302–303 *Nucleic Acids Research, 2000, Vol. 28, No. 1*

The Eukaryotic Promoter Database (EPD)

Rouaïda Cavin Périer, Viviane Praz, Thomas Junier, 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 October 6, 1999; Accepted October 8, 1999

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. WWW-based interfaces have been developed that 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. In order to be included, a promoter must fulfill a number of conditions laid down in the user manual. Most importantly, the transcription start site must be mapped experimentally with an estimated precision of ±5 bp or higher. All information in EPD originates from a critical examination and independent interpretation of the experimental data presented in the cited research publications. Published conclusions and feature table annotations in EMBL entries are never blindly relied upon. 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. The general rule is that one entry corresponds to one transcription initiation site in a genome. Organisms are distinguished at the taxonomic level of the species. According to this policy, data from different literature sources pertaining to the same transcription initiation sites are represented by the same entry. Likewise, promoters belonging to different alleles of the same gene, or to the same gene in different subspecies, are covered by the same entry regardless of whether they differ in sequence. The user manual provides more details about how certain non-trivial cases such as promoters of tandemly repeated genes or retrotransposable elements, are handled.

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

RECENT DEVELOPMENTS

Database

The objective of exhaustive cross-referencing between EPD promoters and EMBL sequences is being given high priority at the moment, especially with regard to genomes that are complete (*Caenorhabditis elegans*) or at an advanced stage of sequencing (*Arabidopsis*, *Drosophila*, human). As a consequence, the number of EMBL cross-references has increased by >1000 since last year (Table 1). Moreover, the internal EPD cross-references have been revised. Until now, such links were only used to connect alternative promoters of the same gene. In future releases, promoters of different genes occurring at a short distance from each other (<1000 bp), will also be cross-referenced. Such pairs of promoters usually promote transcription in opposite directions and often share upstream regulatory elements. As a new format feature, a keyword (KW) line type

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

has been introduced and so far been populated with keywords imported from SWISS-PROT (12). This feature is intended to enhance the query capabilities of various access tools. Additional keywords relating to properties of the promoter rather than to properties of the corresponding gene product will be added in the near future.

Table 1. Database cross-references in EPD release 60

Database	Number of links
EPD internal	188
EMBL (9)	2978
TRANSFAC (1)	1700
SWISS-PROT (12)	1058
FlyBase (16)	116
MIM (17)	234
MGD (18)	126
MEDLINE	2393

Documentation

The user manual has been extensively revised. Bibliographic references have been added to the section explaining the representation of transcript mapping data. Some of them are accompanied by direct hyperlinks to figures in online journals exemplifying a particular technique. Several additional documents have recently been made available over the web. One contains a list of all 'homology groups' defined in EPD. Such groups consist of homologous promoters exhibiting significant sequence similarity in the –79 to +20 region among themselves. Another document presents the hierarchical promoter classification system of EPD.

Promoter element descriptions

Weight matrix descriptions of four major eukaryotic promoter elements (TATA-box, initiator, GC-box and CCAAT-box) have previously been derived from EPD release 17 (7). We have now decided to make updated versions of such matrices available on a yearly basis from the EPD web pages. The latest versions were produced from EPD release 60 using a Baum–Welch hidden Markov model training algorithm (program buildmodel of SAM release 1.3.3, Hughey & Krogh 1998, http://www.cse.ucsc.edu/research/compbio/sam.html).

ACCESS

FTP

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

- Flat-files containing the EPD database in the new and in the old format.
- EPD user manual.
- Sequence libraries in EMBL and FASTA format containing promoter sequences from –499 to +100 relative to the transcription start site.
- A slightly reduced version of EPD in ASN.1 format designed for import into the GenBank–Entrez data environment (13), including a formal data description in ASN1.
- Icarus scripts for indexing EPD by SRS (14).

WWW

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

- Access to EPD 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 (15).
- A page for downloading promoter sequence subsets defined in EPD.
- Access to EPD entries and corresponding promoter sequences via a query form.

Access to EPD via SRS is provided by the Swiss EMBNet node at http://www.ch.embnet.org/

SUPPLEMENTARY MATERIAL

Relevant URL links are available at NAR Online.

ACKNOWLEDGEMENT

EPD is funded in part by grant 31-54782.98 from the Swiss National Science Foundation.

REFERENCES

1. Heinemeyer,T., Chen,X., Karas,H., Kel,A.E., Kel,O.V., Liebich,I., Meinhardt,T., Reuter,I., Schacherer,F. and Wingender,E. (1999) *Nucleic Acids Res.*, **27**, 318–322. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 316–319.
2. Ghosh,D. (1999) *Nucleic Acids Res.*, **27**, 315–317. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 308–310.
3. Kolchanov,N.A., Ananko,E.A., Podkolodnaya,O.A., Ignatieva,E.V., Stepanenko,I.L., Kel-Margoulis,O.V., Kel,A.E., Merkulova,T.I., Goryachkovskaya,T.N., Busygina,T.V., Kolpakov,F.A., Podkolodny,N.L., Naumochkin,A.N. and Romashchenko,A.G. (1999) *Nucleic Acids Res.*, **27**, 303–306. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 298–301.
4. Rombauts,S., Dehais,P., Van Montagu,M. and Rouze,P. (1999) *Nucleic Acids Res.*, **27**, 295–296.
5. Higo,K., Ugawa,Y., Iwamoto,M. and Korenaga,T. (1999) *Nucleic Acids Res.*, **27**, 297–300.
6. Bucher,P. and Trifonov,E.N. (1986) *Nucleic Acids Res.*, **22**, 10009–10026.
7. Bucher,P. (1990) *J. Mol. Biol.*, **212**, 563–578.
8. Fickett,J.W. and Hatzigeorgiou,A.G. (1997) *Genome Res.*, **7**, 861–878.
9. Stoesser,G., Tuli,M.A., Lopez,R. and Sterk,P. (1999) *Nucleic Acids Res.*, **27**, 18–24. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 19–23.
10. Cavin Périer,R., Junier,T. and Bucher,P. (1998) *Nucleic Acids Res.*, **26**, 353–357.
11. Cavin Périer,R., Junier,T., Bonnard,C. and Bucher,P. (1999) *Nucleic Acids Res.*, **27**, 307–309.
12. Bairoch,A. and Apweiler,R. (1999) *Nucleic Acids Res.*, **27**, 49–54. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 45–48.
13. Benson,D.A., Boguski,M., Lipman,D.J. and Ostell,J. (1994) *Nucleic Acids Res.*, **22**, 3441–3444.
14. Etzold,T., Ulyanov,A. and Argos,P. (1996) *Methods Enzymol.*, **266**, 114–128.
15. Junier,T. and Bucher,P. (1998) *In Silico Biol.*, **1**, 13–20.
16. The Flybase Consortium (1999) *Nucleic Acids Res.*, **27**, 85–88.
17. Pearson,P., Francomano,C., Foster,P., Bocchini,C., Li,P. and McKusick,V. (1994) *Nucleic Acids Res.*, **22**, 3470–3473.
18. Blake,J.A., Richardson,J.E., Davisson,M.T. and Eppig,J.T. (1999) *Nucleic Acids Res.*, **27**, 95–98. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 108–111.