 Nucleic Acids Research, 1999, Vol. 27, No. 1 307–309

The Eukaryotic Promoter Database (EPD): recent developments

Rouaïda Cavin Périer, Thomas Junier, Claude Bonnard and Philipp Bucher*

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

Received October 8, 1998; Revised October 13, 1998; Accepted October 22, 1998

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 description of the initiation site mapping data, cross-references to other databases, and bibliographic references. EPD is structured in a way that facilitates dynamic extraction of biologically meaningful promoter subsets for comparative sequence analysis. Recent efforts have focused on exhaustive cross-referencing to the EMBL nucleotide sequence database, and on the improvement of the WWW-based user interfaces and data retrieval mechanisms. EPD can be accessed at http://www.epd.isb-sib.ch

DATABASE DESCRIPTION

EPD is a database of gene function which keeps track of experimental evidence defining the initiation sites of eukaryotic RNA POL II genes. This information is linked to promoter sequence data via machine readable pointers to corresponding positions in entries of the EMBL nucleotide sequence database (1). Note that EPD does not provide information on promoters in the sense of genetically defined transcription regulatory elements. Such information can be found in TRANSFAC and COMPEL (2) and other databases described in this issue.

EPD was originally designed as a resource for comparative sequence analysis and as such has played an instrumental role in the development of eukaryotic promoter prediction algorithms (3). Recently, its scope has been expanded to meet the requirements of the TRADAT project (see http://www.itba.mi.cnr.it/tradat/), a European consortium effort to develop integrated tools for the interpretation of genomic DNA sequences with emphasis on regulatory regions. The extensions comprise many cross-references to data collections covering other aspects of genes and promoters.

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. For instance, its transcription start site must be mapped with a certain accuracy and certainty, the corresponding gene must be functional, and corresponding sequence data must be available in the public databases. EPD is further confined to promoters recognized by the RNA POL II systems of higher eukaryotes, excluding fungi, algae and protists. However, since promoters are viewed as physiological elements dependent on the correct interpretation by a *trans*-acting environment, many viral promoters are included and classified with their host species.

A strict non-redundancy policy is applied based on the principle that one entry should correspond to one biological entity. Data from different literature sources pertaining to the same transcription initiation sites are thus always combined in a single entry. Orthologous promoters from different species are linked by internal cross-references, as are alternative promoters of the same gene. 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 sequence entries are never blindly relied upon.

A more detailed description of the contents and format of EPD has been published in last year's database issue (4).

RECENT DEVELOPMENTS

Cross-references to other databases

In order to facilitate the development of integrated tools within the TRADAT consortium, a concentrated effort has been made to add many new cross-references to related data collections (Table 1). This value-addition was made possible by the major format revision accomplished last year. Breaking with the former tradition that one promoter entry refers to a single EMBL sequence, an initiative was started to systematically cross-reference all EPD entries to all corresponding EMBL entries. The decision to change the former policy was taken when it was realized that many potential links between EPD and TRANSFAC were missed because the two databases referred to different EMBL entries containing the same promoter sequence. With one database comprehensively linked to EMBL, such omissions should no longer occur.

*To whom correspondence should be addressed. Tel: +41 21 692 5892; Fax: +41 21 652 6933; Email: philipp.bucher@isrec.unil.ch

Table 1. Database cross-references in EPD

Database	Number of links
EPD internal	162
EMBL (1)	1849
TRANSFAC (2)	1157
SWISS-PROT (7)	929
FLYBASE (8)	127
MIM (9)	222
MGD (10)	44
MEDLINE	2126

Improvement of the access procedures

The WWW-based user interfaces have been improved in several ways. The sequence download page, which, for instance, can be used for retrieving training sets for promoter prediction algorithms, is shown in Figure 1. Note that the sequence range relative to the transcription initiation site is totally flexible because the sequences are extracted from the corresponding EMBL entries on the fly. The revised EPD query form (Fig. 2) now supports string searches with wildcard characters. The following options have been implemented for field restricted searches: All Text, ID, AccNumber, Description, Organism, Authors, Title, Citation, Homology group number, FLYBASE number, MIM number, TRANSFAC ID, SWISS-PROT ID, EMBL ID. Furthermore, the query results can newly be used for retrieving promoter sequences using the same mechanism as the sequence download page.

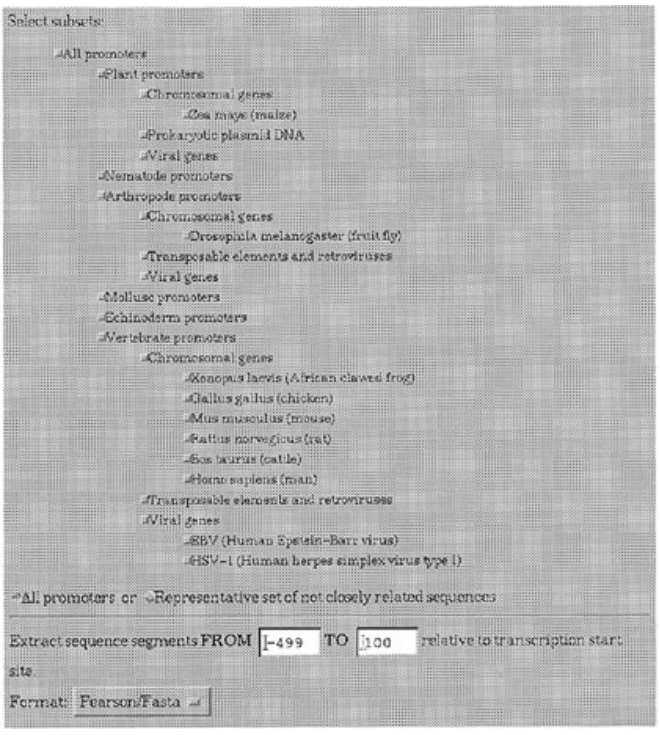

Figure 1. EPD sequence download form. The user can select promoter subsets from various species and higher order taxonomic groups. If the 'Representative set' option at the bottom is actived, a maximal set of representative sequences including no pair with more than 50% sequence identity will be retrieved.

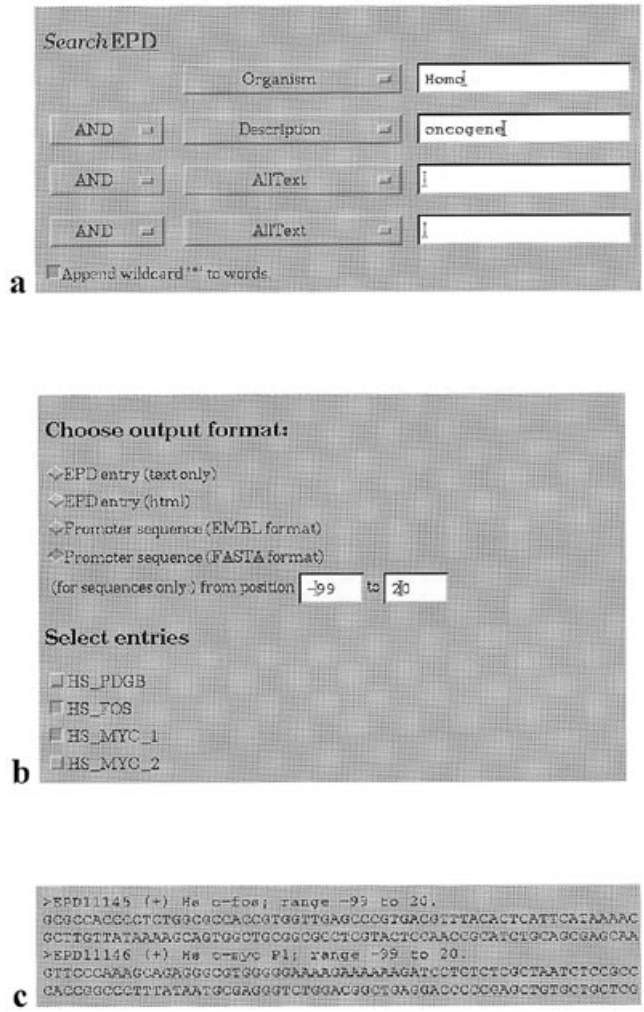

Figure 2. EPD query form and entry access procedures. (**a**) Query form. The design was inspired by the SRS query form at EBI. In the example shown, the user tries to find EPD promoter entries corresponding to human oncogenes. (**b**) Query results. The lower part shows the beginning of the list of entries satisfying the specified criteria, from which the user can select those he is interested in. The upper part offers several alternative output options. The selected entries can either be viewed by the browser, or corresponding sequences can be downloaded to a local file. Here, the user attempts to download promoter sequences from the human c-fos and c-myc oncogenes extending from positions –99 to 20 relative to the transcription start site. (**c**) Contents of the sequence file downloaded by the previous operation. Note that the sequences are in FASTA format as requested by the user.

SRS support for EPD

The existing Icarus configuration files used by SRS version 5 (5) were extensively modified in order to exploit the features of the new EPD format introduced last year. The new versions of these scripts are available from the ftp address given below (files epd.i, epd.is and epd.it). The SRS indexing system takes advantage of many of the recently introduced new fields, especially the cross-references to other databases.

ACCESS

EPD is distributed and maintained as a single ASCII flat file which can be obtained via anonymous ftp from ftp.epd.isb-sib.ch/pub/databases/epd. The following additional files are available:

(i) Sequence containing views in EMBL and FASTA format. These files contain promoter sequences in a range from –499 to +100 relative to the transcription start site plus excerpts from the promoter annotation.

(ii) A slightly reduced version of EPD in ASN.1 format designed for import into the GenBank–Entrez data environment (6).

(iii) Documentation files including the EPD user manual and a formal data description of the ASN.1 version.

(iv) Icarus scripts for indexing EPD by SRS.

The URL for online access to EPD is: http://www.epd.isb-sib.ch . This site offers the following services:

(i) Access to EPD entries by ID or accession number. The following formats are offered: text only, HTML, and HTML combined with a graphic representation of sequence objects by a Java applet (Junier & Bucher 1998, http://www.bioinfo.de/isb/1998/01/0003/).

(ii) A page for downloading promoter sequence subsets defined in EPD (for instance all human promoters from 100 bases upstream to 100 bases downstream of the initiation site).

(iii) Access to EPD entries or corresponding promoter sequences via a query form allowing for field-restricted character string searches.

SRS access to EPD is available at the Swiss EMBNet node: http://www.ch.embnet.org/

ACKNOWLEDGEMENT

EPD is funded in part by grant 95.0236-1 from the Swiss Federal Office for Education and Research.

REFERENCES

1 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk P. (1998) *Nucleic Acids Res.*, **26**, 8–15.
2 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.
3 Fickett,J.W. and Hatzigeorgiou,A.G. (1997) *Genome Res.*, **7**, 861–878.
4 Cavin-Périer,R., Junier,T. and Bucher,P. (1998) *Nucleic Acids Res.*, **26**, 353–357.
5 Etzold,T., Ulyanov,A. and Argos,P. (1996) *Methods Enzymol.*, **266**, 114–128.
6 Benson,D.A., Boguski,M., Lipman,D.J. and Ostell,J. (1994) *Nucleic Acids Res.*, **22**, 3441–3444.
7 Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.*, **25**, 31–36.
8 Gelbart,W.M., Crosby,M., Matthews,B., Rindone,W.P., Chillemi,J., Russo Twombly,S., Emmert,D., Ashburner,M., Drysdale,R.A., Whitfield,E., Millburn,G.H., de Grey,A., Kaufman,T., Matthews,K., Gilbert,D., Strelets,V. and Tolstoshev,C. (1997) *Nucleic Acids Res.*, **25**, 63–66.
9 Pearson,P., Francomano,C., Foster,P., Bocchini,C., Li,P. and McKusick,V. (1994) *Nucleic Acids Res.*, **22**, 3470–3473.
10 Blake,J.A., Richardson,J.E., Davisson,M.A. T, Eppig,J.T. and the Mouse Genome Informatics Group (1997) *Nucleic Acids Res.*, **25**, 85–91.