200–203 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

EpoDB: a prototype database for the analysis of genes expressed during vertebrate erythropoiesis

Christian J. Stoeckert, Jr*, Fidel Salas[1], Brian Brunk[2] and G. Christian Overton[2]

Division of Hematology, The Children's Hospital of Philadelphia, 316E Abramson Research Center, 34th and Civic Center Boulevard, Philadelphia, PA 19104, USA, [1]Pangea Systems, Inc., 1999 Harrison Street, Suite 1100, Oakland, CA 94612, USA and [2]Department of Genetics, University of Pennsylvania School of Medicine, 1312 Blockley Hall, 418 Guardian Drive, Philadelphia, PA 19104-6021, USA

Received October 2, 1998; Revised October 8, 1998; Accepted October 14, 1998

ABSTRACT

EpoDB is a database of genes expressed in vertebrate red blood cells. It is also a prototype for the creation of cell and tissue-specific databases from multiple external sources. The information in EpoDB obtained from GenBank, SWISS-PROT, Transfac, TRRD and GERD is curated to provide high quality data for sequence analysis aimed at understanding gene regulation during erythropoiesis. New protocols have been developed for data integration and updating entries. Using a BLAST-based algorithm, we have grouped GenBank entries representing the same gene together. This sequence similarity protocol was also used to identify new entries to be included in EpoDB. We have recently implemented our database in Sybase (relational tables) in addition to SICStus Prolog to provide us with greater flexibility in asking complex queries that utilize information from multiple sources. New additions to the public web site (http://www.cbil.upenn.edu/epodb) for accessing EpoDB are the ability to retrieve groups of entries representing different variants of the same gene and to retrieve gene expression data. The BLAST query has been enhanced by incorporating BLAST-View, an interactive and graphical display of BLAST results. We have also enhanced the queries for retrieving sequence from specified genes by the addition of MEME, a motif discovery tool, to the integrated analysis tools which include CLUSTALW and TESS.

INTRODUCTION

EpoDB is a database that seeks to capture all available information about genes expressed in vertebrate red blood cells from the first committed progenitor (BFU-E) to the end stage erythrocyte. The overall goal of the EpoDB project is to understand how blood cells develop by using the information in EpoDB to model gene regulation during red blood cell differentiation (erythropoiesis). The information required will encompass ubiquitously-expressed genes as well as those specifically associated with red blood cells. It also includes genes for cytokines and iron transport such as erythropoietin and transferrin which play critical roles in red blood cell development although not expressed in red blood cells. The types of information required for understanding how the patterns of genes expressed in red blood cells lead to their distinctive morphology and make-up include gene sequence, gene function, gene regulation and gene expression. In building EpoDB, we have extracted relevant information from GenBank (1), SWISS-PROT (2), Transfac and TRRD (3) and GERD (4) using a set of keywords to search for candidate entries and manually checked each one for appropriateness of inclusion in EpoDB (5).

The purpose of EpoDB is not just to provide information about red blood cell genes but to utilize that information for computational analyses (e.g., identifying common promoter elements). Thus, selection of entries to be included in EpoDB is only the first step toward making this information 'computer friendly.' Also required is that the information be complete and accurate and that entries to be analyzed can be unambiguously specified. A parser program, SSP, is used to check whether GenBank-derived features are consistent and to deduce missing features where possible (6). The concept 'transcription unit' has been applied to distinguish genes from the entries containing them because an entry may contain multiple genes or transcript units. Reference genes (transcription units) have been identified which are completely and accurately annotated. These genes are given names from controlled vocabularies and links provided to information from other data sources adding to the value of the entries in EpoDB. It allows us to study (for example) the developmental regulation of globin genes by extracting the promoter sequences for fetally-expressed γ-globin genes and comparing them to embryonically-expressed γ-globin genes. The ease with which this 'Gene Landmark' query can be performed is demonstrated on a tutorial page provided at the EpoDB web site and a screen shot collage illustrating the result is given in Figure 1.

EpoDB can also be viewed as a prototype system for the creation of cell or tissue-specific databases from multiple sources. In building EpoDB we are addressing issues of data collection, data integration, curation, updates and maintenance and data analysis. The goal is to develop a set of tools that can be applied to other biological systems of interest with a minimum of manual

*To whom correspondence should be addressed. Tel: +1 215 590 2139; Fax: +1 215 590 4834; Email: stoeckert@email.chop.edu

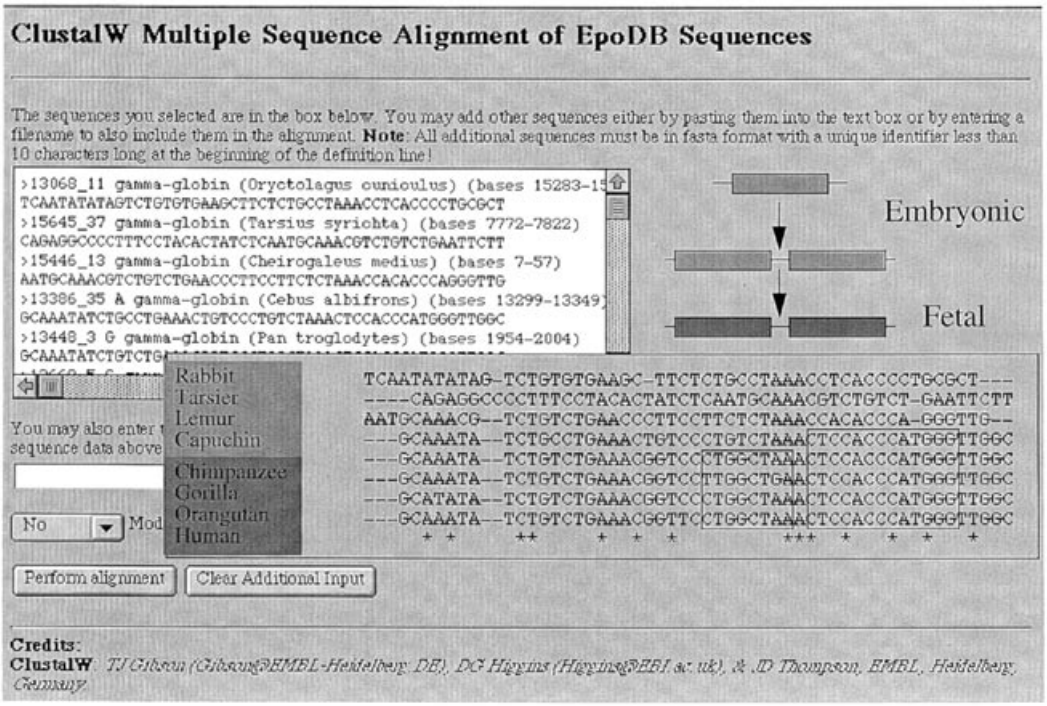

Figure 1. Comparative analysis of γ-globin gene promoters which are active during embryonic or fetal development. Sequences were extracted using the EpoDB Landmark query and aligned using CLUSTALW. Boxed are transcription factor binding sites (identified using TESS) which are selectively conserved in fetally-expressed genes.

involvement. In version 2.2 of EpoDB we have continued the process of collecting relevant information about red blood cells and made improvements to the way information in EpoDB can be acquired as well as accessed.

EpoDB DATA

Data from the multiple external databases are transformed and integrated into a single conceptual database. The data is checked for consistency of syntax and semantics and links are created between attributes of the same entry (and of the same gene). The information can be materialized as a flat-file, as a Prolog database, or as a Sybase relational database. Prolog is a declarative language and well suited for heuristic analyses. However, simple retrieval of data based on specified values can be more easily managed using a relational database server such as Sybase. We have made relational representations of the Prolog schema and have the current version of EpoDB stored in relational Sybase tables. Querying the data can be done using SQL and the Java database API JDBC (http://java.sun.com/products/jdbc/index.html) for transformation of the data into Java objects. It is our goal to use this technology to seamlessly integrate the data in EpoDB to the bioWidget visualization tools (http://www.cbil.upenn.edu/bioWidgets) which use Java objects. The bioWidgets are a reusable set of software components that display and manipulate commonly used biological objects such as sequence and gene maps (7).

The value added to GenBank entries in EpoDB consists of identifying entries expressed in vertebrate red blood cells, error-checking given annotation of features and deducing missing annotation, and applying controlled vocabularies. From 3715 GenBank entries in EpoDB (v2.1) we were able to identify 185 entries that contained a complete gene sequence and could be fully-annotated. The genes in these entries were termed reference genes or 'transcription units' and made available for sequence analysis. An individual non-reference entry may contain little useful information on its own but combined with other entries for the same gene may contribute to a (more) complete description of the gene. Furthermore, reference genes may not contain extended flanking sequence that is available in a non-reference entry that covers only the 5′ end of a gene. Clearly, it would be desirable to have that information somehow linked to the reference gene. To this end, we have clustered all entries referring to the same gene by sequence similarity using the BLASTN program (8). We have restricted the sequences used for this analysis to transcribed regions to prevent entries containing multiple genes (such as the human β-globin gene cluster) from linking all globin genes in the same group. A cut-off of 98% identity over a 50 bp region was used as the criteria for representing the same gene; entries also had to be from the same organism. This procedure resulted in the construction of 353 groups with two or more members representing 2108 of the 3715 entries used. For those groups that contained a manually-identified reference gene, that reference gene became the representative gene of the group. The controlled vocabulary name of the reference gene then can be assigned to all members of the group. In those groups where no reference gene exists, keywords were retrieved from each of the RNA entries of the group to create a group description; only RNA entries were used to avoid inclusion of keywords which describe extragenic features (e.g., alu repeats) or additional genes in an entry. Furthermore, controlled vocabulary gene names can be assigned to each group enabling efficient

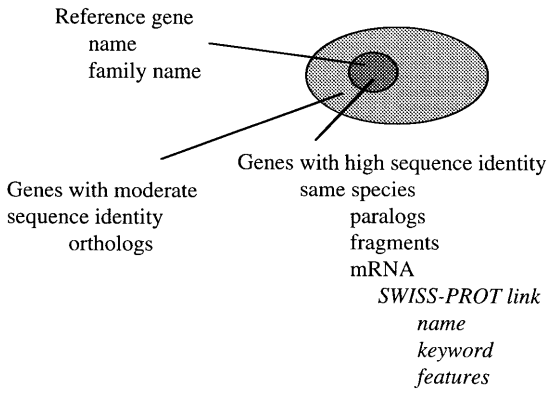

Figure 2. Data integration based on sequence similarity using BLASTN. The set of GenBank-derived entries representing the same gene (dark circle) or closely-related paralogs can be grouped together pooling the information associated with each entry such as a SWISS-PROT link. This pool of information can be accessed through a reference gene belonging to each group which has a name and family name drawn from a controlled vocabulary. The set of entries representing orthologs (light circle) or closely-related homologs can be obtained by relaxing the sequence similarity criteria.

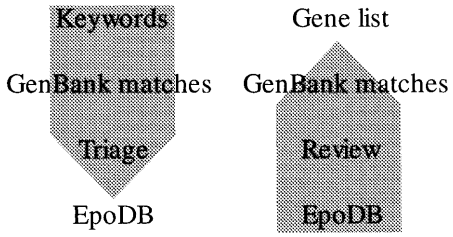

Figure 3. Strategies used to populate EpoDB. A list of keywords was used to search GenBank for candidates (left) that required manual triaging of many entries for appropriateness. Triaging could be avoided if gene sequences instead of keywords were used (right). Manual review is still necessary but on fewer entries than with keywords and the entries are grouped.

retrieval of information about a gene of interest. The net result is that more of the information in EpoDB is directly accessible for both browsing and sequence analysis. This approach can be extended to group genes into families of orthologs and paralogs as well. Figure 2 illustrates the relationship of gene groups and gene family groups.

The use of sequence similarity to group entries from the same gene together was used to identify new GenBank entries for inclusion into EpoDB. To update GenBank entries for version 2.2 of EpoDB, candidate entries were identified in GenBank release 107 as before using a keyword list which returned nearly 15 000 accessions not currently in EpoDB. To avoid the tedious process of manual triage of the 15 000 entries for true positives, false positives, and undecideables, a sequence similarity protocol was employed to identify those new entries which represented the same genes currently found in EpoDB and therefore should also be included in EpoDB. This automated approach does not lead to the inclusion of new genes (unless closely related). These can still be manually identified if it is desired to include a specific gene of interest. However, only one representation of a gene of interest need be included because the automated approach should then pick out other representations. This approach identified 1407 entries with a *P*-value of e-50 or smaller. Manual review is still necessary to ensure that the matches are between different versions of the same genes or their orthologs although the number of entries to review is greatly reduced and easy to assess. By adding the additional criteria of 98% identity, the number of entries was reduced to a set of 210 which were new representations of genes currently in EpoDB. These included instances where the new entries did not include descriptions of the matched gene and raises the possibility of gene discovery by this approach. These entries are grouped with current EpoDB entries and are accessible through them. Candidates for new reference genes in a group can be identified by looking at the size of entries, when last updated, and the number of features contained. In fact, creation of a new database may be efficiently achieved starting with a list of key genes identified by an expert rather than with a list of keywords as originally done for EpoDB. The gene sequences could then be used to search target databases to identify all representations of those genes based on sequence similarity. Figure 3 illustrates the different approaches used to originally create EpoDB and how new entries are to be identified (or new databases created) in the future.

QUERIES, VISUALIZATION AND ANALYSIS TOOLS

EpoDB is accessible at http://www.cbil.upenn.edu/epodb . Queries are available for browsing the database by text search or by specifying a gene name (or family name). Graphic views are provided of the entries using bioWidgets. New in version 2.2 is the ability to retrieve gene groups (based on sequence identity) using gene name or keywords and organism. Also new is the ability to access gene expression information for a specified gene drawn from data derived from the GERD database related to red blood cells.

Extraction of specified sequence (relative to a transcriptional landmark or feature) for analyses is provided at the web site. Integrated with these queries are the CLUSTALW program (9) for sequence alignment, the MEME program for identifying common motifs (10) and the TESS (Transcription Element Search System) site (http://www.cbil.upenn.edu/tess) which searches for potential transcription factor binding sites using the Transfac database. Specific sequence motifs can be directly searched individually or as a pair in a promoter sequence query which utilizes the GenLang pattern recognition system (11). The EpoDB database (or specified parts of it) can be searched for similarity to a submitted sequence using BLAST. A recent enhancement is the incorporation of the BLASTView widget that provides a graphic and interactive display of BLAST results.

SUMMARY

EpoDB provides the ability to not only browse for information about red blood cell genes but also to perform computational analyses of highly curated data obtained from external databases. Three major advances have taken place in the version 2.2 update. The first is the storage of EpoDB data in a relational database which facilitates access to all of the data**.** The second major advance is the use of a sequence homology protocol to group different entries representing the same gene. The third major advance is an update of GenBank-derived entries which utilizes the sequence similarity protocol to identify entries to be included in EpoDB.

Future developments will include the integration of gene expression data from highly parallel analyses to drive both the population and analysis of genes in EpoDB. A list of genes experimentally found to be expressed in cells of interest could be used to search target databases to acquire all relevant information. This information can then be used to provide a context and verification for the gene expression results. Genes with similar expression patterns could be subjected to computational analyses aimed at discovering common regulatory elements. These approaches will enhance EpoDB as a resource for understanding gene regulation during erythropoiesis and as a prototype for creating speciality biological databases.

ACKNOWLEDGEMENTS

We wish to thank Dr Jian Wang for her help with the recent update and Arthur Lin and Zhaoliang Lu for their help on the EpoDB web site. This work was supported by grant number R01-RR04026-08 from the National Center for Research Resources, NIH.

REFERENCES

1 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F. (1998) *Nucleic Acids Res.*, **26**, 1–7.
2 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
3 Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,A.E., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A., Podkolodny,N.L. and Kolchanov,N.A. (1998) *Nucleic Acids Res.*, **26**, 362–367.
4 Stoeckert,C., Podkolodnaya,O.A., Kel,A.E., Brunk,B., Haas,J., Salas,F., Stepanenko,I.L., Ignatieva,E.V., Kel-Margoulis,O.V., Ananko,E.A., Podkolodny,N.L., Overton,G.C. and Kolchanov,N.A. (1998) *Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'98)*, **1**, 20–24.
5 Salas,F., Haas,J., Brunk,B., Stoeckert,C.J.,Jr and Overton,G.C. (1998) *Nucleic Acids Res.*, **26**, 288–289.
6 Overton,G.C., Aaronson,J., Haas,J. and Adams,J. (1994) *J. Comput. Biol.*, **1**, 3–13.
7 Searls,D.B. (1995) *Gene*, **163**, GC 1–6.
8 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
9 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
10 Bailey,T.L.and Elkan,C. (1995) *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology (ISMB'95)*, 21–29.
11 Dong,S. and Searls,D.B. (1994) *Genomics*, **23**, 540–551.