MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity

Ikuo Uchiyama1,*, Toshio Higuchi2 and Mikihiko Kawai1

1National Institute for Basic Biology, National Institutes of Natural Sciences, Nishigonaka 38, Myodaiji, Okazaki, Aichi 444-8585, Japan and 2Intec Systems Institute, Inc., 1-3-3 Shinsuna, Koto-ku, Tokyo 136-0075, Japan

Received September 14, 2009; Revised October 9, 2009; Accepted October 12, 2009

ABSTRACT

The microbial genome database (MBGD) for comparative analysis is a platform for microbial comparative genomics based on automated ortholog group identification. A prominent feature of MBGD is that it allows users to create ortholog groups using a specified subgroup of organisms. The database is constantly updated and now contains almost 1000 genomes. To utilize the MBGD database as a comprehensive resource for investigating microbial genome diversity, we have developed the following advanced functionalities: (i) enhanced assignment of functional annotation, including external database links to each orthologous group, (ii) interface for choosing a set of genomes to compare based on phenotypic properties, (iii) the addition of more eukaryotic microbial genomes (fungi and protists) and some higher eukaryotes as references and (iv) enhancement of the MyMBGD mode, which allows users to add their own genomes to MBGD and now accepts raw genomic sequences without any annotation (in such a case, it runs a gene-finding procedure before identifying the orthologs). Some analysis functions, such as the function to find orthologs with similar phylogenetic patterns, have also been improved. MBGD is accessible at http://mbgd.genome.ad.jp/.

INTRODUCTION

Nearly 1000 microbial genomes have been completely sequenced, and the number of sequences is still growing exponentially. The growth of this number will be even further accelerated by the recent advancement of next-generation sequencing technologies. Thanks to this vast amount of information, much progress has been made in genomics studies toward understanding microbial diversity. One of the promising approaches is the comparison of dozens of closely related or moderately related genomes, which is effective for analyzing critical differences among organisms and understanding the evolutionary process generating such diversity. Another important new advancement is the metagenomic approach, by which researchers can investigate the community structures of microbes and their gene contents in various environmental samples. To facilitate genomic diversity studies, however, effective utilization of the existing genomic data in terms of comparative genomics is crucial, although this becomes more difficult as the size of the genomic database increases.

MBGD (1,2) is a microbial genome database for large-scale comparative genomics based on comprehensive ortholog classification generated by a hierarchical clustering method, DomClust (3). As compared to other comparative genomics resources covering complete microbial genomes, such as CMR (4), MicrobesOnline (5), IMG (6), eggNOG (7) and OMA (8), a prominent feature of MBGD is that it allows users to create ortholog groups using a specified subgroup of organisms. This feature is useful for various types of comparative analysis, including comparisons among closely related as well as among distantly related organisms (1).

In addition to the flexible ortholog analysis functionality, we have recently enhanced the database content by incorporating various types of information regarding gene function and organism phenotype, adding more eukaryotic genomes and implementing several additional functionalities to facilitate large-scale comparative genome analysis. Here, we describe the recent enhancement of MBGD.

DEFAULT ORTHOLOG TABLE

Although one of the significant features of MBGD is that it allows users to create their own ortholog tables by specifying any set of organisms, MBGD also holds a precalculated ‘default ortholog table’, which is now extended to cover all the organisms stored in the database. Actually, the default ortholog table is created using the default set of organisms that contains one representative genome from each genus, but in the ‘extended’

*To whom correspondence should be addressed. Tel: +81 564 55 7629; Fax: +81 564 55 7625; Email: uchiyama@nibb.ac.jp

© The Author(s) 2009. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Table, genes of unselected genomes are also classified into an appropriate ortholog group as follows: each gene is classified into the ortholog group giving the best average similarity score if (i) that score is better than the smallest within-cluster score (i.e. the score assigned at the cluster root node) or (ii) that gene is also the most similar to that ortholog group in that genome (i.e. they are in a bidirectional best-hit relationship). Based on this extended default ortholog table, users can now access the ortholog cluster information from any genome information page. Users can also download the entire default ortholog table as a flat text file.

**ANNOTATION ASSIGNMENT TO EACH CLUSTER**

The ortholog cluster information page has been redesigned (Figure 1). Several types of information are generated from the annotation of its member genes and are added to this page as cluster annotation. The following procedure determines the title of each ortholog cluster: the occurrence of words in the title lines of the member genes are counted, and the words whose occurrence is above or equal to 30% of the most frequent words are retained as frequent words; after scoring each title line based on the frequency of frequent words, the cluster title is constructed by extracting the frequent words from the best-scoring title.

Each cluster entry also contains cross-references to the corresponding entries of COGs, KEGG Orthology, TIGRFAMs and Gene Ontology terms. A correspondence between an MBGD group and a group of another database is classified into ‘equivalent’, ‘supergroup’ and ‘subgroup’, which are defined based on the following set-comparison procedure: let \( A \) and \( B \) be the MBGD group and the other group, respectively, containing only organisms commonly included in both sets, and let \( \alpha = |A \cap B|/|A| \), \( \beta = |A \cap B|/|B| \) and \( F = 2\alpha\beta/(\alpha + \beta) \); we defined \( B \) as being equivalent to \( A \) if \( F \geq 0.7 \); otherwise, \( B \) is a supergroup of \( A \) if \( \alpha \geq 0.7 \) or a subgroup of \( A \) if \( \beta \geq 0.7 \). In these cases, \( B \) is assigned as a cross-reference entry of \( A \) (Figure 1).

As previously, each cluster entry is assigned functional categories, but the definition of functional category has been extended: in addition to the original definition (1), users can now choose a functional category system from among those defined in other databases (COG, KEGG and TIGR); category assignment to each cluster is based on a majority vote of categories assigned to individual genes referring to the cross-reference data.

In the cluster entry page, several comparison functions are available, such as multiple map comparison and multiple sequence alignment (Figure 1). In addition, a function to search for clusters with similar phylogenetic patterns is now available. Here, the cluster table is ordered

---

**Figure 1.** Ortholog cluster entry page displaying the orthologous group of the cobalamin biosynthetic gene, *cobI*, of the default cluster set. The page contains a cluster annotation table showing the gene name, title and cross-references to other databases, and a table showing the list of member genes.
according to the dissimilarity in phylogenetic pattern between each cluster and the original cluster (Figure 2), where the dissimilarities are calculated based on a correlation coefficient, the hamming distance or mutual information (9). This function is useful for predicting functional linkages (10) and similar functions are implemented in some more specialized databases (11,12). In MBGD, users can combine this type of analysis with more flexible ortholog analysis.

ORGANISM SELECTION BASED ON PHENOTYPIC PROPERTIES

The utilization of information on the phenotypic properties of individual organisms is becoming more important for comparative genomics studies as the number of genomes increases. The current version of MBGD provides an interface for specifying a set of organisms to be analyzed using phenotypic information as well as taxonomic information as a reference, where the phenotypic properties are taken from the organism metadata collection in the GOLD database (13) that includes cell shape, motility, oxygen requirements, temperature range and so on (Figure 3). Users can also use a similar interface to specify a phylogenetic pattern in order to search for orthologous groups having similar phylogenetic patterns or specify species colors to interpret the phylogenetic patterns displayed in the header of the ortholog table (Figure 2).

ADDITION OF EUKARYOTIC GENOMES

MBGD is periodically updated using the complete genome data in the RefSeq database as a data source. Although previous MBGD versions mainly contained prokaryotic genomes (except *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* for reference purposes), we have now added more complete genome sequences of eukaryotic microbes belonging to the fungi and protists such as *Plasmodium falciparum*, *Dictyostelium discoideum*, *Aspergillus nidulans* and *Candida glabrata*. In addition, the complete genome sequences of some higher eukaryotes, including *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Homo sapiens*, have also been added as a reference. To incorporate the eukaryotic genomic data, we have modified our genome map viewer as well as the database schema to correctly display eukaryotic genomes where the genes are interrupted by introns.

ENHANCEMENT OF THE MyMBGD FUNCTIONALITY

The MyMBGD functionality, which allows users to add their own genome sequences to MBGD, has been enhanced: we now provide the ‘gene prediction mode’, in which users can submit genomic nucleotide sequences without any annotation and ask the system to predict the genes within them. The gene-finding procedure implemented here uses both the GeneMarkS program (14) and the Glimmer3 program (15) and merges their
outputs by taking the longer region when two programs predict different start positions in the same reading frame. The predicted genes are then subjected to the DomClust procedure (3) after an all-against-all similarity search, which is the usual MyMBGD procedure described previously (2). MyMBGD also provides the `metagenome mode', which accepts a set of nucleotide or protein sequences from a mixture of genomes and applies an ortholog assignment procedure similar to that for extending the default ortholog table described above, except for omitting the secondary condition for testing the bidirectional best hit. With this enhancement, users can now use the MyMBGD functionality as a tool to annotate a newly sequenced genome or new metagenome data.

FUNDING

Institute for Bioinformatics Research and Development, Japan Science and Technology Agency; Grant-in-Aid for Publication of Scientific Research Results from Japan Society for the Promotion of Science (218061). Funding for open access charge: Institute for Bioinformatics Research and Development, Japan Science and Technology Agency.

REFERENCES

1. Uchiyama, I. (2003) MBGD: microbial genome database for comparative analysis. Nucleic Acids Res., 31, 58–62.
2. Uchiyama, I. (2007) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. Nucleic Acids Res., 35, D343–D346.
3. Uchiyama, I. (2006) Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. Nucleic Acids Res., 34, 647–658.
4. Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K. and White, O. (2001) The comprehensive microbial resource. Nucleic Acids Res., 29, 123–125.
5. Alm, E.J., Huang, K.H., Price, M.N., Koche, R.P., Keller, K., Dubchak, I.L. and Arkin, A.P. (2005) The MicrobesOnline Web site for comparative genomics. Genome Res., 15, 1015–1022.
6. Markowitz, V.M., Szeto, E., Palaniappan, K., Grechkin, Y., Chu, K., Chen, I.M., Dubchak, I., Anderson, L., Lykidis, A., Mavromatis, K., et al. (2008) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res., 36, D258–D253.
7. Jensen, L.J., Julien, P., Kuhn, M., van Mering, C., Muller, J., Doerks, T. and Bork, P. (2008) eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res., 36, D250–D254.
8. Schneider, A., Dessiorms, C. and Gonnet, G.H. (2007) OMA Browser–exploring orthologous relations across 352 complete genomes. Bioinformatics, 23, 2180–2182.

Conflict of interest statement. None declared.
9. Wu, J., Kasif, S., and DeLisi, C. (2003) Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, 19, 1524–1530.

10. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, 96, 4285–4288.

11. Enault, F., Sahre, K., and Claverie, J. M. (2005) Phydbac ‘Gene Function Predictor’: a gene annotation tool based on genomic context analysis. *BMC Bioinformatics*, 6, 247.

12. Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M. *et al.* (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, 37, D412–D416.

13. Liolios, K., Mavromatis, K., Tavernarakis, N., and Kyrpides, N. C. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, 36, D475–D479.

14. Besemer, J., Lomsadze, A., and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, 29, 2607–2618.

15. Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, 27, 4636–4641.