D82–D85 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh122

The Eukaryotic Promoter Database EPD: the impact of *in silico* primer extension

Christoph D. Schmid[1], Viviane Praz[1,2], Mauro Delorenzi[1,2,3], Rouaïda Périer[1] and Philipp Bucher[1,2,*]

[1]Swiss Institute of Bioinformatics, [2]Swiss Institute for Experimental Cancer Research and [3]NCCR Molecular Oncology, Ch. des Boveresses 155, 1066 Epalinges s/Lausanne, Switzerland

Received September 25, 2003; Revised and Accepted October 20, 2003

ABSTRACT

The Eukaryotic Promoter Database (EPD) is an annotated non-redundant collection of eukaryotic POL II promoters, experimentally defined by a transcription start site (TSS). There may be multiple promoter entries for a single gene. The underlying experimental evidence comes from journal articles and, starting from release 73, from 5′ ESTs of full-length cDNA clones used for so-called *in silico* primer extension. Access to promoter sequences is provided by pointers to TSS positions in nucleotide sequence entries. The annotation part of an EPD entry includes a description of the type and source of the initiation site mapping data, links to other biological databases and bibliographic references. EPD is structured in a way that facilitates dynamic extraction of biologically meaningful promoter subsets for comparative sequence analysis. Web-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. Tools for analysing sequence motifs around TSSs defined in EPD are provided by the signal search analysis server. EPD can be accessed at http://www.epd.isb-sib.ch.

OVERVIEW

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 (1,2), as well as in the development of eukaryotic promoter prediction algorithms (3). 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 (4).

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 earlier (5). User interfaces and software support for local installations were described in (6,7). Information on the regulation of promoters is provided through cross-references to CleanEx (previously named EPDEX), a database that maps promoters via genes to public expression profile (8).

RECENT DEVELOPMENTS

New entries based on *in silico* primer extension

Up to release 72 (dated October 2002) EPD was a manually compiled database, relying exclusively on experimental evidence published in scientific journals. With release 73, we started to exploit 5′ ESTs from full-length cDNA clones as a new resource for defining promoters. These data are automatically processed by computer programs and have rapidly revolutionized the way EPD is produced. Already a year after the introduction of this new method, more than half of the EPD entries (1634) are based on 5′ EST sequences.

We call this new technique of transcription start site (TSS) mapping '*in silico* primer extension'. The principle is the same as for conventional primer extension. In both cases, one attempts to synthesize cDNA molecules that extend to the 5′end of a transcript with the aid of a primer that hybridizes to an internal part of the mRNA sequence. However, there are two important differences. *In silico* primer extension uses 5′end sequences from cloned cDNAs generated for an entire mRNA population of a cell using a non-specific primer [usually oligo(dT), which hybridizes to the 3′end of the transcripts]. Conventional primer extension is carried out for one gene at a time with a gene-specific primer that hybridizes to a complementary sequence region near the 5′end of the mRNA. With the latter technique, the expected cDNA products are short, and thus likely to extend the 5′end of the target mRNA. Conversely, with poly(dT) as primer, the full-length cDNA products are expected to be long. Therefore specific cloning techniques have to be applied in order to

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

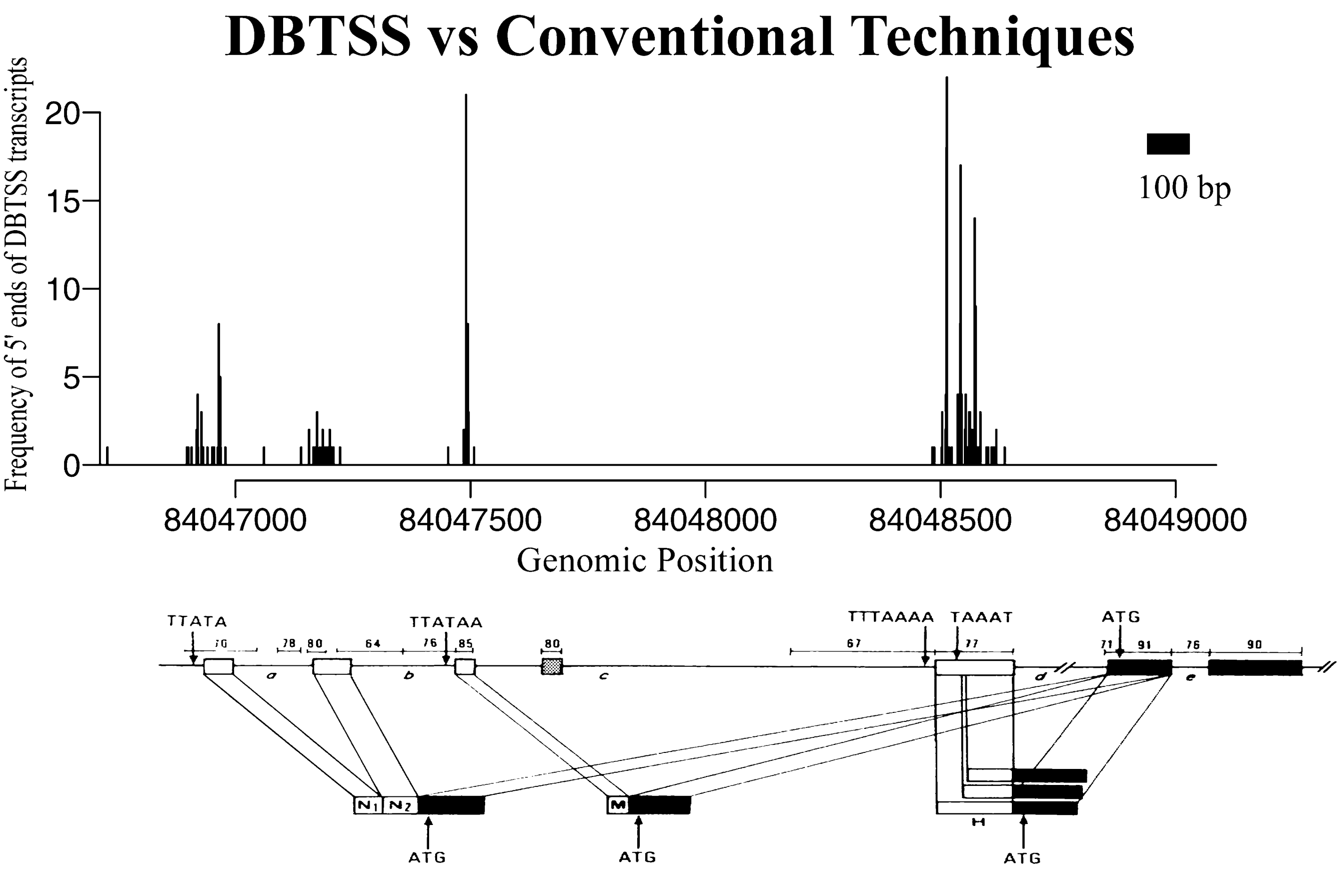

Figure 1. *In silico* primer extension yields results comparable to those of conventional methods. The upper panel displays the frequency distribution of the 5′ends of transcripts of the human aldolase A gene as derived from DBTSS (9). The figure in the lower panel [reprinted from (11) with permission from Elsevier] summarizes the results of mRNA 5′end mapping experiments carried out by conventional techniques for the same gene. Note that *in silico* primer extension successfully identified all four promoter regions reported before.

enrich for cDNAs that extend to the mRNA 5′end. The oligo-capping method, pioneered by the DBTSS team (9) has proved to be very effective to this end.

The second difference concerns the way cDNA extension products are analysed. In classical primer extension, the length of these products is determined by gel electrophoresis. With *in silico* primer extension, the cDNAs are analysed by cloning and sequencing. The 5′end sequences are then mapped *in silico* to the corresponding genome sequence with programs such as Blast or Sim4. We currently use procedures developed for the trome database (10) for this purpose. This mapping leads to a so-called cDNA 5′end profile, a digital structure that essentially contains the same information as the picture of a lane of a polyacrylamide gel documenting the length distribution of cDNAs obtained by conventional primer extension. It records how many times the 5′end of a cDNA clone from a particular gene has been found at each base position within a genome region of ~2 kb.

An automatic procedure has been developed for the identification of clusters within a cDNA 5′end profile that are likely to represent true TSSs. This procedure relies on a new clustering algorithm implemented in a program called madap (ftp://ftp.isrec.isb-sib.ch/pub/software/unix/madap), which attempts to fit a cDNA 5′end profile to a mixture of Gaussian distributions using an Expectation-Maximization (EM) algorithm. The EM algorithm can be forced to respect some user-defined constraints, in our case that: (i) a cluster must contain at least 10 cDNA 5′ends, (ii) it must comprise at least 10% of all 5′ends recorded in the profile, and (iii) it must obey a minimal centre-to-centre distance of 50 bp to its nearest neighbour. These criteria correspond roughly to the guidelines established earlier for the interpretation of data published in journal articles (5). Note that our way of exploiting 5′ends of cDNA differs in two important ways from the approach taken by DBTSS. First we allow for multiple promoters for the same gene. Second, we take as reference position for an EPD promoter entry the most frequently observed rather than the most upstream-located cDNA 5′end. An example of a cDNA 5′end profile for a gene having multiple promoters is shown in Figure 1 along with a gene model derived from conventional transcript mapping data.

So far, we have used the following data sources for *in silico* primer extension: (i) DBTSS (http://dbtss.hgc.jp/index.html), providing human full-length cDNA sequences from libraries constructed with the oligo-capping method, (ii) additional sequences of high-quality human cDNAs from the MGC project (12), and (iii) 5′ EST sequences of two *Drosophila* clone libraries of the Berkley *Drosophila* Genome project constructed with the oligo-capping method (13). Although not generated with the oligo-capping method, we accepted the

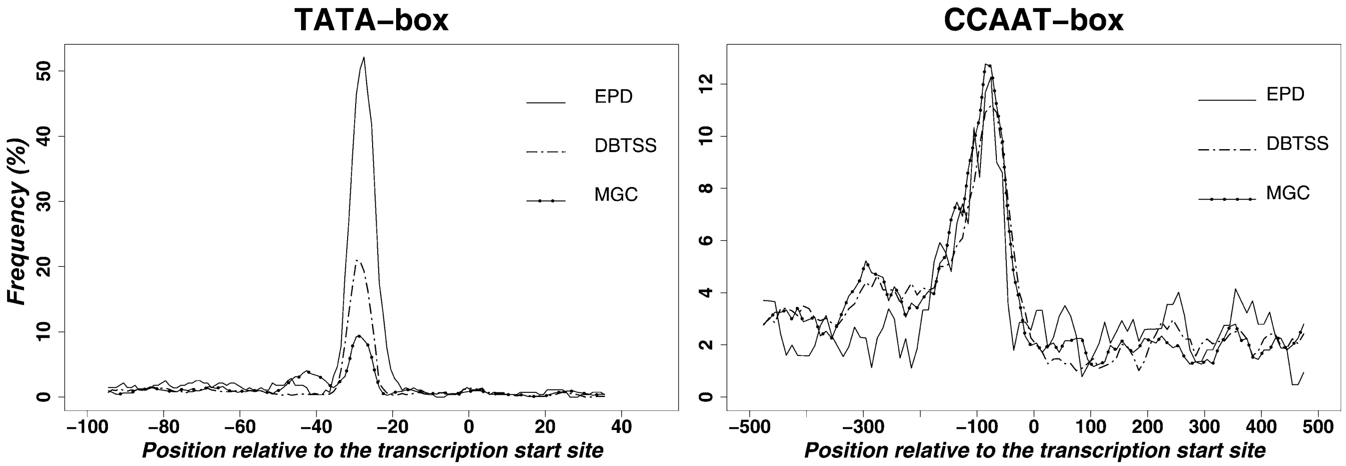

Figure 2. TATA and CCAAT box occurrence profiles for three classes of promoter entry. The DBTSS and MGC subsets were derived by *in silico* primer extension. The definitions of the sequence motifs were taken from (2). The TATA and CCAAT box signals were searched for in sliding windows of 20 and 50 bp, respectively. Theses profiles have been produced with the Signal search analysis server (14).

5′ ESTs from the MGC project because rigorous quality checks (see below) indicated that they are highly enriched in full-length sequences. The data from the two sources were nevertheless processed separately for reasons of transparency. In the processing of the MGC data, we started from the chromatograms (available at http://mgc.nci.nih.gov) as we noticed that the sequences deposited in EMBL often start several bases downstream of the true 5′end of the cDNA insert (which can be precisely identified in the chromatograms).

The newly generated promoter entries resulting from *in silico* primer extension were subjected to extensive quality controls before they were accepted for EPD. We first looked at the new TSS assignments of those promoters that were already in EPD. With very few exceptions, the TSS positions derived with the new and old methods were the same within experimental error. As a second test, we analysed the occurrence profiles of known promoter signals around the TSS. A previous study based on EPD led to the conclusion that ~70% of human promoters contain a TATA box located ~27 bp upstream of the TSS (2). Another promoter element, the CCAAT box, was found to be over-represented in a large upstream region of ~200 bases, with a peak frequency at –80. We analysed the positional distributions of these two signals around TSSs in three promoter subsets: old entries, new entries based on oligo-capped cDNA sequences from DBTSS and new entries based on MGC ESTs. If we assume that the promoter entries defined by the three different methods all correspond to true promoters, and that the TSS is determined with the same precision, then we would expect to see exactly the same positional distributions for the three subsets. This is indeed the case for the CCAAT box (see Fig. 2). The picture obtained for the TATA box is slightly different. Whereas the location and shape of the three peaks are largely identical, the heights are unequal. We explain this by the probable fact that the promoter subsets based on *in silico* primer extension are enriched in a subclass of TATA-less, CpG-island-associated promoters typical of abundant and ubiquitously expressed genes. The condition that a TSS must be documented by at least 10 cDNA 5′ends (see above) excludes weakly expressed genes with a narrow tissue distribution. Overall, we take the signal occurrence profiles shown in Figure 2 as proof that *in silico* primer extension is equally reliable and precise in identifying true transcription start sites as conventional methods.

Other developments

We have recently started to revise the naming of genes for several model organisms in order to conform to internationally approved gene nomenclatures. Thus all names for human genes are now based on Genew (15). We also started to provide TSS position references to genome contig sequences from RefSeq (16) in addition to EMBL sequence pointers.

Promoter sequences from EPD and several subsets of it can now be analysed at the Signal Search Analysis server (14) (http://www.isrec.isb-sib.ch/ssa/), which offers various types of algorithms to identify, localize and characterize promoter elements. For instance, Figure 2 has been produced with the OProf (Occurrence Profile) function of this server.

ACCESS

The FTP site (ftp://ftp.epd.isb-sib.ch/pub/databases/epd) provides the EPD database in flat file format, the user manual, promoter sequences in EMBL and FASTA format (from –499 to +100 relative to the TSS), an ASN.1 version of EPD designed for import into the GenBank-Entrez data environment (17) and Icarus scripts for indexing EPD by the Sequence Retrieval System (SRS) (18).

The website (http://www.epd.isb-sib.ch) offers the following services: Access to EPD entries through text-based query interfaces, display of EPD entries in text, HTML and the graphical SEView (19) formats, and a page for downloading promoter sequences of any length and location relative to the TSS. Blast and SRS access to EPD is offered by the Swiss EMBnet server (20).

ACKNOWLEDGEMENTS

This work was supported by the National Center of Competence in Research (NCCR) Molecular Oncology, a research instrument of the Swiss National Science Foundation. EPD is funded by grants from the Swiss government and the Swiss National Science Foundation (31-063933).

REFERENCES

1. Bucher,P. and Trifonov,E.N. (1986) Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res.*, **22**, 10009–10026.
2. 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.
3. Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.
4. 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.
5. Cavin Périer,R., Junier,T. and Bucher,P. (1998) The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.*, **26**, 353–357.
6. Cavin Périer,R., Junier,T., Bonnard,C. and Bucher,P. (1999) The Eukaryotic Promoter Database (EPD): recent developments. *Nucleic Acids Res.*, **27**, 307–309.
7. Cavin Périer,R., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The eukaryotic promoter database (EPD). *Nucleic Acids Res.*, **28**, 302–303.
8. Praz,V., Jagannathan,V. and Bucher,P. (2004) CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature. *Nucleic Acids Res.*, **32**, D542–D547.
9. Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002) DBTSS: database of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331
10. Sperisen,P., Iseli,C., Pagni,M., Stevenson,B.J., Bucher,P. and Jongeneel,C.V. (2004) trome, trEST and trGEN: databases of predicted protein sequences. *Nucleic Acids Res.*, **32**, D509–D511.
11. Maire,P., Gautron,S., Hakim,V., Gregori,C., Mennecier,F. and Kahn,A. (1987) Characterization of three optional promoters in the 5′ region of the human aldolase A gene. *J. Mol. Biol.*, **197**, 425–438.
12. Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G., Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D., Altschul,S.F. *et al.* Mammalian Gene Collection Program Team (2002) Generation and initial analysis of more than 15 000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
13. Stapleton,M., Liao,G.C., Brokstein,P., Hong,L., Carninci,P., Shiraki,T., Hayashizaki,Y., Champe,M., Pacleb,J., Wan,K. *et al.* (2002) The *Drosophila* gene collection: Identification of putative full-length cDNAs for 70% of *D. melanogaster* genes. *Genome Res.*, **12**, 1294–1300.
14. Ambrosini,G., Praz,V., Jagannathan,V. and Bucher,P. (2003) Signal search analysis server. *Nucleic Acids Res.*, **31**, 3618–3620.
15. Wain,H.M., Lush,M., Ducluzeau,F. and Povey,S. (2002). Genew: the human gene nomenclature database. *Nucleic Acids Res.*, **30**, 169–171.
16. Pruitt,K., Tatusov,T. and Manglott,D. (2003) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **31**, 34–37.
17. Benson,D.A., Boguski,M., Lipman,D.J. and Ostell,J. (1994) GenBank. *Nucleic Acids Res.*, **22**, 3441–3444.
18. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
19. Junier,T. and Bucher,P. (1998) SEView: a Java applet for browsing molecular sequence data. *In Silico Biol.*, **1**, 13–20.
20. Falquet,L., Bordoli,L., Ioannidis,V., Pagni,M. and Jongeneel,C.V. (2003) Swiss EMBnet node web server. *Nucleic Acids Res.*, **31**, 3782–3783.