entities represent attributes. Double ovals denote multi-value attributes. For example, the name attribute of entity gene can have many alternative names.

In entity gene, intergenic_region defines the upstream region of a gene up to the boundary of its neighboring gene or ORF. There are two types of intergenic regions. One consists of the 5 regions of two genes, which are expected to be transcriptionally co-regulated. The other type consists of the 3 region of one gene and the 5 region of the other. In this case, only one gene is expected to be transcriptionally regulated. However the transcriptional termination signals of the other gene may exist as well. Promoter_region contains a graphical view of a promoter region in which the binding sites are highlighted.
In SCPD, there is no distinction between the name of a transcription factor and the name of its binding site. In entity factor, coordinate is composed of three sub-attributes: start, end and orientation. By default, the orientation is the forward strand. A ‘c’ in orientation indicates the complementary strand.

There are seven relationships defined in SCPD, denoted by diamonds. Correlate represents the correlation between fac-
tors. The correlation between factors is described by the concurrence frequency, which is defined as the number of genes having both factors’ sites over the total number of genes having at least one factor’s site. Predicted by is a ternary relationship among putative site, matrix and consensus pattern. A putative site may be predicted by either a matrix or a consensus pattern. Matrices and consensi are derived from the alignment of available mapped sites of individual factors. The consensus patterns are determined using similar approach described in Fondrat and Kalogeropoulos (1996). Other relationships are obvious in Figure 1a.

The ER model explains the storage of SCPD data. The entries of SCPD are organized either in plain text files or in a relational database operated by mSQL (Jepson and Hughes, 1998). However, the ER model is inadequate to describe the dynamic features of SCPD. Figure 1b shows the object model of SCPD. There are a variety of classes in SCPD. Boxes denote classes. The name of each class is shown in the first compartment, variables in the second and methods in the third. The lines between classes represent relationships. The numbers above the lines define their cardinalities. The object model maps well onto a WWW interface in which the links represent for objects, buttons for method, and input text field or area for parameters.

**General features of yeast transcription factors’ binding sites and regulatory elements**

SCPD enables us to study the general features of yeast transcription factors’ binding sites.

Many factors have multiple binding sites in their upstream regions. For the 200 genes (455 non-redundant sites) documented in SCPD, 203 sites are in single copy, 69 in two copies, 19 in three, 9 in four, 3 in five and 1 in six. Sites with a large copy number (such as six) are very rare. The lengths of binding sites range from 5 to 30 bp. The majority (78%) are located between 5 and 16 bases. This is the most likely range for detecting other novel sites. It is also worth mentioning that the length of a site may depend on the experimental method used to map it.

To answer a question such as where to find TF sites, one first needs to know where RNA transcription starts. SCPD contains 425 entries of TSS mapped in 172 genes. On average, each gene has ~2–3 mapped transcription start sites. Among these start sites, 183 (43%) start from A, 78 (18%) from C, 71 (17%) from G and 93 (22%) from T. The consensus sequence of TSS is rather loose. Figure 2 shows the distribution of these mapped TSSs. The translational start site is at 0. The median TSS location is ~61 bp upstream of the translational start site, with the mean at 89 bp. Secondly, one may want to know where the binding site for the pre-initiation complex (PIC) is (see, for example, Roeder, 1996). Since TBP (TATA box binding protein) is a part of the PIC, the TATA box should be a good indicator of the PIC location. Figure 3 shows the distribution of 22 known TBP binding sites in the upstream regions of 18 genes. The median TATA box position is 177 bp, and the mean at 125 bp upstream of the translational start site. The distance between TSS and TATA box is defined as the number of base pairs between the first base of TATA box and the TSS. For this distance calculation, we used information on 12 genes (HIS4, UGA4, SUC2, CYC1, CTS1, HSC82, ADH2, ARG1, ARG8, HIS3, CLN2 and GAL80). All of these genes have only one mapped TATA box. We did not include the information on the GCY1 gene, which has two TATA boxes and five TSS. The average
The authors would like to thank Dr J. Tabaska for reading the manuscript and anonymous referees for critical suggestions. This work was supported by Public Health Service grant HG01696 from NIH/NIHGR (to M.Q.Z.).

Acknowledgements

Fig. 5. Main page of SCPD.

distance is $62 \pm 30$ bp. This is consistent with the observation that the distance between yeast TATA element and mRNA initiation site ranges between 40 and 120 bp (Struhl, 1987).

In contrast, for vertebrate promoters, the distance between the TATA box and the TSS is only $\sim 25$–30 bp (Bucher, 1990).

Figure 4 summarizes the distribution of all experimentally mapped sites (not including TSS) documented in SCPD. The majority were found in a range from 10 to 700 bp upstream of the translational start site.

Figure 5 shows the screen shot of SCPD home page. A number of simple but useful analysis tools are provided (see on-line documentation for their usage). They may help users to retrieve promoter sequences, identify known motifs and predict putative sites. Tools like K-tuple relative information and Gibbs sampler can be used to find new promoter elements in the co-regulated gene cluster analysis of large-scale gene expression experiments [examples may be found in Zhang (1999)].

References


