ABSTRACT

Summary: TFExplorer is a web-based integrated database for predicted regulatory elements in human, mouse and rat. It shows predicted binding sites of transcription factors in the promoter regions, along with their phylogenetic footprinting information. In addition, TFExplorer can search for genes that have a given sequence pattern in their promoter regions using the motif-searching method.

Availability: TFExplorer is freely available at http://mars.kribb.re.kr:8080/tfExplorer/
Contact: sskimb@kribb.re.kr

FEATURES

TFExplorer provides various useful information about cis-transcription regulatory elements. The major features are as follows:

• Upstream promoter sequences for all the RefSeq known genes in human, mouse and rat.
• Predicted transcription factor binding sites in the promoter regions of known genes in graphical display.
• Multiple sequence alignment information of transcription factor binding sites.
• Motif search for genes that have a specific sequence pattern in their promoter region.
• Homologous gene information among human, mouse and rat.
• In addition, several cross reference links such as LocusLink, RefSeq, and the UCSC Genome Browser for the detailed and comprehensive information about genes.

INTRODUCTION

The genes in the genome are selectively expressed by transcription factors that bind to their promoter regions. Transcription factors recognize specific sequence patterns on the promoter region and bind to them for the regulation of gene expression (Pennacchio and Rubin, 2001). The regulatory factors that bind to an interesting gene provide an important information in understanding the function of that gene.

Currently, several databases, such as TRANSFAC (Matys et al., 2003), JASPAR (Sandelin et al., 2004) and EPD (Schmid et al., 2004), provide cis-transcription regulatory information. Whereas they are limited to the experimentally verified data obtained by searching published literature. HomGL (Bluthgen and Kielbasa, 2004) provides promoter sequences with homologous information, but it does not offer in-depth information on binding sites.

TFExplorer provides putative transcription factor binding sites for all the RefSeq (Pruitt and Maglott, 2001) known genes of human, mouse and rat (Fig. 1). Binding sites were predicted by MATCH program in TRANSFAC, which is one of the most popular transcription factor databases. However, the current approaches for the prediction of transcription factor binding sites are plagued with many false positives. The comparative information from multiple species provides a promising approach to alleviate those problems, because biologically functional regions are highly conserved in evolution when compared with other regions (Moses et al., 2004; Wasserman and Sandelin, 2004). For this reason, TFExplorer offers the comparative information derived from the multiple sequence alignments among human, mouse and rat. In addition, it provides a functionality of searching for genes that have a specific sequence pattern in their promoter regions by using the motif-searching method.

METHODS

We used several databases such as the UCSC Genome Browser, TRANSFAC and NCBI databases to construct TFExplorer. For the prediction of transcription factor binding sites, 3000 bp upstream and 2000 bp downstream regions from the transcription start site were obtained from the UCSC Genome Browser for each RefSeq gene. We used these sequences to predict sites bound by transcription factors. The MATCH program in the TRANSFAC 8.1 was used to predict binding sites. We set the parameters for the MATCH program stringently. Only the high-quality matrices in vertebrates are
Fig. 1. Visualization of transcription factor binding sites on the promoter region with comparative information. TFExplorer provides a variety of information on genes, such as promoter sequences, homologous information and many cross links, as well as transcription factor binding site information.

selected for the prediction, and the option for minimizing false positives was chosen.

To get the comparative information on conservation, we used the genome-to-genome multiple alignments among human, mouse and rat that were provided in the UCSC Genome Browser. Their pairwise and multiple alignments were computed using BLASTZ and MULTIZ (Blanchette et al., 2004).

For each binding site, we cut the sequence fragment from the multiple alignments and was scored to show the degree of conservation with the same scoring matrices used by the UCSC Genome Browser. Approximately 63% of predicted binding sites contain comparative information. The mean for the predicted binding site is 33.9, whereas the mean scores of exon regions and non-binding upstream regions are 71.16 and 25.27, respectively. Non-binding upstream regions are promoter areas where no transcription factor binding site was predicted. As expected, exon regions are more conserved than the other regions. The mean score 49.87, the 10% quantile of scores in exon regions is located at the 65% quantile of scores in the predicted binding sites and the 98% quantile in non-binding regions. According to these results, a significant portion of putative transcription factor binding sites can be eliminated by adjusting the cut-off values. More detailed information about the scoring strategy and the results are provided in the website.

Homologous information was derived from the NCBI Homologene database (http://www.ncbi.nlm.nih.gov/Homologene/). A total of 58.4% of the human RefSeq genes have either mouse or rat orthologs, whereas 73.2% of the mouse RefSeq genes and 87.4% of the rat RefSeq genes have their orthologs.

TFExplorer uses MySQL as a database and was implemented as a Java-based web program.

DISCUSSION

TFExplorer is expected to be a useful resource, as the first step, for those biologists who are interested in the regulation of gene expression by providing a variety of information on genes as well as transcriptional regulatory information. The conservation information for each binding site could also help to remove a significant amount of false binding sites. As an application area, it would be a good approach to combine...
the transcription factor binding site data provided here and
the microarray expression data to elucidate complicated gene
regulatory mechanisms.

ACKNOWLEDGEMENT
This work was supported by Grant No. FG-5-01 of the 21C
Frontier Functional Human Genome Project from the Ministry
of Science & Technology, Korea.

REFERENCES
Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F.A.,
Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H.,
genomic sequences with the threaded blockset aligner. Genome
ists across species and with different accession numbers.
Bioinformatics, 20, 125–126.
genomic sequence alignments. Pac. Symp. Biocomput., 2002,
115–126.
Karolchik,D., Baertsch,R., Diekhtans,M., Furey,T.S., Hinrichs,A.,
Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J.
et al. (2003) The UCSC Genome Browser Database. Nucleic Acids
Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R.,
Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. et al.
(2003) TRANSFAC: transcriptional regulation from patterns to
motif detection by expectation-maximization on evolutionary
identify mammalian regulatory sequences. Nat. Rev. Genet., 2,
137–140.
Prüitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink:
NCBI gene-centered resources. Nucleic Acids Res., 29,
137–140.
Sandelin,A., Alkema,W., Engström,P., Wasserman,W. and
Lenhard,B. (2004) JASPAR: an open-access database for euca-
ryotic transcription factor binding profiles. Nucleic Acids Res., 32
(Database issue), D91–D94.
The Eukaryotic Promoter Database EPD: the impact of in silico
for the identification of regulatory elements. Nat. Rev. Genet., 5,
276–287.