# EpoDB: a database of genes expressed during vertebrate erythropoiesis

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

Department of Genetics, University of Pennsylvania School of Medicine, Room 475, Clinical Research Building, 422 Curie Boulevard, Philadelphia, PA 19104-6145, USA and [1]Division of Hematology, The Children's Hospital of Philadelphia, Abramson Center, 34th and Civic Center Boulevard, Philadelphia, PA, USA

Received October 14, 1997; Accepted October 15, 1997

## ABSTRACT

**EpoDB is a database designed for the study of gene regulation during differentiation and development of vertebrate red blood cells. In building EpoDB, we have taken the in advance approach to the data integration problem: we have extracted data relevant to red blood cells from GenBank, SWISS-PROT, TRRD (transcriptional regulation data) and GERD (expression levels data) to create a single integrated, highly curated view. Tools have been developed to automate data extraction from online resources, cleanse data of errors, enter information manually from the primary literature, generate a uniform, canonical representation of information and maintain data currency. The database is organized around biological features, e.g., genes, rather than sequences, which are supported by a controlled and consistent vocabulary for gene names and gene family names. Beyond the standard database queries, the functionality of EpoDB includes the ability to extract features and subsequences, display sequences and features graphically using bioWidget viewers and integrated analysis tools. EpoDB may be accessed at: http://cbil.humgen.upenn.edu/epodb/**

## INTRODUCTION

Elucidating the detailed mechanisms that regulate gene expression during development and differentiation remains one of the central challenges of biology. A vast literature detailing many of the structural components involved in gene regulation (transcription factors, transcription elements, signal transduction pathways, and so on) and to a lesser extent the detailed biological processes has accumulated over the last several decades. Complementing this foundational information, recent genome-scale projects in functional analysis are generating data relevant to understanding gene expression at an unprecedented rate. The existence of hundreds of databases and other information sources available today presents the biologist with the problem of how to integrate, analyze, and visualize the various sets of data of interest. Finally, much key information on gene function and gene expression remains or ends up in the primary literature only and is currently inaccessible for analysis. The volume and complexity of data mandate the development of a new generation of database systems capable of handling the temporal and spatial, as well as the structural aspects of gene expression. We have implemented a prototype system, EpoDB, which focuses on gene expression during erythropoiesis, the differentiation of red blood cells from the hematopoietic stem cell. With EpoDB, we will advance both the study of erythropoieis and issues regarding data management, analysis and visualization of spatio-temporal information.

In order to facilitate functional analysis of developmental pathways, we have taken the 'eager' approach to data integration (commonly called data warehousing). That is, we extract in advance data of interest from various sources, translate, cleanse, filter as appropriate, and finally store them in a central repository. We have chosen the pathway of differentiation that leads to the mature red cell as a system to study gene expression and as a testbed for a general approach to building data warehouses that integrate data for functional analysis. We chose erythropoiesis (1) because there is already a wealth of information on the regulation of genes (especially the globins) and their expression levels during erythropoiesis. Also, interspecies information can be used for functional and evolutionary analyses.

EpoDB (2) contains information on gene structure, function, regulation and expression levels during vertebrate erythropoiesis. It is available online as a reference source to help investigators design experiments and assist in developing models of gene regulation. In this paper, we describe the content of EpoDB, the value we have added to data, provide examples of the queries we provide over the World Wide Web (WWW), and illustrate the ways in which the results are displayed.

## EpoDB DATA

The EpoDB data management and analysis system, and companion tools for multiple database access and integration, are implemented in SICStus Prolog and Sybase. EpoDB contains information on housekeeping genes as well as genes expressed differentially during erythropoiesis. The content of EpoDB is derived by a semi-automated process: query the source database

*To whom correspondence should be addressed. Tel: +1 215 573 3105; Fax: +1 215 573 3111; Email: coverton@cbil.humgen.upenn.edu with a set of erythropoiesis relevant keywords to identify source entries; extract and transform the source entries to the EpoDB schema; triage the entries to remove unwanted entries; cleanse the data to correct errors. The source databases currently include GenBank (3), SWISS-PROT (4), TRRD (5) (Transcription Regulatory Region DB), and Medline. In addition, a separate database, GERD (Gene Expression and Regulation DB), which records information on gene expression extracted directly from the primary literature, has been developed. GERD is managed jointly by our group and our collaborators Drs N. Kolchanov, A. Kel and O. Kel at the Institute of Cytology and Genetics (CGI, Novosibirsk, Russia).

In the case of GenBank data, accession numbers of entries matched by the keywords are retrieved from the flat files. Using these accession numbers, entries in ASN.1 format are retrieved and converted to the EpoDB format. To perform cleansing of the data, computational augmentation of the feature table, and to create an uniform structure, the entries are passed through a parser/expert system (6). In the case of the SWISS-PROT, GERD and TRRD data, the matched entries are analyzed to add more structure and then transformed into the EpoDB schema. What few (syntactic) errors were detected in SWISS-PROT entries were quickly corrected by the SWISS-PROT team. Similarly, TRRD and GERD errors were corrected by the CGI team.

The numbers of current entries extracted from each source database are shown in Table 1. Each entry was manually checked to see if our criteria for inclusion in EpoDB was met. Entries were marked as ‘yes’ if they corresponded to genes known to be expressed in erythropoiesis; ‘no’ if they were known not to be expressed in erythropoiesis; and ‘maybe’ if it could not be unequivocally determined one way or the other.

**Table 1.** Entries in EpoDB by source

| | GenBank | SWISS-PROT | TRRD | GERD |
|---|---|---|---|---|
| total | 7819 | 2381 | 171 | 65 |
| yes | 3715 | 1241 | 80 | 65 |
| maybe | 1807 | 668 | 0 | 0 |
| no | 2297 | 472 | 91 | 0 |

In addition to the computational augmentation performed by the system, our immediate goal has been to add a consistent, standard convention for naming genes, gene products and gene families. This information extends the capabilities of the system so that queries of the form ‘retrieve all alpha hemoglobin genes’ are not only possible, but are exhaustive and correct. To support this effort, controlled vocabularies have been developed that include genes, gene families, gene products, developmental stages, differentiation stages, expression levels and experiment types. This allows, for example, information to be extracted based on gene name and gene family names, something that cannot be done with GenBank entries.

## QUERIES, VISUALIZATION AND ANALYSIS TOOLS

The EpoDB server, version 2.1, can be accessed through the EpoDB home page at http://cbil.humgen.upenn.edu/epodb/ . The site is best viewed using a Java enabled browser. A graphic visualization of the data in EpoDB is made available using the map and sequence bioWidgets (7), implemented in Java. These comprise a set of graphical units designed for the creation of adaptable, reusable graphical user interfaces, deployed in modules that are easily incorporated in a variety of applications, and in such a way as to promote interaction between those applications (8).

Currently, the EpoDB WWW server provides the following types of queries:

- Entries can be retrieved by, e.g., gene and gene family name, and features and sequence displayed textually or graphically using the bioWidget Map and Sequence viewers. Queries return a link to the protein entry which can also be viewed textually or graphically.
- Nucleic acid subsequences can be specified for genes (or gene families) by reference to features (5′UTR, CDS, exon, intron, etc.) or feature boundaries (start/end of transcription/translation) and retrieved over user defined intervals. Retrieved subsequences can then be passed to analysis programs such as TESS (9) or GenLang (10) for promoter sequence and motif analysis.
- Text searches over the nucleic acid and protein entries are supported.
- Nucleic acid and protein sequences can be searched by BLAST queries.

## SUMMARY

In conclusion, our immediate goal in building EpoDB has been to provide more powerful access, analysis and visualization of information relevant to gene expression. In the future, we expect EpoDB to support simulation of gene expression and gene networks. While we have focused on the process of erythropoiesis, the tools we have implemented are generic and can be applied to the study of any well-defined differentiating system.

## ACKNOWLEDGEMENTS

This work was supported by a grant number R01-RR04026-07A2 from the National Center for Research Resources, NIH.

## REFERENCES

1 Stamatoyannopoulos,G., Nienhuis,A.W., Majerus,P.W. and Varmus,H. (eds) (1994) *Molecular Basis of Blood Diseases.* W.B. Saunders Co., Philadelphia, PA.
2 Salas,F., Haas,J., Stoeckert,C.J. and Overton,G.C. (1997) *Lecture Notes Comput. Sci*., **1278**, 52–61.
3 Benson,D.A., Boguski,M.S., Lipman,D.J. and Ostell,J. (1997) *Nucleic Acids Res*., **25**, 1–6 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 1–7].
4 Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res*., **25**, 31–36 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 38–42].
5 Wingender,E., Kel,A.E., Kel,O.V., Karas,H., Heinemeyer,T., Dietze,P., Knüppel,R., Romaschenko,A.G. and Kolchanov,N.A. (1997) *Nucleic Acids Res*., **25**, 265–268 [see also this issue, *Nucleic Acids Res.* (1998) **26**, 362–367].
6 Overton,G.C., Aaronson,J., Haas,J. and Adams,J. (1994) *J. Comput. Biol*., **1**, 3–13.
7 CBIL bioWidgets (1996) http://agave.humgen.upenn.edu/bioWidgets/ Computational Biology and Informatics Laboratory, University of Pennsylvania.
8 Searls,D.B. (1995) *Gene*, **163**, GC 1–16.
9 Schug,J. and Overton,G.C. (1997) Technical Report CBIL-TR-1997-1001-v0.0, Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania.
10 Dong,S. and Searls,D.B. (1994) *Genomics*, **23**, 540–551.