322–324 Nucleic Acids Research, 2002, Vol. 30, No. 1

The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data

Viviane Praz, Rouaïda Périer, Claude Bonnard and Philipp Bucher*

Swiss Institute of Bioinformatics and Swiss Institute for Experimental Cancer Research, Ch. des Boveresses 155, 1066-Epalinges s/Lausanne, Switzerland

Received September 24, 2001; Accepted September 28, 2001

ABSTRACT

The Eukaryotic Promoter Database (EPD) is an annotated, non-redundant collection of eukaryotic Pol II promoters, for which the transcription start site has been determined experimentally. Access to promoter sequences is provided by pointers to positions in nucleotide sequence entries. The annotation part of an entry includes a description of the initiation site mapping data, exhaustive cross-references to the EMBL nucleotide sequence database, SWISS-PROT, TRANSFAC and other databases, as well as bibliographic references. EPD is structured in a way that facilitates dynamic extraction of biologically meaningful promoter subsets for comparative sequence analysis. World Wide Web-based interfaces have been developed which enable the user to view EPD entries in different formats, to select and extract promoter sequences according to a variety of criteria, and to navigate to related databases exploiting different cross-references. The EPD web site also features yearly updated base frequency matrices for major eukaryotic promoter elements. EPD can be accessed at http://www.epd.isb-sib.ch.

DATABASE DESCRIPTION

The term 'promoter' has two different meanings in biology: (i) a gene region immediately upstream of a transcription initiation site; and (ii) a *cis*-acting genetic element controlling the rate of transcription initiation of a gene. The Eukaryotic Promoter Database (EPD) is a database of promoters in the former sense. Information about promoters in the latter sense can be found in other databases, such as TRANSFAC (1), ooTFD (2), TRRD (3), PlantCARE (4) and PLACE (5).

EPD was originally designed as a resource for comparative sequence analysis and as such has played an instrumental role in the characterization of eukaryotic transcription control elements (6,7), as well as in the development of eukaryotic promoter prediction algorithms (8). The main purpose of the database is to keep track of experimental data that define transcription initiation sites of eukaryotic genes. This type of functional information is linked to promoter sequences via machine-readable pointers to positions within sequences of the EMBL nucleotide sequence database (9).

EPD is a rigorously selected, curated and quality-controlled database. At present, EPD is confined to promoters recognized by the RNA Pol II system of higher eukaryotes (multicellular plants and animals). Note that this restriction does not a priori exclude viral promoters. EPD is also a strictly non-redundant database.

A comprehensive description of the contents and format of EPD has been published previously (10). User interfaces and software support for local installations were also described previously (11,12).

RECENT DEVELOPMENTS

EPDEX

The recent publication of large gene expression data sets generated with microarrays or SAGE technology prompted us to explore various mechanisms to link EPD entries to this new type of information. The solution we have chosen is a companion database called EPDEX, which maps promoters via genes to expression profiles. There is one EPDEX entry per gene. Each entry contains a brief description of the gene plus cross-references to EPD, gene expression data sets and RNA sequences in EMBL (EPD refers only to genomic DNA sequences). The references to expression data are structured in such a way that they can be hyperlinked both to the original Web servers (if existing) or to a local gene expression data archive at the Swiss Institute of Bioinformatics. So far, EPDEX is confined to human genes, and the entry names obey the guidelines published by the HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/). We plan to extend the database to other organisms, in particular *Caenorhabditis elegans*, *Drosophila* and mouse. The format of EPDEX is described in the corresponding user manual available at http://www.epd.isb-sib.ch/current/EPDEX_manual.html.

New EPD entries

Recently, we have started to process data directly submitted from genome and cDNA sequence centers, including promoter sets contributed by the *Drosophila* genome annotation team and by collaborators from the MGC (13) and NEDO full-length cDNA sequencing projects (14,15). The latter promoter collection is based on RACE cloning and 5′ EST sequencing. These

*To whom correspondence should be addressed. Tel: +41 21 692 5892; Fax: +41 21 692 5945; Email:philipp.bucher@isb-sib.ch

Table 1. Database cross-references in EPD release 67

Database	Number of links
EPD internal	252
EPDEX	79
EMBL (9)	3340
TRANSFAC (1)	1788
SWISS-PROT (19)	1152
FlyBase (20)	123
MIM (21)	258
MGD (22)	133
MEDLINE	2425

entries, which are generated in a semi-automatic fashion, will be added to the next EPD release, 68.

New features in EPD

A new line type, with line code NP (neighboring promoter), has been introduced which cross-references promoters of different genes occurring at a short distance from each other (<1000 bp). Such pairs of promoters usually promote transcription in opposite directions and often share upstream regulatory elements. As a consequence, the number of EPD internal cross-references has considerably increased since the last release (Table 1). Another new line type, IF (initiation frequency), was added to reflect the frequency at which each nucleotide within the initiation region is found at the 5′ end of bona fide full-length cDNA clone inserts originating from large-scale cDNA sequencing project (see Supplementary Material, Example 2).

ACCESS

FTP

The following files are available from ftp://ftp.epd.isb-sib.ch/pub/databases/epd.

1. Flatfiles containing the EPD database in the new and in the old format.
2. EPD user manual.
3. Copyright statement.
4. Sequence libraries in EMBL and FASTA format containing promoter sequences from –499 to +100 relative to the transcription start site.
5. A slightly reduced version of EPD in ASN.1 format designed for import into the GenBank–Entrez data environment (16), including a formal data description in ASN.1.
6. Icarus scripts for indexing EPD by SRS (17).
7. Flatfile containing the EPDEX database.
8. EPDEX user manual.

World Wide Web

The following services are offered at http://www.epd.isb-sib.ch.

1. Access to EPD and EPDEX entries by ID or accession number. The following formats are available: text only, HTML and HTML combined with a graphic representation of sequence objects by a Java applet (18).
2. A page for downloading promoter sequence subsets defined in EPD.
3. Access to EPD entries and corresponding promoter sequences via a query form.
4. Access to EPD via SRS is provided by the Swiss EMBNet node: http://www.ch.embnet.org/.

SUPPLEMENTARY MATERIAL

The following Supplementary Material is available at NAR Online: snapshot of the EPD homepage; README file of the FTP archive; Appendix A, taxonomic break down of EPD release 67; EPDEX gene list; Example 1, example of an EPDEX entry; Example 2, new EPD entry (HS_CAH2, data from the NEDO project); a general TATA-box matrix derived from EPD release 60 plus subclass-specific versions for plants, insects and vertebrates.

ACKNOWLEDGEMENTS

EPD is funded by grants from the Swiss government and the Swiss National Science Foundation (grant 31-54782.98).

REFERENCES

1. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
2. Ghosh,D. (2000) Object-oriented transcription factors database (ooTFD). *Nucleic Acids Res.*, **28**, 308–310.
3. Kolchanov,N.A., Podkolodnaya,O.A., Ananko,E.A., Ignatieva,E.V., Stepanenko,I.L., Kel-Margoulis,O.V., Kel,A.E., Merkulova,T.I., Goryachkovskaya,T.N., Busygina,T.V. *et al.* (2000) Transcription regulatory regions database (TRRD): its status in 2000. *Nucleic Acids Res.*, **28**, 298–301. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 312–317.
4. Rombauts,S., Dehais,P., Van Montagu,M. and Rouze,P. (1999) PlantCARE, a plant cis-acting regulatory element database. *Nucleic Acids Res.*, **27**, 295–296. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 325–327.
5. Higo,K., Ugawa,Y., Iwamoto,M. and Korenaga,T. (1999) Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.*, **27**, 297–300.
6. Bucher,P. and Trifonov,E.N. (1986) Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res.*, **22**, 10009–10026.
7. Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.
8. Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.
9. Baker,W., van den Broek,A., Camon,E., Hingamp,P., Sterk,P., Stoesser,G. and Tuli,M.A. (2000) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **28**, 19–23. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 21–26.
10. Cavin Périer,R., Junier,T. and Bucher,P. (1998) The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.*, **26**, 353–357.
11. Cavin Périer,R., Junier,T., Bonnard,C. and Bucher,P. (1999) The Eukaryotic Promoter Database (EPD): recent developments. *Nucleic Acids Res.*, **27**, 307–309.
12. Cavin Périer,R., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The eukaryotic promoter database (EPD). *Nucleic Acids Res.*, **28**, 302–303.
13. Strausberg,R.L., Feingold,E.A., Klausner,R.D. and Collins,F.S. (1999) The Mammalian Gene Collection. *Science*, **286**, 455–457.
14. Suzuki,Y., Taira,H., Tsunoda,T., Junko,J., Mizushima-Sugano,J., Sese,J., Hata,H., Ota,T., Isogai,T., Tanaka,T. *et al.* (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.*, **2**, 388–393.

15. Suzuki,Y., Tsunoda,T., Sese,J., Taira,H., Mizushima-Sugano,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Nakamura,Y. *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.*, **11**, 677–684.
16. Benson,D.A., Boguski,M., Lipman,D.J. and Ostell,J. (1994) GenBank. *Nucleic Acids Res.*, **22**, 3441–3444. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 17–20.
17. Etzold,T., Ulyanov,A. and Argos P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
18. Junier,T. and Bucher,P. (1998) SEView: a Java applet for browsing molecular sequence data. *In Silico Biol.*, **1**, 13–20.
19. Bairoch,A. and Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
20. The FlyBase Consortium (1999) The FlyBase database of the Drosophila Genome Projects and community literature. *Nucleic Acids Res.*, **27**, 85–88. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 106–108.
21. Hamosh,A., Scott,A.F., Amberger,J., Valle,D. and McKusick,V.A. (2000) Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.*, **15**, 57–61.
22. Blake,J.A., Richardson,J.E. and Davisson,M.T. (2000) The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. The Mouse Genome Database Group. *Nucleic Acids Res.*, **28**, 108–111. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 113–115.