EPD in its twentieth year: towards complete promoter coverage of selected model organisms

Christoph D. Schmid[1], Rouaïda Perier[1], Viviane Praz[1] and Philipp Bucher[1,2,*]

[1]Swiss Institute of Bioinformatics and [2]Swiss Institute for Experimental Cancer Research, Chemin des Boveresses 155, CH-1066 Epalinges, Switzerland

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). Access to promoter sequences is provided by pointers to positions in the corresponding genomes. Promoter evidence comes from conventional TSS mapping experiments for individual genes, or, starting from release 73, from mass genome annotation projects. Subsets of promoter sequences with customized 5′ and 3′ extensions can be downloaded from the EPD website. The focus of current development efforts is to reach complete promoter coverage for important model organisms as soon as possible. To speed up this process, a new class of preliminary promoter entries has been introduced as of release 83, which requires less stringent admission criteria. As part of a continuous integration process, new web-based interfaces have been developed, which allow joint analysis of promoter sequences with other bioinformatics resources developed by our group, in particular programs offered by the Signal Search Analysis Server, and gene expression data stored in the CleanEx database. EPD can be accessed at http://www.epd.isb-sib.ch.

HISTORICAL BACKGROUND

The Eukaryotic Promoter Database (EPD) originates from a promoter compilation published in this journal 20 years ago (1). Two years later, this collection became available in machine-readable form as an accessory database of the EMBL nucleotide sequence data library. Since then, EPD has undergone many changes, but its primary objective has remained the same: to provide access to experimentally mapped eukaryotic promoter sequences and to keep track of transcription start site (TSS) mapping data.

Not only our database has evolved over the last 20 years, but also the biologist's view of promoters, the experimental protocols to map TSSs, and the biological data environment have changed over this period. When we started to compile promoter sequences, commonly held views were that (i) each gene has one promoter, (ii) transcription always initiates at the same nucleotide and (iii) there is one sequence motif, the TATA-box, common to all promoters recognized by the eukaryotic polymerase II system. None of these assumptions have turned out to be true. Today we know that many human genes are transcribed from multiple promoters, not necessarily close to each other on the genome, and often giving rise to alternative first exons. Moreover, transcription initiation mechanisms appear to be less precise than initially assumed. In the human genome, it is not uncommon that the 5′ ends of mRNAs transcribed from the same promoter region are spread over >50 bp (2). Finally, promoters turned out to be heterogeneous with regard to sequence motif content. According to recent surveys, the once considered universal TATA-box element occurs only in about a third of all promoters in the systematically analyzed genomes of human (3), *Drosophila melanogaster* (4) and *Arabidopsis thaliana* (5).

The experimental procedures for mapping promoters, as well as the way EPD entries are constructed from public data, have undergone drastic changes at the beginning of the functional genomics era. Before, promoters were mapped for one gene at a time by techniques such as nuclease protection assay and primer extension analysis. The corresponding EPD entries were the result of a critical examination and independent interpretation of data published in paper-based journal articles. Today, TSSs are mapped at once for a whole genome with high-throughput technologies such as 5′ SAGE (6) or CAGE (7). The resulting data are disseminated in machine-readable form over the internet. As a consequence, EPD entries are now largely generated by intelligent Perl scripts with built-in quality control procedures rather than by critical readers of scientific articles. An overview of

Table 1. Summary of currently accessible mass genome annotation data for promoter mapping

5′ EST sequences from oligo-capped cDNA libraries			
Human	http://dbtss.hgc.jp/	400 225	Suzuki *et al.* (11)
Mouse	http://dbtss.hgc.jp/	580 209	Suzuki *et al.* (11)
Drosophila	Sequences available from Genbank/EMBL, accession numbers extractable from Unigene (23), Unilib IDs 23941 or 23942	102 617	Stapleton *et al.* (27)
Arabidopsis	ftp://pfgweb.gsc.riken.jp/rafl/	92 654	Seki *et al.* (28)
5′ sequences tags (5′SAGE, CAGE, GIS ditag)			
Human	http://5sage.gi.k.u-tokyo.ac.jp/	22 546	Hashimoto *et al.* (6)
Human	http://fantom31p.gsc.riken.jp/cage/download/hg17/	5 992 395	Carninci *et al.* (7)
Mouse	http://fantom31p.gsc.riken.jp/cage/download/mm5/	11 567 973	Carninci *et al.* (7)
Mouse	ftp://fantom.gsc.riken.jp/FANTOM3/GIS/	225 914	Ng *et al.* (29)
Reference sequence collections from oligo-capped cDNA libraries			
Rice	ftp://cdna01.dna.affrc.go.jp/pub/data/CURRENT	30 598	Kikuchi *et al.* (22)

The third column indicates the number of available sequences or tags.

publicly available mass genome annotation data useful for promoter mapping is given in Table 1.

Undoubtedly, the sequence data environment has undergone the most spectacular revolution during EPD's life span. When we started to compile promoters, sequences were available for only a few hundred short pieces of the human genome, most of them barely exceeding a thousand base pairs in length. Today, we have access to several complete genomes of higher eukaryotes totaling billions of nucleotides.

Despite these changes, the conceptual organization and data representation of EPD has remained remarkably stable. As a matter of fact, we anticipated many of the forthcoming changes in the initial design of EPD. For instance, we distinguished from the very beginning three classes of promoters, characterized by (i) single initiation sites, (ii) clustered multiple initiation sites and (iii) transcription initiation regions. We also allowed for multiple promoters per gene, being aware of a few such examples known at that time. The decision to provide access to promoter sequences indirectly through machine-readable pointers to sequences stored elsewhere turned out to be very helpful during the transition phase from the old-style nucleotide sequence database to the whole genome environment.

EPD is not anymore the only public database maintained by our group. The gene expression database CleanEx (8) and the Signal Search Analysis (SSA) server (9) are complementary bioinformatics resources developed in close coordination by partly overlapping teams. Note that CleanEx originated from a companion database of EPD called EPDEX (10), which by now has become largely obsolete. Whereas the source file distributions of the three products via ftp will remain self-contained and stand-alone, efforts are underway to integrate the corresponding web access tools into a tightly interconnected system for gene regulatory sequence analysis.

EPD is also not anymore the only database providing information about experimentally mapped TSSs. DBTSS (11) and PromoSer (12) are comprehensive collections of mammalian promoters based on clustering of expressed sequence tag (EST) and full-length cDNA sequences. These resources define the TSS as the furthest 5′ position in the genome which can be aligned with the 5′ end of a cDNA from the corresponding gene. In contrast, EPD considers the most frequent cDNA 5′ end as the TSS and further applies a specialized algorithm to infer multiple promoters for a given gene. Arguments and results in favor of our approach were presented in a previous article (13). PlantProm (14) is a smaller

volume database of plant promoters based on published TSS mapping data. HemoPDB (15) is a more specialized resource for promoters of genes of the hematopoietic system, providing information on transcription factor binding sites in addition to TSS annotation. OMGProm (16), DoOP (17) and CORG (18) are databases of orthologous promoters with a comparative genomics focus.

A detailed description of the contents and format of EPD was given in Ref. (19). Information about interfaces and support for local installations can be found in Ref. (20,21). New format features for promoter entries derived from mass genome annotation data are described in Ref. (10). The *in silico* primer extension protocol used for generating promoter entries from mass genome annotation data is detailed in Ref. (13).

TOWARDS COMPLETE COVERAGE FOR MODEL ORGANISMS

In the past, the maintenance policy of EPD was to guarantee high-quality standards. In order to be included, a promoter had to satisfy stringent criteria regarding its experimental characterization (13). Undoubtedly, the user community of EPD (mostly computational biologists) has appreciated this focus on quality rather than quantity in the past. The backside nevertheless is that promoter coverage of important model genomes has remained modest.

Today, the demand is slowly changing. As a result of the Human Genome Sequencing Project, the so-called global approach to organisms has become fashionable. Intensive efforts are currently made to functionally annotate the complete genomes of various model organisms by experimental as well as computational methods. In response to these trends, we redefined the priorities of our development efforts. Our stated objective is now to reach complete promoter coverage for three model organisms (human, *D.melanogaster* and rice) as soon as possible.

To conciliate the contrasting objectives of high quality and quantity, we introduced a new class of promoter entries called 'preliminary'. Such entries fulfill less stringent admission criteria and are generated automatically from mass genome annotation data and other genome information resources. Some of these entries are based on external annotation efforts. There are several potential reasons why a preliminary entry may not be acceptable as a standard, high-quality entry: insufficient experimental data, missing information about

computational TSS inference procedures, format incompatibilities with third party annotations, or uncertainties about the identity of the corresponding genes. The latter happens, for instance, with scarcely annotated new genomes, such as the rice genome.

The inclusion of preliminary promoter entries was encouraged by the successful development of sequence motif-based tests to assess the quality of automatically generated promoter sets (13). These tests take into account the occurrence frequency and positional distribution of predicted promoter elements in the evaluated promoter set in order to estimate the amount of contaminating non-promoter sequences and the average error of TSS positions. Corresponding results obtained from a high-quality promoter set from the same organism are used for calibration. Since preliminary promoter entries are always generated in large numbers by the same procedure, statistically robust quality estimates can be obtained for groups of such entries, but obviously not for individual promoters.

The above outlined quality evaluation procedure is also used for choosing the acceptance threshold, and for the fine-tuning of certain parameters of the data processing pipeline for preliminary entries. To illustrate this principle, let's consider the *in silico* primer extension method for inferring TSSs. Our currently implemented procedure relies on the program madap (ftp://ftp.isrec.isb-sib.ch/pub/software/unix/madap) for identifying clusters of cDNA 5′ ends mapped to the genome. For standard EPD entries, we require at least 10 cDNAs per cluster. For preliminary entries, we could simply lower the threshold number. Sequence motif-based tests as described above suggest that a threshold as low as three cDNAs would still yield acceptable quality for preliminary promoter entries.

The first set of preliminary entries ready for inclusion in EPD happened to be a collection of 13 046 rice promoters, derived from a reference collection of ∼30 500 mRNA sequences published by the rice full-length cDNA Consortium (22). This reference collection was generated by clustering and genome mapping of ∼170 000 initial cDNA sequences, all from libraries generated with the oligo-capping method. There were several reasons which prevented us from making standard EPD entries from this external genome annotation resource: (i) we had no access to the primary data, (ii) in the general case, we had to rely on one full-length cDNA sequence per gene and consequently were unable to assign the promoter to one of the three TSS classes, single, multiple or region and (iii) the preliminary annotation status of the rice genome made it impossible for many promoters to provide meaningful gene descriptions.

In our local data processing pipeline, we first subjected the rice mRNA sequence collection to additional quality control steps. Sequences whose 5′ terminal 11 nt did not match the rice genome with at most one mismatch were discarded. By checking the assignment of the corresponding GenBank/EMBL accession numbers to Unigene clusters (23) we were able to eliminate hitherto undiscovered redundancy in the original collection. For the remaining entries, the rudimentary gene annotation provided by the consortium was complemented with information from the genome annotation project at TIGR [release 3.0: December 30, 2004 (24)].

Application of the sequence motif-based evaluation procedure to this preliminary promoter set indicated that the TSS assignment was of similar precisions as in standard EPD entries. However, we cannot exclude for the moment the possibility that the collection is contaminated with a sizable
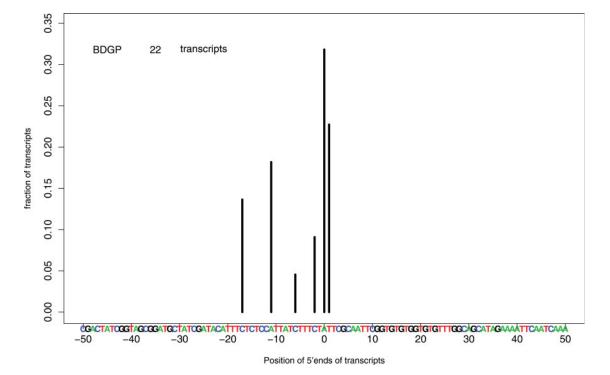

Figure 1. Graphical representation of the distribution of 5′ ends of full-length transcripts. The diagram is based on data from the Berkley *Drosophila* Genome Project for gene ARF79F and is part of the 'niceview' display of EPD entry DM_ARF1_2 (http://www.epd.isb-sib.ch/cgi-bin/get_doc?db=epd&format=nice&entry=DM_ARF1_2).

fraction of non-promoter sites. In an additional test, we tried to compare the newly generated entries with already existing EPD entries for the same promoters. We found only seven examples suitable for this purpose. Of those, five preliminary entries matched their high-quality homologs with TSS position shifts of −4, −2, +2, +2 and +25 bp.

Additional preliminary promoter sets are in preparation. Most of them will be based on *in silico* primer extension protocols with relaxed constraints, as described above.

Preliminary EPD entries are available in a separate file named epd_bulk.dat from our FTP server. The web-based pages provide access to both standard and preliminary entries. Note that preliminary entries are identified by the keyword 'preliminary' on the ID line.

OTHER RECENT DEVELOPMENTS

In response to numerous requests, we included new Fasta-formatted promoter sequence library files with an extended range of −9999 to +6000 relative to TSS in the FTP release. The popular sequence download page, which can be used for retrieval of biologically meaningful promoter sequence subsets of user-defined extension, has a new feature allowing direct sequence transfer to the SSA server (9). The web-based entry viewers were equipped with genome position hyperlinks to the 'ENSEMBL ContigView' (25) and 'UCSC Genome Browser' (26). Moreover, a graphical representation of the initiation site patterns (Figure 1) was added to the 'niceview' display for those EPD entries which include a cDNA 5′ end profile derived by *in silico* primer extension.

ACKNOWLEDGEMENTS

Conflict of interest statement. None declared.

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. Suzuki,Y., Taira,H., Tsunoda,T., Mizushima-Sugano,J., Sese,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Morishita,S. *et al.* (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.*, **2**, 388–393.
3. 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.
4. Ohler,U., Liao,G.C., Niemann,H. and Rubin,G.M. (2002) Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.*, **3**, RESEARCH0087.
5. Molina,C. and Grotewold,E. (2005) Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics*, **6**, 25.
6. Hashimoto,S., Suzuki,Y., Kasai,Y., Morohoshi,K., Yamada,T., Sese,J., Morishita,S., Sugano,S. and Matsushima,K. (2004) 5′-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.*, **22**, 1146–1149.
7. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The

transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
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. Ambrosini,G., Praz,V., Jagannathan,V. and Bucher,P. (2003) Signal search analysis server. *Nucleic Acids Res.*, **31**, 3618–3620.
10. Praz,V., Périer,R.C., Bonnard,C. and Bucher,P. (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res.*, **30**, 322–324.
11. 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.
12. Halees,A.S., Leyfer,D. and Weng,Z. (2003) PromoSer: a large-scale mammalian promoter and transcription start site identification service. *Nucleic Acids Res.*, **31**, 3554–3559.
13. Schmid,C.D., Praz,V., Delorenzi,M., Perier,R. and Bucher,P. (2004) The Eukaryotic Promoter Database EPD: the impact of *in silico* primer extension. *Nucleic Acids Res.*, **32**, D82–D85.
14. Shahmuradov,I.A., Gammerman,A.J., Hancock,J.M., Bramley,P.M. and Solovyev,V.V. (2003) PlantProm: a database of plant promoter sequences. *Nucleic Acids Res.*, **31**, 114–117.
15. Pohar,T.T., Sun,H. and Davuluri,R.V. (2004) HemoPDB: Hematopoiesis Promoter Database, an information resource of transcriptional regulation in blood cell development. *Nucleic Acids Res.*, **32**, D86–D90.
16. Palaniswamy,S.K., Jin,V.X., Sun,H. and Davuluri,R.V. (2005) OMGProm: a database of orthologous mammalian gene promoters. *Bioinformatics*, **21**, 835–836.
17. Barta,E., Sebestyen,E., Palfy,T.B., Toth,G., Ortutay,C.P. and Patthy,L. (2005) DoOP: Databases of Orthologous Promoters, collections of clusters of upstream sequences from chordates and plants. *Nucleic Acids Res.*, **33**, D86–D90.
18. Dieterich,C., Wang,H., Rateitschak,K., Luz,H. and Vingron,M. (2003) CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Res.*, **31**, 55–57.
19. Cavin Périer,R., Junier,T. and Bucher,P. (1998) The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.*, **26**, 353–357.
20. Cavin Périer,R., Junier,T., Bonnard,C. and Bucher,P. (1999) The Eukaryotic Promoter Database (EPD): recent developments. *Nucleic Acids Res.*, **27**, 307–309.
21. Cavin Périer,R., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The eukaryotic promoter database (EPD). *Nucleic Acids Res.*, **28**, 302–303.
22. Kikuchi,S., Satoh,K., Nagata,T., Kawagashira,N., Doi,K., Kishimoto,N., Yazaki,J., Ishikawa,M., Yamada,H., Ooka,H. *et al.* (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*, **301**, 376–379.
23. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–45.
24. Yuan,Q., Ouyang,S., Liu,J., Suh,B., Cheung,F., Sultana,R., Lee,D., Quackenbush,J. and Buell,C.R. (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res.*, **31**, 229–233.
25. 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**, D447–453.
26. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
27. Stapleton,M., Liao,G., 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.
28. Seki,M., Narusaka,M., Kamiya,A., Ishida,J., Satou,M., Sakurai,T., Nakajima,M., Enju,A., Akiyama,K., Oono,Y. *et al.* (2002) Functional annotation of a full-length Arabidopsis cDNA collection. *Science*, **296**, 141–145.
29. Ng,P., Wei,C.L., Sung,W.K., Chiu,K.P., Lipovich,L., Ang,C.C., Gupta,S., Shahab,A., Ridwan,A., Wong,C.H. *et al.* (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods*, **2**, 105–111.