

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://cgsig-ma.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 transcrip-

tion 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)

	TRANSFAC	SCPD
Genes	139	205
Sites	312 (include artificial sites)	580
Factors	159 (include those with no mapped sites)	103 (with at least one mapped site)
TSS	N/A	425
Coordinates	As published	First base of coding region at +1
Sequence	As published	Referred to SGD
Consensus sequences	21	48
Matrices	21	24
Affinity data	N/A	7

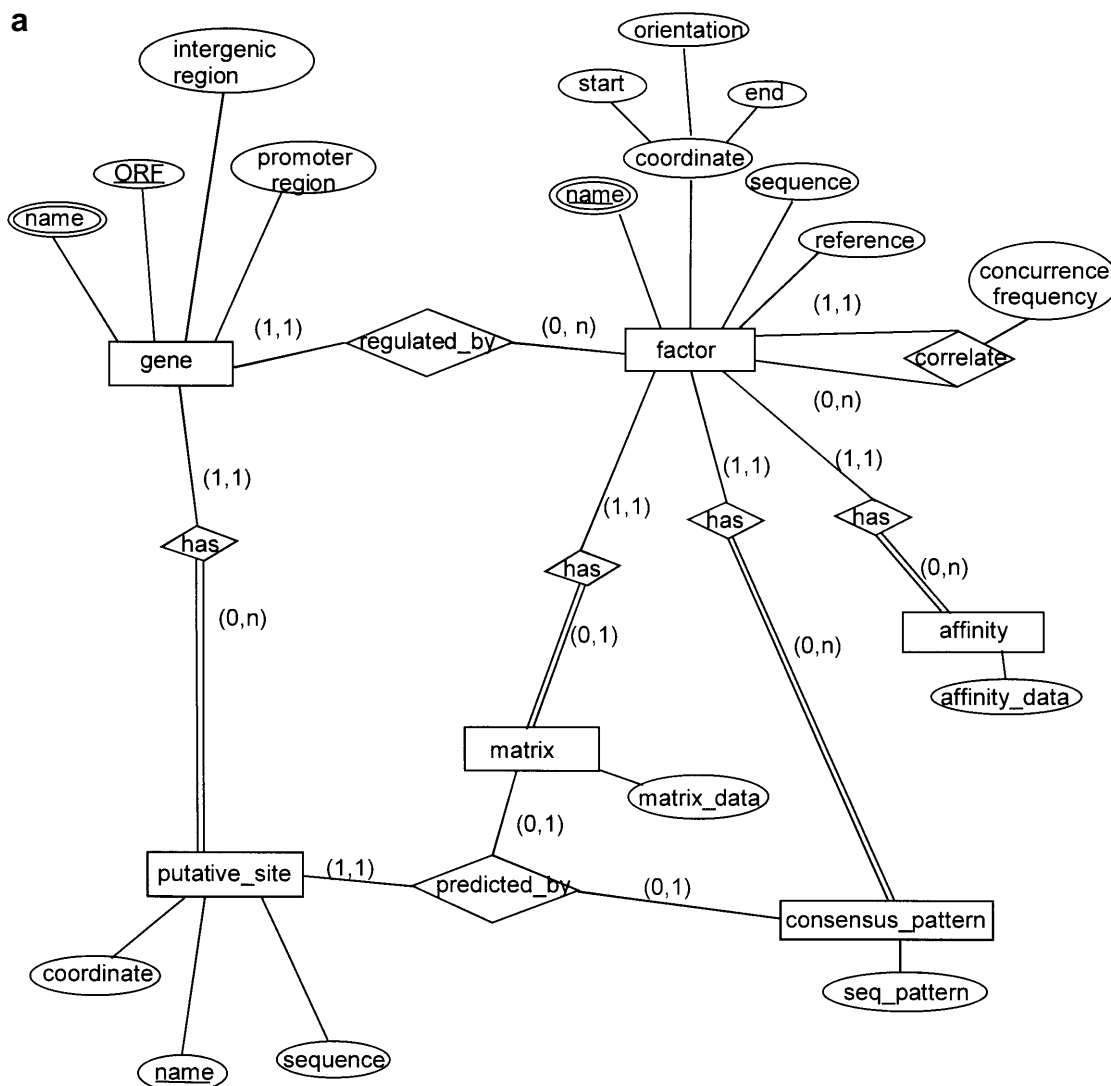


Fig. 1. Design of SCPD. (a) The ER model of SCPD. Box, entity; diamond, relationship; single oval, single value attribute; double oval, multi-value attribute. Entities and relationships are connected by lines. Double lines denote a dependent relationship. (b) The object model of SCPD. Classes are denoted by boxes. For each class, the first compartment denotes the name of the class, the second compartment contains the variables and the third compartment contains methods.

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 Figure 1a. The ovals linked to entities represent attributes. Double ovals denote multivalued 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 transcriptional 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.

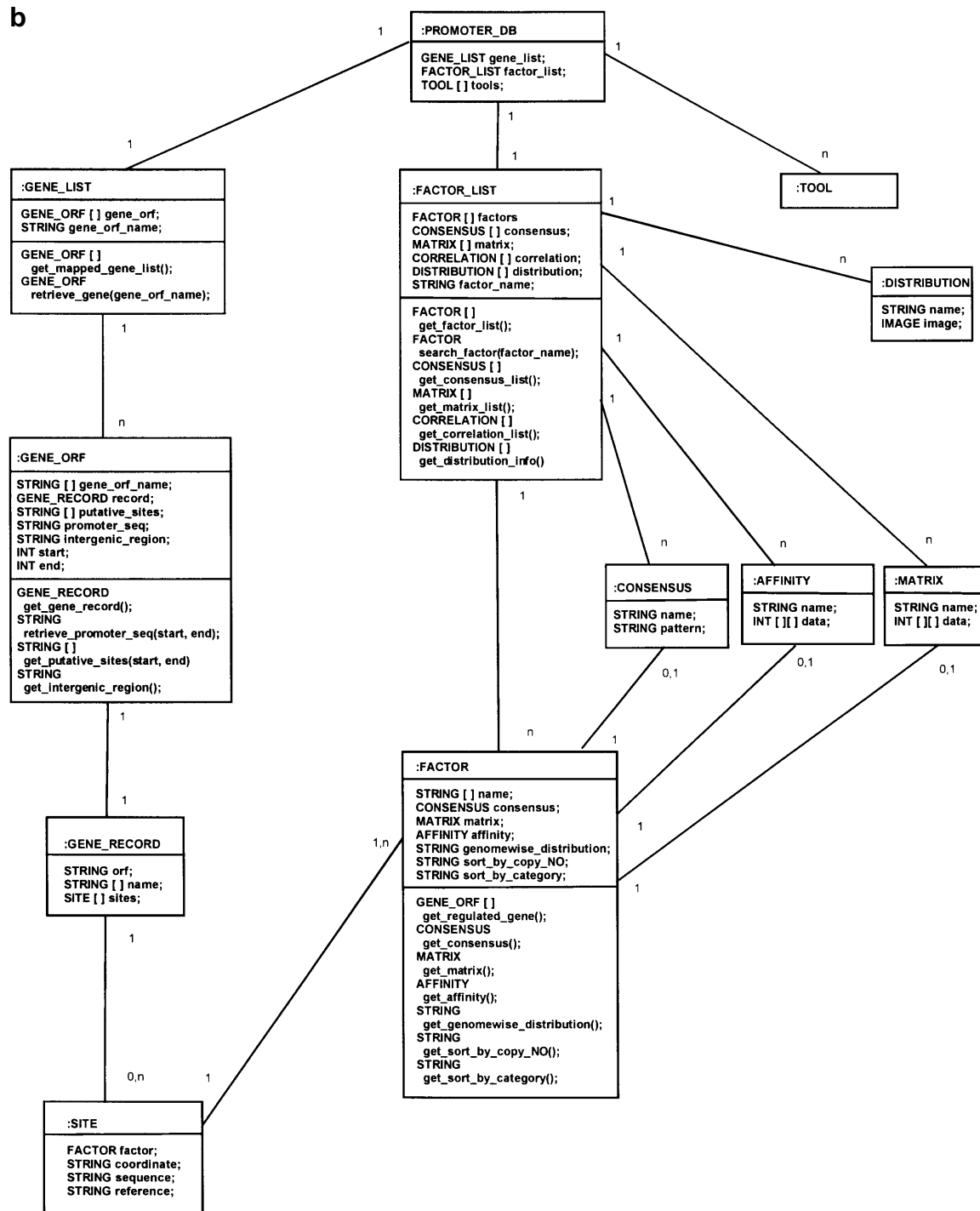


Figure 1. Continued

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-

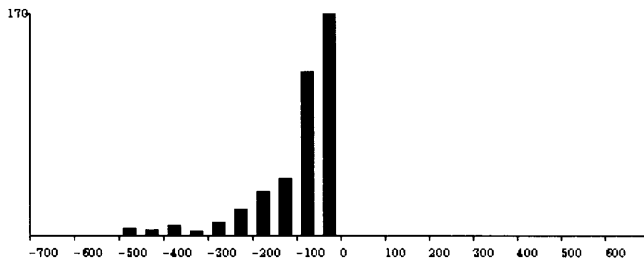


Fig. 2. Distribution of transcription start sites. The start of the translational start site is at 0. Most transcription start sites locate 50–100 bases upstream of the coding region. The median location is –61, while the mean is –89.

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

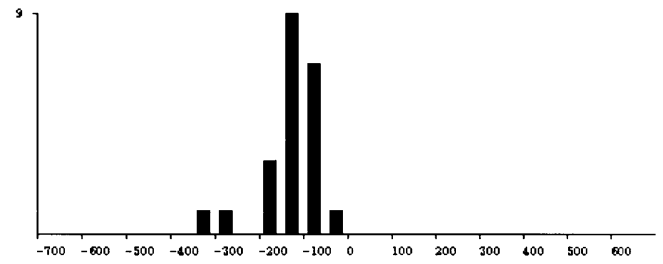


Fig. 3. Distribution of TATA boxes from 18 genes. The start of translation is at 0. The median location is –177, the mean is –125.

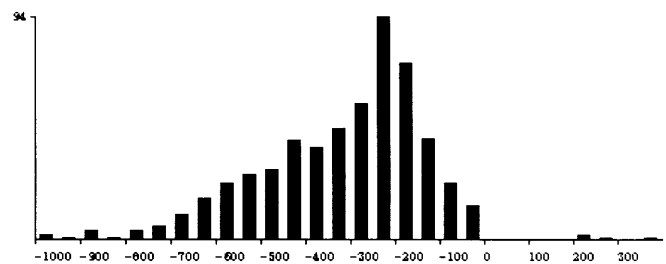


Fig. 4. Distribution of all mapped sites documented in SCPD. The translational start is at 0.

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



SCP The Promoter Database of *Saccharomyces cerevisiae*

- **Genes:** Explore the promoter regions of ~6000 genes and ORFs in yeast genome
 - Provide information on genes with mapped regulatory regions
 - Annotate putative regulatory sites of all genes and ORFs
 - Locate intergenic regions
 - Retrieve sequence of the promoter region
- **Regulatory elements and transcriptional factors**
 - Provide information on transcriptionally related genes
 - Matrix and Consensus sequence
 - Correlation between elements
 - Binding affinity and expression
 - Genomewise distribution
 - [List of transcription factors by categories](#)
 - [List of transcription factors by names](#)
- **Analysis tools**
 - [Retrieve promoter sequences](#)
 - [Search existing motifs](#)
 - [Search putative regulatory elements](#)
 - [Using predefined matrix and consensus](#)
 - [Using self-defined consensus](#)
 - [Using self-defined matrix](#)
 - [Group genes according to function categories](#)
 - [Repetitive sequence analyzer \(Java\)](#)
 - [Motif distribution \(Java\)](#)
 - [K-tuple relative information](#)
 - [Palindromic sequences](#)
 - [Multisequence alignment by GibbsDNA](#)
- **Submit records to SCPD**
 - [Submit a gene record](#)
 - [Submit a consensus record](#)
 - [Submit a matrix record](#)
- **Miscellaneous features**

Fig. 5. Main page of SCPD.

distance is 62 ± 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 ~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)].

Acknowledgements

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.).

References

- Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J. and Ouellette, B.F.F. (1998) GenBank. *Nucleic Acids Res.*, **26**, 1–7.
- Bucher, P. (1990) Weight matrix descriptions of four eukaryotic RNA II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.
- Cavin Perier, R., Junier, T. and Bucher, P. (1998) The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.*, **26**, 353–357.
- Cherry, M. *et al.* (1998) SGD: *Saccharomyces Genome Database*. *Nucleic Acids Res.*, **26**, 73–79. <http://genome-www.stanford.edu/Saccharomyces/>
- DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
- Fondrat, C. and Kalogeropoulos, A. (1996) Approaching the function of new genes by detection of their potential upstream activation sequences in *Saccharomyces cerevisiae*: application to chromosome III. *Comput. Appl. Biosci.*, **12**, 363–374.
- Ghosh, D. (1998) OOTFD (Object-Oriented Transcription Factors Database): an object-oriented successor to TFD. *Nucleic Acids Res.*, **26**, 360–362.
- Goffeau, A. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546.
- Hampsey, M. (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol. Mol. Biol. Rev.*, **62**, 465–503.
- Heinemeyer, T. *et al.* (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.*, **26**, 362–367.
- Jepson, B. and Hughes, D. (1998) *Official Guide to Mini SQL*. John Wiley & Sons, New York. The URL is <http://www.Hughes.com.au/>
- Roeder, R.G. (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.*, **21**, 327–335.
- Struhl, K. (1987) Promoter, activator proteins, and the mechanism of transcriptional initiation in yeast. *Cell*, **49**, 295–297.
- Wingender, E., Dietze, P., Karas, H. and Knuppel, R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
- Zhang, M.Q. (1999) Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem.*, in press.