D104–D107 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj113

TiProD: the Tissue-specific Promoter Database

Xin Chen, Jian-min Wu, Klaus Hornischer[1], Alexander Kel[1] and Edgar Wingender[1,2,*]

Centre of Bioinformatics, College of Life Sciences, National Laboratory of Protein Engineering and Plant Genetic Engineering, Peking University, Beijing 100871, China, [1]BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbüttel, Germany and [2]Department of Bioinformatics, UKG, University of Göttingen, Goldschmidtstrasse 1, D-37077 Göttingen, Germany

Received August 12, 2005; Revised October 7, 2005; Accepted October 18, 2005

ABSTRACT

TiProD is a database of human promoter sequences for which some functional features are known. It allows a user to query individual promoters and the expression pattern they mediate, gene expression signatures of individual tissues, and to retrieve sets of promoters according to their tissue-specific activity or according to individual Gene Ontology terms the corresponding genes are assigned to. We have defined a measure for tissue-specificity that allows the user to discriminate between ubiquitously and specifically expressed genes. The database is accessible at http://tiprod.cbi.pku.edu.cn:8080/index.html.

INTRODUCTION

Promoters are genomic DNA sequences that enable and control transcription of the gene(s) they are associated with. In particular in multicellular organisms, they are involved in a complex coordination of transcription under all conceivable spatio-temporal-conditional circumstances. This is achieved by their internal structure, consisting of arrays of individual protein (transcription factor) binding sites, that form a hierarchical structure of modules, i.e. functionally important and transferable TFBS combinations. In the last few years, several authors have published approaches to systematically identify modules in promoters of co-regulated, or at least co-expressed, genes (1–5). These systematic approaches have used (i) the manually annotated promoters provided by the Eukaryotic Promoter Database (EPD) (6), which is very reliable but has relatively low coverage; (ii) regions around gene starts from Ensembl (7), which ensures high coverage but provides mixed quality; or obtaining regions around the experimentally determined transcription start sites (TSSs) from DBTSS (8,9), which at present provides the best combination of coverage and quality. However, it is still cumbersome to retrieve the promoter sequences of genes that share a certain activity. As a step toward facilitating these kind of investigations, we have constructed a promoter database that allows easy retrieval of promoter sequences which share a certain tissue-specificity or any Gene Ontology (GO)-assignment, e.g. to a given biological process.

DATABASE CONTENTS

Promoter data

As anchor points for promoters, we assigned the TSSs of as many human genes as possible. Based on information available in EPD (6), DBTSS (8,9) and Ensembl (7), we assigned 'Virtual TSSs' by summarizing information from these three resources. This was necessary because collected TSSs for a given gene may be located on a sequence fragment spanning several thousand nucleotides, in some cases even more than 100 kb. Sometimes, these TSSs occur in clusters of only a few dozen nucleotides length, but often they are scattered over a large sequence range.

Therefore, an algorithm was designed to apply a set of rules to the data collection in order to find clusters of TSSs. A window of 1000 nt length was slid along the entire sequence fragment. A 'clustering score' was calculated by summing up weighted contributions from each TSS in each window. For each TSS derived from a DBTSS one-pass mRNA or an Ensembl mRNA model we give one evidence point. We assume EPD TSSs to have a higher reliability owing to the fact that they are annotated by hand and give 50 evidence points each. The weights of evidence points were additionally multiplied by a distance score: the central position is multiplied by 1, the outer positions are multiplied by 0, and all positions in between by a value taken from a cosine function, according to the distance from the center of the window. The peaks of the resulting clustering score were regarded as potential 'virtual TSSs'. However, for some of the genes only a handful of evidence points are available, resulting in multiple 'virtual TSSs', each consisting of only one or two evidence points. Therefore, for all those genes where less than 20 evidence points are available only the most 5′ 'virtual TSS' was accepted. For all other genes, peaks were accepted as 'virtual TSSs'

*To whom correspondence should be addressed. Tel: +49 (0) 551 39 14912; Fax: +49 (0) 551 39 14914; Email: edgar.wingender@bioinf.med.uni-goettingen.de

Table 1. Statistics of the TiProD data content

General	Total number of entries
Genes	12 785
Promoter sequences[a]	15 384
cDNA libraries	8547
EST expression frequency data	1 965 392
Links to UniGene	12 785
Links to GO	4403
Links to CYTOMER	52

[a]All promoter sequences have a link to the TRANSPRO database.

for which the respective sequence window contains at least 5% of all evidence points. This method of defining the virtual TSSs ensured that the equal weighting of DBTSS and Ensembl TSSs does not unduly neglect their different experimental evidence base: on average, virtual TSSs derived from Ensembl alone map onto a position which is even 1–2 nt more upstream than those derived from DBTSS.

The collection of 'virtual TSSs' determined in this way forms the basis of the commercial database TRANSPRO™ (release 2.1; http://www.biobase.de/pages/products/transpro.html), which is part of the TRANSFAC® family of databases [(10); see separate publication in this issue] and from which the TiProD sequences were derived. For the TiProD database, we have extracted −500 to +60 sequences around these 'virtual TSSs' from the corresponding sequences of the human genome. The exact location of the sequence within the human genome can be retrieved from the corresponding TRANSPRO entry.

Calculation of 'virtual TSSs' and the subsequent data extraction are fully automated processes; whenever conflicts or inconsistencies occur the respective gene is excluded from the database. Presently, TiProD provides data about 15 384 human promoters (Table 1).

Tissue-specificity of the promoters

The sequencing and analysis of expressed sequence tags (ESTs) is one of the most important techniques used to reveal gene expression profiles. Currently there are over 25 million EST sequences in NCBI's dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/). Because the same gene may be represented by many different EST sequences, the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene) was developed to subsume nucleotide sequences into a non-redundant set of gene-oriented clusters (11,12). Furthermore, the EST servers that do exist, including Digital Differential Display (DDD, http://www.ncbi.nlm.nih.gov/UniGene/info_ddd.html), cDNA Digital Gene Expression Displayer (DGED, http://cgap.nci.nih.gov/Tissues/GXS) and xProfiler (http://cgap.nci.nih.gov/Tissues/xProfiler), all aim at finding differentially expressed genes in different pools of tissues or samples.

CGAP represents expression strengths in terms of numbers of ESTs for each gene in each pool. These numbers are used in TiProD as additional filtering criterion. We have parsed the UniGene Library Data and Expression Data files from CGAP (http://cgap.nci.nih.gov/Info/CGAPDownload/) (13,14) and loaded the data into a relational database (MySQL). Since TRANSPRO promoter sequences are assigned to UniGene entries as well, these common links were used to connect the promoter sequences of a gene with the corresponding expression data. The overall schema of the database and the data flow are shown in Figure 1. Altogether, TiProD holds information about 52 tissues and their gene signatures (Table 1). All these tissues (organs and cells) have been mapped onto the proper items of the CYTOMER database (15,16).

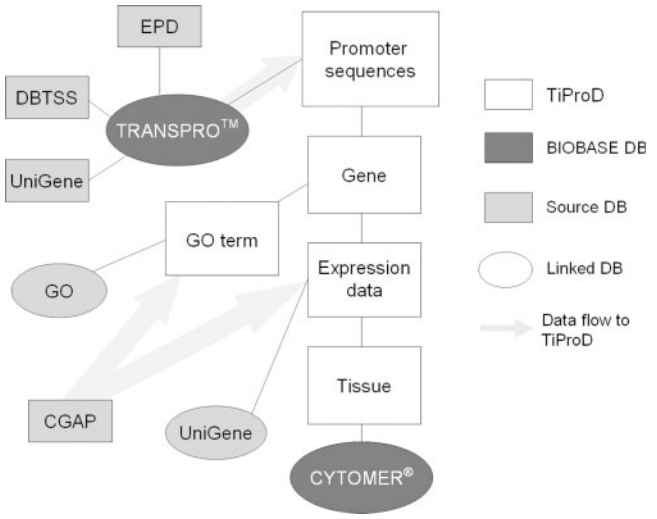

Figure 1. Data flow and schema of the TiProD database. Indicated are the main tables of the database (open rectangles), the used source databases (shaded ovals and rectangles), as well as the links between them (straight lines) and the data flow to TiProD (light gray arrows).

To enable the selection of tissue-specific promoters from the database we have computed an index of tissue-specificity for each gene in each tissue library.

Let n_{ij} be a number of ESTs of gene i ($I = 1, G$) in the tissue library j ($j = 1, L$). First, we normalize frequencies of genes in each library and calculated an abundancy score a_{ij}:

$$a_{ij} = \frac{n_{ij}}{\sum_{i=1,G} n_{ij}}. \quad \mathbf{1}$$

Then we compute the average abundancy score $\bar{a}_i$ for each gene i among all tissues j:

$$\bar{a}_i = \frac{\sum_{j=1,L} a_{ij}}{L}. \quad \mathbf{2}$$

The tissue-specificity index t_{ij} then is the ratio of the abundancy to the average abundancy score:

$$t_{ij} = \frac{a_{ij}}{\bar{a}_i}. \quad \mathbf{3}$$

The value will be close to 1 for a gene that is expressed in a tissue at an average level compared with other tissues, but significantly higher than 1 if a gene is specifically expressed in that tissue. This specificity measure worked well and is equivalent to the 'relative expression' defined by Schug *et al*. (17) multiplied by the number of tissue libraries considered. This type of normalization helps to cope with the problems of low- versus high-expressing genes inherent to the EST methodology.

Thus, in addition to the frequency values, the tissue-specificity index can be used as another selection criteria to retrieve promoter sets.

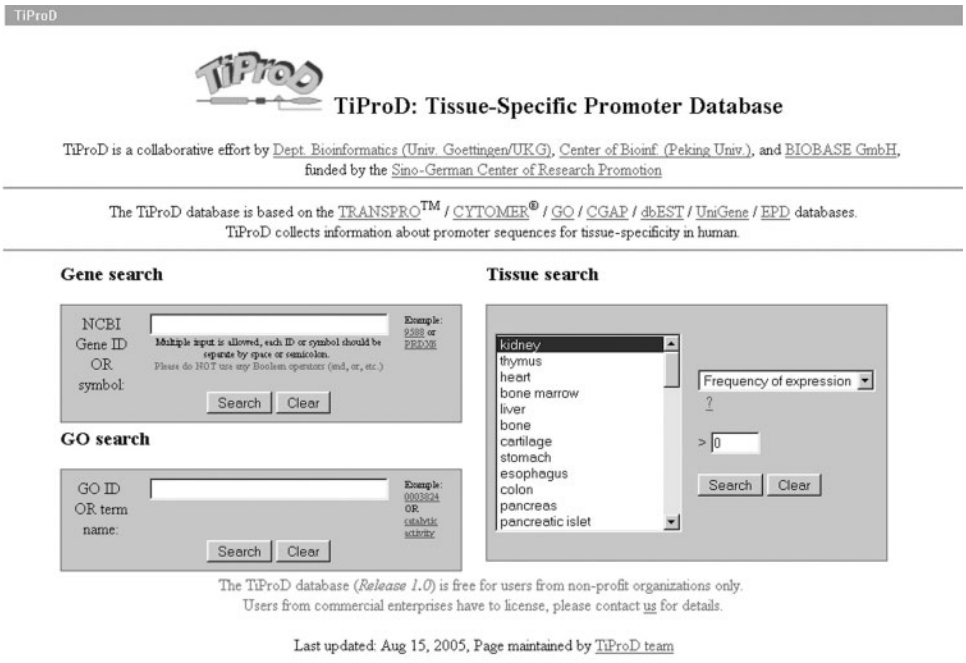

Figure 2. Screenshot of the TiProD interface.

GO assignments

The TiProD database also makes use of the linking of UniGene clusters to Gene Ontology terms of all three subontologies: biological process, cellular compartment and molecular function (18). The corresponding data were obtained from CGAP (http://cgap.nci.nih.gov/Info/CGAPDownload/).

As a result, TiProD allows the user to input an Entrez Gene ID or gene symbol and retrieve extensive functional information about the gene. Alternatively, it also allows a user to input an ID from GO (GO term) and to retrieve all UniGene clusters involved in the particular GO function. At present, TiProD strictly adheres to the hierarchical gene assignments done by the GO consortium and does not summarize subclasses with each other and with the genes directly linked to the corresponding top node. TiProD release 1.0 comprises 4403 GO term assignments (Table 1).

FUNCTIONALITY OF THE DATABASE

The major aim of the TiProD database is to retrieve promoter sequences according to their tissue-specificity or other functional groupings. Each entry in the TiProD database corresponds to a particular promoter, or a set of promoters, of a human gene and contains the gene name, a description, synonyms, Entrez Gene ID, expression information (including cDNA library ID, IDs of the tissue database CYTOMER, tissue name and expression frequency from CGAP), GO terms and the sequence of the corresponding upstream region. The TiProD interface (Figure 2) allows the retrieval of expression patterns and promoter sequences for individual human genes. It also enables the retrieval of the gene expression signature for a certain tissue (organ and cell type), including the selection either of all active genes, or of those that are specifically expressed in this tissue. Similarly, all promoters of genes that are assigned to a certain GO term (or GO ID) can be retrieved.

DISCUSSION AND FUTURE DEVELOPMENTS

Continued efforts will be made to update the promoter sequence and expression data. We are aware that working with EST data has some pitfalls since, e.g. the experimental conditions may significantly affect the detection of low-expressing genes and the ratios of expression levels among different genes in general. Therefore, we plan to include additional high-throughput as well as conventional expression data and to further refine the statistics applied. We also plan to add data about additional organisms such as mouse and rat. Also, in the next release the system will be able to deal with hierarchical dependences, for instance between organs and their substructures, or between the different levels of GO term assignments, in a more flexible manner.

ACKNOWLEDGEMENTS

This project was partially funded by Sino-German Center for Research Promotion (DFG & NFSC, GZ 186 (105/3)), by the National Key Basic Research Program of China (973 No 2003CB715900) and Natural Science Foundation of China (NSFC No 90408015) as well as grants from the China National High-tech Program (863). We also would like to thank X. Gu, J. Luo, E. Shelest, and the members of the groups at PKU and UKG for helpful discussions. We are also indebted to K. Lennon-Hopkins for her linguistic advice. The Open Access publication charges for this article were paid by our Institutional budget.

Conflict of interest statement. None declared.

REFERENCES

1. Frech,K., Quandt,K. and Werner,T. (1998) Muscle actin genes: a first step towards computational classification of tissue specific promoters. *In Silico Biol.*, **1**, 29–38.

2. Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.
3. Bailey,T.L. and Noble,W.S. (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **19** (Suppl. 2), II16–II25.
4. Kel,A., Reymann,S., Matys,V., Nettesheim,P., Wingender,E. and Borlak,J. (2004) A novel computational approach for the prediction of networked transcription factors of aryl hydrocarbon-receptor-regulated genes. *Mol. Pharmacol.*, **66**, 1557–1572.
5. Shelest,E. and Wingender,E. (2005) Construction of predictive promoter models on the example of antibacterial response of human epithelial cells. *Theor. Biol. Med. Model.*, **2**, 2.
6. Cavin Périer,R., Junier,T. and Bucher,P. (1998) The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.*, **26**, 353–357.
7. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
8. Suzuki,Y., Yamashita,R., Sugano,S. and Nakai,K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res.*, **32**, D78–D81.
9. Suzuki,Y., Yamashita,R., Shirota,M., Sakakibara,Y., Chiba,J., Mizushima-Sugano,J., Kel,A.E., Arakawa,T., Carninci,P., Kawai,J. *et al.* (2004) Large-scale collection and characterization of promoters of human and mouse genes. *In Silico Biol.*, **4**, 0036.
10. Matys,V., Fricke,E., Geffers,R., Gößling,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 profiles. *Nucleic Acids Res.*, **31**, 374–378.
11. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
12. Wu,X., Walker,M.G., Luo,J. and Wei,L. (2005) GBA server: EST-based digital gene expression profiling. *Nucleic Acids Res.*, **33**, W673–W676.
13. Lal,A., Lash,A.E., Altschul,S.F., Velculescu,V., Zhang,L., McLendon,R.E., Marra,M.A., Prange,C., Morin,P.J. and Polyak,K. (1999) A public database for gene expression in human cancers. *Cancer Res.*, **59**, 5403–5407.
14. Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.
15. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhäuser,R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.
16. Michael,H., Chen,X., Fricke,E., Haubrock,M., Ricanek,R. and Wingender,E. (2005) Deriving an ontology for human gene expression sources from the CYTOMER database on human organs and cell types. *In Silico Biol.*, **5**, 61–66.
17. Schug,J., Schuller,W.-P., Kappen,C., Salbaum,J.M., Bucan,M. and Stoeckert,C.J.,Jr (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, **6**, R33.
18. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.