D358–D362 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj037

ODB: a database of operons accumulating known operons across multiple genomes

Shujiro Okuda[1,*], Toshiaki Katayama[2], Shuichi Kawashima[2], Susumu Goto[1] and Minoru Kanehisa[1,2]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan and [2]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai Minato-ku, Tokyo 108-8639, Japan

Received August 15, 2005; Revised and Accepted September 26, 2005

ABSTRACT

Operon structures play an important role in co-regulation in prokaryotes. Although over 200 complete genome sequences are now available, databases providing genome-wide operon information have been limited to certain specific genomes. Thus, we have developed an ODB (Operon DataBase), which provides a data retrieval system of known operons among the many complete genomes. Additionally, putative operons that are conserved in terms of known operons are also provided. The current version of our database contains about 2000 known operon information in more than 50 genomes and about 13 000 putative operons in more than 200 genomes. This system integrates four types of associations: genome context, gene co-expression obtained from microarray data, functional links in biological pathways and the conservation of gene order across the genomes. These associations are indicators of the genes that organize an operon, and the combination of these indicators allows us to predict more reliable operons. Furthermore, our system validates these predictions using known operon information obtained from the literature. This database integrates known literature-based information and genomic data. In addition, it provides an operon prediction tool, which make the system useful for both bioinformatics researchers and experimental biologists. Our database is accessible at http://odb.kuicr.kyoto-u.ac.jp/.

INTRODUCTION

With the increasing availability of completely sequenced genomes, comparative genomic approaches are becoming more important to decipher the functions of genes. Methods, which are powerful, using the conservation of gene proximity on genomes (i.e. determining potential operons) can understand functional associations between genes (1–3). Genes in an operon are functionally associated with each other in prokaryotes; thus, various kinds of operon prediction methods have been developed to understand the functional relationships and to annotate genes (4–13). Databases that accumulate the experimentally verified operon information should be useful to validate such prediction methods and also to understand the functional association between genes. However, databases providing genome-wide operon information have been limited to certain specific genomes (14,15). Although the STRING database was developed to identify functional associations between genes for multiple genomes, it uses gene neighborhood based on genome context methods (16). Here, we introduce the database called ODB (Operon DataBase), which provides operon data documented in the literature and putative operons that are conserved in terms of known operons. Furthermore, to characterize operons, it integrates genome context, gene co-expression obtained from microarray data, functional links in biological pathways and data on the conservation of gene order across genomes. ODB also provides operon prediction based on these various types of data as an application of our database. These datasets are fully precomputed so that all information can be quickly accessed. The ODB database integrates known literature-based information and genomic data. In addition, it provides an operon prediction tool, which makes the system useful for both bioinformatics researchers and experimental biologists.

OPERON DATA SOURCES

We have collected information of known operons of multiple genomes from the literature. We note that the experimentally verified operons, which we have collected, have been verified by a variety of means, from direct measurements such as primer extension and northern blots to less direct methods

*To whom correspondence should be addressed. Tel: +81 774 38 3270; Fax: +81 774 38 3269; E-mail: okuda@kuicr.kyoto-u.ac.jp

Table 1. Statistics of operons in major genomes

Species	No. of operons	No. of putative operons
Eukaryotes: 7 species		
Caenorhabditis elegans	628	149
Saccharomyces cerevisiae	–	7
Prokaryotes: 177 species		
Bacillus subtilis	711	60
Escherichia coli	389	61
Pseudomonas aeruginosa	33	156
Agrobacterium tumefaciens	15	172
Synechocystis sp. PCC6803	12	26
Bradyrhizobium japonicum	10	190
Archaea: 19 species		
Methanosarcina acetivorans	–	44
Pyrococcus furiosus	2	13
Total: 203 species	1957	13 258

such as gene knock-out experiments. Our database represents an ongoing effort to increase the coverage of operons. The current version of our database contains about 2000 known operon information in more than 50 genomes obtained from a total of 825 literatures (Table 1). Note that although some of these operons overlap, we use the term 'operon' to refer to a 'transcriptional unit' individually as opposed to the generally understood usage of the term that may include multiple overlapping transcriptional units. These data also include the operons of *Caenorhabditis elegans*. Operon structures are often observed in prokaryotes, but nematodes also have similar transcriptional systems (17,18). Thus, we added the eukaryotic operons into our database. Note that the operons from *Bacillus subtilis* contain operons obtained from transcriptional maps stored in BSORF (http://bacillus.genome.jp/). Because these maps were derived from the results of northern blotting experiments, we added these operons into our database. Note that these entries can be distinguished from the operons obtained from the literature, as the origin of the source (BSORF) is annotated in the database.

Table 1 also shows putative operons that are conserved in terms of known operons. When we calculated these conservations, we used KEGG OC as the ortholog gene set (19), which is ortholog gene clustering based on Smith–Waterman sequence similarity scores. If genes in a known operon have ortholog genes in another genome and these ortholog genes are consecutively located on the same strand of the genome, we regarded them as a putative but highly reliable operon. Note that this is not applied to known mono-cistronic genes. Furthermore, the putative operons were also explored from the viewpoint of paralog genes. These putative operons are also explored in eukaryotes. Usually, we do not use the term 'operon' for the eukaryotic gene clusters, but we use this term operationally in our database. As a result, over 13 000 putative operons were observed in over 200 genomes.

OVERVIEW OF THE DATABASE

ODB uses a relational database management system (MySQL, http://www.mysql.com/) to store and manage all information including not only known and putative operons but also primary data, such as gene location and definition, and associations between genes. This system contains four types of associations between genes that determine an operon: (i) intergenic distances, (ii) functional links in biological pathways, (iii) gene co-expression obtained from microarray data and (iv) the conservation of gene order across multiple genomes. These four types of associations are considered indicators and that the genes linked by them can organize an operon. Therefore, we pre-calculated these associations among all genes in all available genomes to characterize operons. Genes in an operon are often closely located on the genome compared with those between non-operons. Therefore, this is one of the indicators to characterize operons. Intergenic distances are defined as the number of bases between the end position of a gene and the start position of the next gene on the genome.

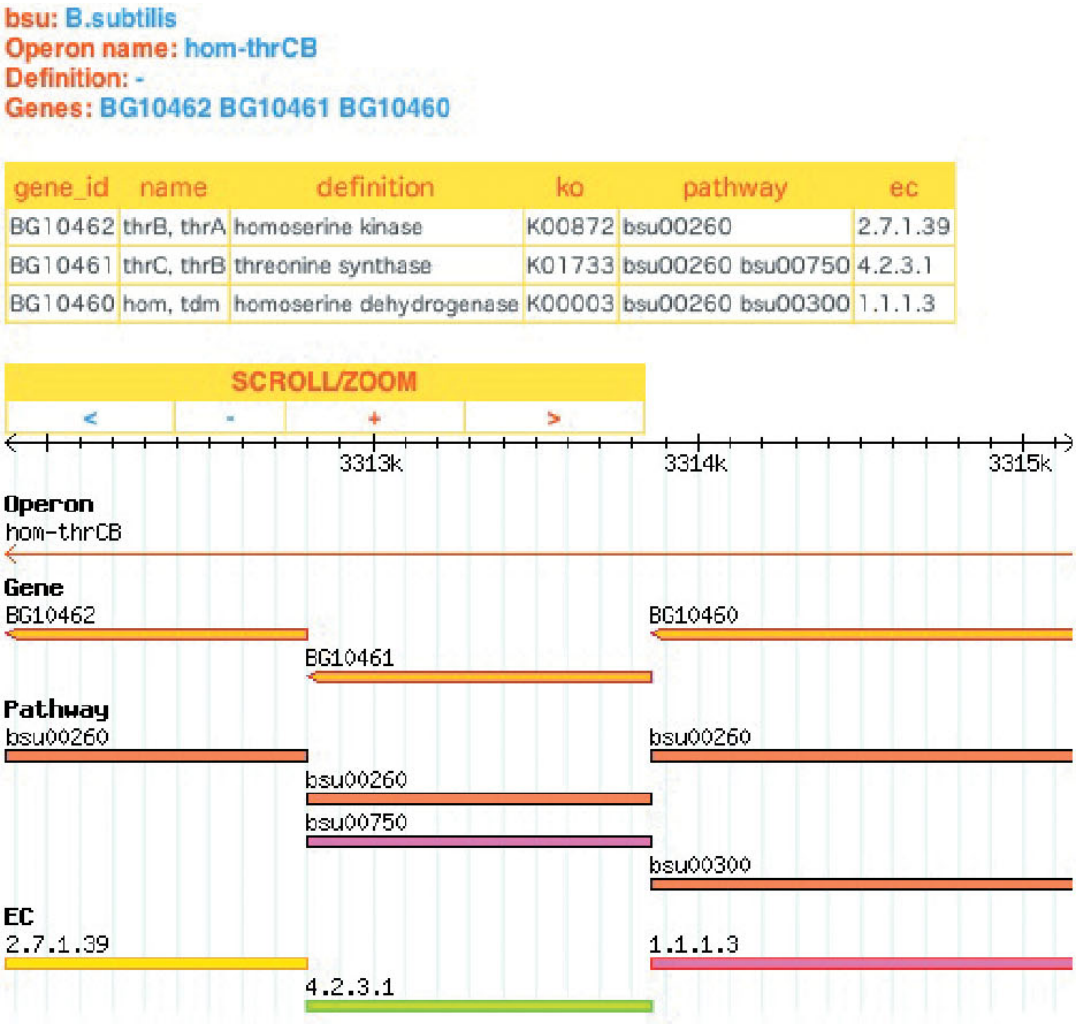
In addition, genes in an operon are often functionally related. For example genes appearing in a metabolic pathway are often clustered on the genome to be co-transcribed (20). Such functional links were obtained from KEGG pathway (19). We calculated the number of steps between genes in the pathway maps. The number of steps indicates that when two genes are linked across a compound, the number of steps is one. In this way, we calculated the number of steps not only in the same pathway map but also across different pathway maps.

The KEGG EXPRESSION database contains the gene expression data derived from microarrays of four organisms, *B.subtilis*, *Escherichia coli* K-12 W3110, *Synechocystis* sp. PCC6803 and *Saccharomyces cerevisiae* (19). We used the information of co-expressed genes from the database. We calculated the Pearson's correlation coefficients between gene expression profiles obtained from these microarray data. Because it is considered that microarray data reflect actual gene transcription and that they are powerful tools to predict operons, co-expressed gene clusters on the genome are possible operons. However, the limitation of experimental conditions and quality of the experiments still leave the issue that certain operons are not transcribed and that the level of gene co-expression is not homogeneous. Therefore, there are cases where genes are not co-expressed even if they are genes in a known operon.

Gene order in an operon is often shuffled and collapsed in evolutionary history (21,22). Therefore, conservation of gene order across genomes is rather rare, especially in distantly related genomes. If such conservation is observed, they are probably related to a physical interaction such as a molecular complex (23). Therefore, this feature is also important in characterizing operons. We calculated the step number between gene pairs. That is, given a gene pair, we took each of their ortholog genes from all genomes, calling this 'ortholog gene pair'. Then we calculated the step number between these two ortholog genes. When the gene pair is adjacently located on the genome, the step number is regarded as one. Here, we ignore the genomes included in the same taxonomic group, which are defined in KEGG (http://www.genome.jp/kegg/catalog/org_list.html).

All-against-all runs of these associations between genes were performed. Pre-computed results are stored in a table, allowing quick retrieval against the query specified by users. Each table in our system corresponds to a particular genome to facilitate efficient access and retrieval of the information.

When users search a gene or an operon of interest, the gene cluster including it can be identified by its name and identifier.



gene_id	name	definition	ko	pathway	ec
BG10462	thrB, thrA	homoserine kinase	K00872	bsu00260	2.7.1.39
BG10461	thrC, thrB	threonine synthase	K01733	bsu00260 bsu00750	4.2.3.1
BG10460	hom, tdm	homoserine dehydrogenase	K00003	bsu00260 bsu00300	1.1.1.3

INTERGENIC DISTANCE				
		0	1	2
BG10462	0	-	-1	1057
BG10461	1	-1	-	2
BG10460	2	1057	2	-

PATHWAY				
		0	1	2
BG10462	0	-	1	1
BG10461	1	1	-	2
BG10460	2	1	2	-

ORTHOLOG				
		0	1	2
BG10462	0	-	1	1
BG10461	1	1	-	1
BG10460	2	1	1	-

EXPRESSION				
		0	1	2
BG10462	0	-	0.80	0.87
BG10461	1	0.80	-	0.81
BG10460	2	0.87	0.81	-

Figure 1. An example view of an operon.

Then, the user is presented with a summary of genes and associations between genes in the region on the genome (Figure 1). Primary data such as gene names, gene IDs, definitions, KO IDs as functional classes, KEGG pathway IDs and EC numbers are presented. These are linked to the KEGG database if available. Additionally, the genomic view of the region of interest is also presented. This view includes graphical symbols of operons, genes, pathways and EC numbers and

each symbol is also linked to the KEGG database. The user can also scroll and zoom the region of interest on the genome. Finally, the four types of associations are shown as separate tables. For the biological pathway table, the shortest step numbers between genes are presented. For the ortholog gene table, the shortest step numbers between the ortholog gene pairs are shown. In these tables, additional pages are accessible which show the detail of the information. For the gene expression table, the correlation coefficients between gene expression profiles are shown, and the strength of co-expression is illustrated by a color gradient ranging from blue to red.

OPERON PREDICTION

Because the conditions to determine putative operons are very strict and are not genome-wide, ODB also provides a system to predict operons, using the four associations. Given a specific species, predicted operons that may exist within that species are returned. There are two options that are available: simple and advanced prediction mode. For a simple mode, users can obtain prediction results based on default parameter values that have been validated by known operons. However, in advanced prediction mode, users can freely change these parameter values, which are based on the four types of associations described above. When genes linked by these associations are clustered on the genome, they are likely to be an operon. Thus, we benchmarked the accuracy of the predictions based on combinations of various values of intergenic distances, step numbers between ortholog genes and the number of the genomes having conserved ortholog genes that are linked within a specific range of step numbers. Therefore, the optimal values that predict the largest number of operons while keeping the accuracy high is provided as default values in simple prediction mode (Supplementary data). When there is little or no known operon information in a genome, the default values of another genome in the same taxonomic group and having sufficient operon information is used as an alternative to the genome. If such genomes are also unavailable, we used the values of *B.subtilis* (see Supplementary Data for details).

CONCLUDING REMARKS

ODB provides a platform for searching known operons and consequent putative operons and for predicting operons with high accuracy validated by literature-based operon data. It includes about 2000 literature-based operons in over 50 genomes and about 13 000 putative operons in over 200 genomes. In addition, the data from KEGG pathway and related resources that are provided allow analyses not only based on a specific genomic context but also across genomes. Thus, it is the first of its kind to integrate operon data from a variety of genomes, providing a wide-ranging coverage of operons. This integrated system of both known literature-based and genomic data is a useful system for bioinformatics researchers and experimental biologists.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Kiyoko F. Aoki-Kinoshita for critical reading of our manuscript. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, the Japan Society for the Promotion of Science, and the Japan Science and Technology Agency. The computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University and the Super Computer System, Human Genome Center, Institute of Medical Science, the University of Tokyo. Funding to pay the Open Access publication charges for this article was provided by the grant-in-aid for scientific research from the Ministry of Education.

Conflict of interest statement. None declared.

REFERENCES

1. Tamames,J., Casari,G., Ouzounis,C. and Valencia,A. (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol*., **44**, 66–73.
2. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
3. Huynen,M., Snel,B., Lathe,W.,III and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res*., **10**, 1204–1210.
4. Bockhorst,J., Craven,M., Page,D., Shavlik,J. and Glasner,J. (2003) A Bayesian network approach to operon prediction. *Bioinformatics*, **19**, 1227–1235.
5. Bockhorst,J., Qiu,Y., Glasner,J., Liu,M., Blattner,F. and Craven,M. (2003) Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, **19** (Suppl. 1), i34–i43.
6. Craven,M., Page,D., Shavlik,J., Bockhorst,J. and Glasner,J. (2000) A probabilistic learning approach to whole-genome operon prediction. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **8**, 116–127.
7. Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res*., **29**, 1216–1221.
8. Sabatti,C., Rohlin,L., Oh,M.K. and Liao,J.C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res*., **30**, 2886–2893.
9. Yada,T., Nakao,M., Totoki,Y. and Nakai,K. (1999) Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics*, **15**, 987–993.
10. Zheng,Y., Szustakowski,J.D., Fortnow,L., Roberts,R.J. and Kasif,S. (2002) Computational identification of operons in microbial genomes. *Genome Res*., **12**, 1221–1230.
11. De Hoon,M.J., Imoto,S., Kobayashi,K., Ogasawara,N. and Miyano,S. (2004) Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. *Pac. Symp. Biocomput*., 276–287.
12. de Hoon,M.J., Imoto,S., Kobayashi,K., Ogasawara,N. and Miyano,S. (2003) Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. *Pac. Symp. Biocomput*., 17–28.
13. Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*., **33**, 880–892.
14. Makita,Y., Nakao,M., Ogasawara,N. and Nakai,K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res*., **32**, D75–D77.
15. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al*. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res*., **32**, D303–D306.
16. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005)

STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res*., **33**, D433–D437.
17. Blumenthal,T., Evans,D., Link,C.D., Guffanti,A., Lawson,D., Thierry-Mieg,J., Thierry-Mieg,D., Chiu,W.L., Duke,K., Kiraly,M. *et al*. (2002) A global analysis of *Caenorhabditis elegans* operons. *Nature*, **417**, 851–854.
18. Lercher,M.J., Blumenthal,T. and Hurst,L.D. (2003) Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. *Genome Res*., **13**, 238–243.
19. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
20. Ogata,H., Fujibuchi,W., Goto,S. and Kanehisa,M. (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res*., **28**, 4021–4028.
21. Teichmann,S.A. and Babu,M.M. (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol*., **20**, 407–410; Discussion 410.
22. Itoh,T., Takemoto,K., Mori,H. and Gojobori,T. (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol*., **16**, 332–346.
23. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci*., **23**, 324–328.