

EchoBASE: an integrated post-genomic database for *Escherichia coli*

Raju V. Misra, Richard S. P. Horler, Wolfgang Reindl, Igor I. Goryanin¹ and Gavin H. Thomas*

Department of Biology (Area 10), University of York, PO Box 373, York, YO10 5YW, UK and ¹Scientific Computing and Mathematical Modelling, GlaxoSmithKline, Gunnels Wood Road, Stevenage, SG1 2NY, UK

Received August 13, 2004; Accepted September 21, 2004

ABSTRACT

EchoBASE (<http://www.ecoli-york.org>) is a relational database designed to contain and manipulate information from post-genomic experiments using the model bacterium *Escherichia coli* K-12. Its aim is to collate information from a wide range of sources to provide clues to the functions of the approximately 1500 gene products that have no confirmed cellular function. The database is built on an enhanced annotation of the updated genome sequence of strain MG1655 and the association of experimental data with the *E.coli* genes and their products. Experiments that can be held within EchoBASE include proteomics studies, microarray data, protein–protein interaction data, structural data and bioinformatics studies. EchoBASE also contains annotated information on ‘orphan’ enzyme activities from this microbe to aid characterization of the proteins that catalyse these elusive biochemical reactions.

INTRODUCTION

The bacterium *Escherichia coli* K-12 is the most thoroughly studied free-living organism and the completion of the genome sequence of strain MG1655 in 1997 was a landmark event in the study of this model organism (1). Our understanding of this species has been boosted by the genome sequences of the pathogenic strains *E.coli* O157:H7 (2,3) and *E.coli* CFT073 (4). The genome sequence has provided the opportunity of being able to identify all the potential components of the cell, and the annotation of *E.coli* K-12 strain MG1655 has recently been updated with some sequence and resulting gene annotation changes. One striking fact remains 7 years after the sequence was finished: the physiological functions of nearly 35% of its gene products are unknown, while almost 20% of the genome contains genes for which we cannot make a confident prediction of function (functionally unknown or FUN genes) (5–7).

Data is continuously being generated about FUN genes using many traditional and post-genomic research techniques, including proteomics, biochemical studies, microarrays, structural and bioinformatics approaches. At present, tracing all the published data on a particular FUN gene is difficult and time consuming and there is a need to collate and organize such data into a manageable and efficient database system. Integration of a wide range of information about a particular FUN gene can work synergistically to help in prediction of its function.

We describe herein the creation of EchoBASE, a new database that integrates information from post-genomic experiments into a single resource. For basic curation of gene product information, the database relies on features from other selected *E.coli* databases, its novelty being the linkage of curated experimental data to individual genes and ability to manipulate data from genome-wide experiments. While we aim to predict biological functions for uncharacterized gene products, there are existing lists of biochemical activities that have been identified in *E.coli* but not mapped to a gene product, so-called ‘orphan enzymes’. To complete our knowledge of metabolic pathways in *E.coli*, it is essential to identify the genes that encode these ‘orphan’ enzymes and we describe our curation and analysis of ‘orphan’ enzymes in *E.coli* as a component of EchoBASE.

EchoBASE is the major evolution of a simple HTML catalogue of functional updates of uncharacterized gene products that was begun in 1998 (7), and is a component of the *E.coli* index WWW site (<http://ecoli.bham.ac.uk/>) (8).

DESCRIPTION OF EchoBASE

EchoBASE is a relational database that was designed and implemented using the MySQL database management system and Macromedia ColdFusion MX server technology. The original version was launched in January 2004 and this paper describes version 1.2, released on September 1. Although the database focuses on holding data from post-genomic experiments, it also provides a basic annotation of the *E.coli* K-12 genome. This annotation was based on EcoGene 16, which we considered the most accurate annotation in 2003 (9). However,

*To whom correspondence should be addressed. Tel: +44 1904 328678; Fax: +44 1904 328825; Email: ght2@york.ac.uk

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use permissions, please contact journals.permissions@oupjournals.org.

in version 1.2, we have added a number of changes to the sequence as a result of additional sequence data presented in version m56 of the GenBank entry U00096, the first movement towards a single united annotation of the MG1655 genome sequence which will be held in the ASAP database (10).

Each gene record contains a functional description of the product, the location and direction of the gene on the chromosome and predictions of the properties of the gene product (Figure 1). The nucleotide and amino acid sequences are also held in the database. A simple graphical genome navigation tool has been incorporated into the gene page in version 1.2, which shows information relating to relative gene size, chromosomal position and uses a colour coding scheme that reflects the type of features being illustrated, e.g protein-coding sequence, rRNA, sRNA. For additional information on gene products, *EchoBASE* links out to a suite of other databases that provide a variety of complementary data. For existing literature we link to EcoGene (9), for information on transcription units and metabolic pathways we link to EcoCyc (11), for data on protein domains and families we link to GenProtEC (12) and for comparative genomics tools we link to *coliBASE* (13).

We have added some novel in-house whole-genome annotation, the most useful being a genome-wide survey of predicted subcellular location, the *EchoLOCATION* feature.

The subcellular location of a protein can often provide insight into its functional role in the cell, and we have combined and manually processed data from SignalP v. 2.0 (14), LipOP v. 1.0 (15), TMHMM v. 2.0 (16) and HMMTOP (17) to make a prediction for each gene product.

CURATION OF EXPERIMENTAL INFORMATION

The focus of *EchoBASE* is on the integration of post-genomic data to provide greater insight into the potential functions of the many FUN genes. Currently, data is manually curated from published experimental research and sorted into one of the following ‘types’ of experiments: proteomics; microarray; transcript level; genetics; biochemistry; bioinformatics, protein-protein interactions and structure. Each of these experiment types have been created to hold a wide range of different data types that are commonly determined in each of the different experiment types. The bioinformatics experiment type is the most generic schema due to the large variety of data types that could be usefully included in the database. After an experiment is sorted into an experiment type, it is then handled differently depending on whether it is a ‘single or few gene experiment’ which provides data for up to 15 different gene products, or a ‘genome-wide’ experiment. Data from the

The screenshot shows the *EchoBASE* interface for the *cueO* gene. The top box displays the gene name (*cueO*), Blattner ID (b0123), synonyms (*yucK*, *cutD*, *(RL)*), and function (Multi-copper oxidase (aerobic)). It also includes a table of genomic coordinates and a list of databases where the gene is annotated. The middle section, *EchoLOCATION*, provides a signal peptide prediction. The right-hand boxes contain detailed experimental data, including proteomics evidence for 233 gene products and transcript level data from 1945.00. The proteomics section includes a table with columns for M_{av} obs, pI obs, N-Term seq, and processing details. The transcript level section includes a table with columns for M_{av} obs, pI obs, N-Term seq, and processing details. The right-hand boxes also include a table of protein properties such as pI predicted, pI observed, MW predicted, and MW observed.

Figure 1. Screenshot of the *EchoBASE* gene page for *cueO*. The gene and protein sequence are linked from the gene name (top box). There are a number of different experiments associated with this gene and information for a proteomics experiment by Link *et al.* is displayed (top right box). This is the experiment overview page, with details of the experiment and our basic comment, and below this are the options for displaying different subsets of the data from this paper and to rank this by particular properties. Specific information in this paper that relates to CueO can be viewed in detail by following another link which includes our full annotation of the papers and all relevant information for the selected gene (right box).

EchoORPHAN

Orphan Enzymes List

The coloured boxes within the enzyme rating column, refer to the data source used to identify the orphan enzymes.

■ = EcoCyc list ■ = Serres *et al.* list ■ = Bernhard Pallson lab list

Rank	Enzyme Name	EC Number	Enzyme Rating	Status	Gene
1	Aminobutyraldehyde dehydrogenase	1.2.1.19	83 ■ ■ ■		ydcW
2	Methionine adenosyltransferase 2	2.5.1.6	82 ■		
3	Succinate semialdehyde dehydrogenase, NAD-dependent	1.2.1.24	78 ■ ■ ■		
4	ADP-sugar pyrophosphatase	EchoORPHAN Aminobutyraldehyde dehydrogenase			
5	Arabinose-5-phosphate isomera	Alternative names gamma-aminobutyraldehyde dehydrogenase 4-aminobutyraldehyde dehydrogenase 4-aminobutanal dehydrogenase dehydrogenase, aminobutyraldehyde ABAL dehydrogenase 4-aminobutanal:NAD+ 1-oxidoreductase Enzyme rating 83 Rank order in list 1 Present in the lists of: EcoCyc. Serres. Pallson lab.			
6	Isoamylase	Information gathered to support this orphan enzyme			
7	3-deoxy-D-manno-octulosonate phosphatase	Supporting evidence Purification and mapping to region			
8	Glutaminase B	MW (kDa)	95 +/- 1 (monomer), 195 +/- 10 (from gel filtration)	Fold purification	350-fold
9	D-galactose 1-dehydrogenase	Km (uM)	31.3 +/- 6.8 uM (for delta-1-pyrroline) 53.8 +/- 7.4 uM (for NAD)	Temp. optimum	
10	NMN amidohydrolase	Inhibitors	NADH	Activators	
		Pathway	Putrescine degradation	EC code	1.2.1.19
		Substrates	4-aminobutyraldehyde	Cellular localisation	
		Locus	Mapped between 28 and 32 minutes	<i>E. coli</i> strain	K12 (mutants able to grow on putrescine as sole carbon and nitrogen source)
				Gene Name	<i>prf</i>
		EchoBASE analysis and predictions			
	Methods used	Colibri-search [combined locus-Mw-search]. BLAST search with known enzymes with this activity from <i>H. sapiens</i> , <i>B. melitensis</i> and <i>P. sativum</i> .			
	Predictions	A Colibri-search with the Mw published by Prieto <i>et al.</i> in 1987 (Mw 92.5-102.5 kDa) and the locus (28-32) proposed by Shaabe <i>et al.</i> in 1985 showed ydbH as possible gene as possible gene for this enzyme. The gene product YdbH is a hypothetical periplasmic protein. A BLAST-search of the <i>H. sapiens</i> enzyme showed ydcW (1* <i>e</i> -78) and yneI (7* <i>e</i> -60), of the <i>B. melitensis</i> enzyme ygaF (1* <i>e</i> -115) and of the <i>P. sativum</i> enzyme again yneI (2* <i>e</i> -53) as possible targets. yneI is located at 34.733 and codes for a putative aldehyde dehydrogenase, however the Mw of the respective protein is only 49.5 kDa (would fit, if enzyme is no dimer but tetramer).			
	Further comments	In a review from the group of Larry Reitzer in May 2003 it is reported that <i>ydcW</i> is the gene for this orphan enzyme. This is one of the genes that we predicted from doing BLAST searches. It maps to 32.63 minutes which was just out of our search range of 28-32 minutes from the literature. Also, the Mw of YdcW is 50 kDa which suggests it was purified as a dimer not a monomer and is a tetramer under native conditions.			
	Reference data for Aminobutyraldehyde dehydrogenase				
	Publication title	Properties of gamma-Aminobutyraldehyde Dehydrogenase from Escherichia coli			
	Authors	Prieto MI, Martin J, Balana-Fouce R, Garrido-Pertierra A			
	Journal details	Biochimie 1987 Nov-Dec;69(11-12):1161-8			
	Publication title	Metabolic Pathway for the Utilization of L-Arginine, L-Ornithine, Agmatine, and Putrescine as Nitrogen Sources in Escherichia coli K-12			
	Authors	Shaabe E, Metzger E, Halpern YS			
	Journal details	J Bacteriol 1985 Sep;163(3):933-7			
	Data Entered By	W.Reindl	Entry Date	2003/06/17	Amendment History

Figure 2. Screenshot of the EchoORPHAN page from the database. The initial page displays the 'Enzyme rating', its current status and the EC code linking to BRENDA (14). This ranked list is superimposed with the detail page for the orphan enzyme aminobutyraldehyde dehydrogenase which was our highest scoring orphan enzyme and has recently been mapped to the *ydcW* gene product.

former type are usually manually curated into the database like the previous catalogue version (7), and data from the latter are parsed from data sets either provided directly within the publication, from supplementary information available from the publisher's WWW site, or from direct contact with the authors.

SEARCHING AND MANIPULATING EXPERIMENTAL DATA IN EchoBASE

The data held within EchoBASE can be navigated either by browsing/searching different experiment types or by searching for a particular gene. Browsing different experiment types

returns a summary list of all experiments of a particular type, e.g. proteomics, from which one can be selected to view our full annotation of the paper. As well as a textual description of the paper and the methods used, the data presented in the manuscript can be displayed in a number of ways. For example, for a proteomics experiment, the data can be sorted by different molecular properties, such as pI observed or relative abundance. This allows the user new ways to look at published results to enable faster sorting and extraction of data useful to them. Once an interesting piece of data has been spotted for a gene of interest, the data relating to this particular gene and experiment can then be viewed in full.

An alternative route into the database is through a particular researcher. All experimental annotations are linked to individual principle investigators and the data can be browsed to see which experiments from a particular group are contained within the database. Currently, there are over 400 research groups in the database, which has been built from a list previously compiled in the *E.coli* index (8).

In version 1.2, we have implemented the 'Complex search' which increases the power of the database by allowing complex questioning of the data set. For example, if a researcher was looking for candidate gene for a periplasmic binding protein involved in transporting an alternative nitrogen source, they could search the data for proteins that were (i) predicted to be periplasmic binding proteins (bioinformatics), (ii) demonstrated to be located in the periplasm (proteomic) and (iii) induced during nitrogen limiting conditions (microarray). This could be combined with searches for proteins of a certain molecular weight range that are encoded by genes located at a certain position on the genome sequence.

ORPHAN ENZYMES

Information on orphan enzymes in *E.coli* has been extracted from three sources, which were initially merged into a single comprehensive list. Data was taken from a list published by Riley and Serres (18), a list from the EcoCyc database (11) and a list used to construct an *in silico* metabolic genotype by Edwards and Palsson (19). This list had 100 different enzyme/protein names that could be considered to be orphan enzymes, and 36 of these were immediately removed as they had actually been linked to genes. To assess the remaining 64 activities, a marking scheme was created (out of 100) that was used to estimate how strong the evidence was to support each orphan enzyme. Details of the scheme can be found in the database, but consider factors like whether the activity had been purified with a known molecular weight and whether the locus linked to the activity had been mapped to a region on the genome. A score was given to 57 of the activities, varying between 3 and 83. Since creating the list in July 2003, 4 orphan enzymes have been mapped to their genes and 3 of these are in the top 7 scoring orphan enzymes in our list, strongly supporting our scoring system. The fourth activity that has been mapped was surprisingly rank order 29, mainly due to the only evidence coming from mapping the locus to a small region on the genome.

Given the likelihood that some of these activities are side reactions of known enzymes and that some probably are not present in MG1655, we estimate that there are around

25 genuine orphan enzymes to be mapped to genes in *E.coli* K-12 MG1655. The list is continuously updated and newly mapped orphan enzymes are highlighted within the list and a link to the appropriate gene is created (Figure 2).

ADDITIONAL FEATURES OF *Echo*BASE

One feature we consider important is data tracking within the database so that all changes can be traced. Therefore most pages in the database contain an 'Annotation history' field. *Echo*BASE will be one of the first databases to have implemented the new sequence and annotation of the *E.coli* genome (version m56), a significant sequence and annotation change released in 2004. All the sequence changes that alter coding regions, which number over 150, can be found associated with each gene and also have been described in the 'Annotation' page of the database that keeps a record of all changes to genes. Sequence changes result in addition and removal of Blattner numbers as genes are split and fused, changes in lengths of coding sequences as gene starts and stops are changed, and amino acid changes that have occurred as a result of base substitutions within protein-encoding genes.

*Echo*BASE also has a series of help pages to guide the user around the database and can provide data sets for other users that relate our unique identifier (the EB number) to those of the other resources we link out to from the gene page (Figure 1). Currently, we encourage researchers to send us details of their discoveries that we curate, but eventually we would like to move to a more community-based annotation.

ACKNOWLEDGEMENTS

We would like to acknowledge CNAP and the University of York for technical support and hosting *Echo*BASE, and Louise Fairweather for collecting information for *Echo*LOCATION during her MRes project. We thank GlaxoSmithKline and the BBSRC for financial support.

REFERENCES

1. Blattner, F.R., Plunkett, G., III, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
2. Perna, N.T., Plunkett, G., III, Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.
3. Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.G., Ohtsubo, E., Nakayama, K., Murata, T. *et al.* (2001) Complete genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.*, **8**, 11–22.
4. Welch, R.A., Burland, V., Plunkett, G., III, Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J. *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 17020–17024.
5. Hinton, J.C. (1997) The *Escherichia coli* genome sequence: the end of an era or the start of the FUN? *Mol. Microbiol.*, **26**, 417–422.
6. Serres, M.H., Gopal, S., Nahum, L.A., Liang, P., Gaasterland, T. and Riley, M. (2001) A functional update of the *Escherichia coli* K-12 genome. *Genome Biol.*, **2**, RESEARCH0035.

7. Thomas, G.H. (1999) Completing the *E. coli* proteome: a database of gene products characterised since the completion of the genome sequence. *Bioinformatics*, **15**, 860–861.
8. Thomas, G.H. and Bettelheim, K.A. (1998) *Escherichia coli* on the WWW. *Lett. Appl. Microbiol.*, **27**, 122–123.
9. Rudd, K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.
10. Glasner, J.D., Liss, P., Plunkett, G., III, Darling, A., Prasad, T., Rusch, M., Byrnes, A., Gilson, M., Biehl, B., Blattner, F.R. *et al.* (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.*, **31**, 147–151.
11. Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C. and Gama-Castro, S. (2002) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56–58.
12. Serres, M.H., Goswami, S. and Riley, M. (2004) GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res.*, **32**, D300–D302.
13. Chaudhuri, R.R., Khan, A.M. and Pallen, M.J. (2004) coliBASE: an online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics. *Nucleic Acids Res.*, **32**, D296–D299.
14. Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, **8**, 581–599.
15. Juncker, A.S., Willenbrock, H., von Heijne, G., Brunak, S., Nielsen, H. and Krogh, A. (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.*, **12**, 1652–1662.
16. Sonnhammer, E.L., von Heijne, G. and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
17. Tusnady, G.E. and Simon, I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics.*, **17**, 849–850.
18. Riley, M. and Serres, M.H. (2000) Interim report on genomics of *Escherichia coli*. *Annu. Rev. Microbiol.*, **54**, 341–411.
19. Edwards, J.S. and Pálsson, B.O. (2000) The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. USA*, **97**, 5528–5533.