Published online 30 October 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D631–D635***
doi:10.1093/nar/gks1006

MBGD update 2013: the microbial genome database for exploring the diversity of microbial world

Ikuo Uchiyama[1,2,*], Motohiro Mihara[3], Hiroyo Nishide[2] and Hirokazu Chiba[1]

[1]Laboratory of Genome Informatics, [2]Data Integration and Analysis Facility, National Institute for Basic Biology, National Institutes of Natural Sciences, Nishigonaka 38, Myodaiji, Okazaki, Aichi 444-8585 and [3]Dynacom Co. Ltd., Kobe Incubation Office 202, 9-1, Minatojima, Chuo-ku, Kobe, Hyogo 650-0045, Japan

Received September 15, 2012; Accepted October 1, 2012

ABSTRACT

The microbial genome database for comparative analysis (MBGD, available at http://mbgd.genome.ad.jp/) is a platform for microbial genome comparison based on orthology analysis. As its unique feature, MBGD allows users to conduct orthology analysis among any specified set of organisms; this flexibility allows MBGD to adapt to a variety of microbial genomic study. Reflecting the huge diversity of microbial world, the number of microbial genome projects now becomes several thousands. To efficiently explore the diversity of the entire microbial genomic data, MBGD now provides summary pages for pre-calculated ortholog tables among various taxonomic groups. For some closely related taxa, MBGD also provides the conserved synteny information (core genome alignment) pre-calculated using the CoreAligner program. In addition, efficient incremental updating procedure can create extended ortholog table by adding additional genomes to the default ortholog table generated from the representative set of genomes. Combining with the functionalities of the dynamic orthology calculation of any specified set of organisms, MBGD is an efficient and flexible tool for exploring the microbial genome diversity.

INTRODUCTION

Thanks to the advancement of DNA sequencing technologies that has drastically reduced sequencing cost, microbial genome sequencing has become a common task for every researcher in microbiology. As a result, public sequence databases now contain nearly 2000 complete sequences and several thousands more incomplete sequences of microbial genomes. This vast amount of sequence data is valuable resource for studies of microbial diversity including comparative analysis of closely related microbes as well as metagenome analysis of various environmental samples, although effective use of this resource is becoming more difficult due to the rapid data accumulation.

MBGD is a microbial genome database for comparative analysis based on large-scale ortholog analysis conducted through a hierarchical clustering method, DomClust (1). Among many microbial genome databases such as CMR (2), MicrobesOnline (3), IMG (4) as well as comprehensive ortholog database including prokaryotic genomes such as eggNOG (5) and OMA (6), MBGD has a unique feature that it allows users to choose any subset of organisms to create ortholog table among them (7), in addition to providing the default ortholog table that contains a representative organism from each genus covering the entire taxonomic range. This feature makes MBGD useful not only for comprehensive comparison of distantly related organisms but also for any kind of comparison including detailed comparisons of closely or moderately related organisms. In fact, one of the most common usages of MBGD is to compare genomes in a given taxonomic group that the user is interested in.

Since its launch in 1997 with a few genomes, MBGD has been constantly updated (usually two times per year in recent years) and now it contains well over 1000 genomes. The accumulation of genomic data that covers a broad range of taxonomic groups motivated us to reconsider the system to be more effective for users who are interested in microbial diversity in general, and now we decided to provide pre-calculated ortholog tables for several major taxonomic groups in addition to the default ortholog table. This enhancement allows users to see and compare genomic features of various taxonomic groups (in various taxonomic ranks such as species, genus and family) without creating ortholog table dynamically. To facilitate this type of analysis, the page for ortholog table overview has been modified to allow users to switch ortholog tables of different taxonomic groups. For closely related microbial genome comparisons, syntenic conservation is important information to identify correct

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

orthologs inherited through vertical transfers. In addition to the conventional ortholog tables created only from similarity relationship, MBGD now provides the 'core structure' of each taxonomic group that represents syntenically conserved regions among the given group extracted using the CoreAligner program (8).

Here, we introduce recent enhancements of the MBGD database especially focusing on utilizing taxonomic information in order to effectively represent and explore the microbial diversity.

DATA SOURCES

MBGD originally used the complete microbial genome section of the Reference Sequences (RefSeq) database (9) of National Center for Biotechnology Information (NCBI), whose annotation is partly assigned by NCBI and partly propagated from the original GenBank entry. The DNA Data Bank of Japan (DDBJ) has also provided a microbial genome database named Gene Trek in Prokaryote Space (GTPS) (10), whose annotation is assigned using its own pipeline. MBGD now combines RefSeq, GenBank and GTPS entries for its genomic data sources by integrating the annotations of the open reading frames (ORFs) in different databases that have the same 3′-end position. We took all GTPS ORFs that have a grade other than 'X' (no information available) followed by ORFs in RefSeq and GenBank in this order that are not overlapped with any gene previously included in the set. By this way, MBGD provides comprehensive collection of putative genes from the up-to-date collection of genomes for comparative analysis.

PRE-COMPUTED ORTHOLOG TABLES FOR TAXOMIC GROUPS

In MBGD, the default ortholog table was constructed using the set of genomes containing one genome from each genus, according to the NCBI Taxonomy database (11). In addition to the default ortholog table, MBGD now provides an ortholog table for each major taxon that contains at least six representative genomes that are selected one from each taxon in a given rank defined under the target taxon (Table 1).

Table 1. Pre-calculated ortholog tables in MBGD

Target rank	Rank for representative selection	CoreAlign is available	No. of taxa[a]
All (default)	Genus	No	1
Superkingdom	Genus	No	2
Phylum	Genus	No	13
Class	Genus	No	18
Order	Species	No	50
Family	Species	Yes	58
Genus	Species	Yes	31
Species	Genome	Yes	21

[a]The number of taxa in which the number of representative genomes ≥6 in MBGD release 2012–01.

Figure 1 shows the 'Ortholog Cluster Overview' page that can be shown by clicking the 'Ortholog Table' menu item from the top page. The page contains several views for summarizing the clustering result including a histogram of cluster size ('Cluster size'), a bar graph showing the relationship between occurrence pattern (representing presence or absence of the orthologs in each genome/taxon) and functional category of each orthologous group ('Occurrence pattern'), a similarity matrix for pairwise genome comparisons from which one can invoke the CGAT program (12) to display dotplot between any pair of genomes ('Pairwise comparison') (13), and a diagram showing the syntenically conserved core structure (CoreAligner; see the next section). Users can change the target taxonomy group for the current ortholog analysis by clicking an appropriate taxonomy node in the taxonomy tree in the left-hand panel.

The 'Occurrence pattern' display is the original summarization form of the entire clustering result (7), but the length of an occurrence pattern can be quite long when the number of organisms to compare is large, so that it became not effective to display in a limited space. To display such a large occurrence pattern effectively, MBGD now provides a compressed form of occurrence pattern using taxonomy information: an original occurrence pattern is first converted into a real-valued vector, where each element corresponds to a taxon in a specified taxonomic rank (rather than genome) that has a value between 0 and 1 representing the ratio of the genomes that have the orthologs in that taxon, and then each value r is quantized into four characters `0`, `-`, `+` and `1`, according to $r = 0$, $0 < r < 0.5$, $0.5 \leq r < 1$, and $r = 1$, respectively (Figure 2). The extent of aggregation in the compressed pattern can be controlled by choosing a taxonomic rank for aggregation unit from the 'Group by' option menu followed by pressing the 'Redraw the map' button. The names of taxa in the occurrence pattern are shown in the table below the histogram, where users can specify a condition of presence or absence of each taxon to filter occurrence patterns to display, in addition to the previous occurrence pattern specification window, where users can choose set of genomes by specifying phenotypic properties as well as taxonomic information (14).

CORE GENOME STRUCTURES OF TAXONIC GROUPS

In prokaryotic genome evolution, homologous genes are generated not only by speciation and duplication [called orthologs and paralogs, respectively (15)], but also by horizontal transfers [called xenologs (16)], but it is generally not easy to distinguish xenologs from orthologs or paralogs using sequence similarity information without exact knowledge of species tree, which itself is difficult to obtain for prokaryotic taxa. On the other hand, for closely related genome comparisons, syntenic conservation can be used to identify true orthologs. We previously developed a method, named CoreAligner, to identify conserved core genome structure among related organisms using gene order conservation among orthologs (8).

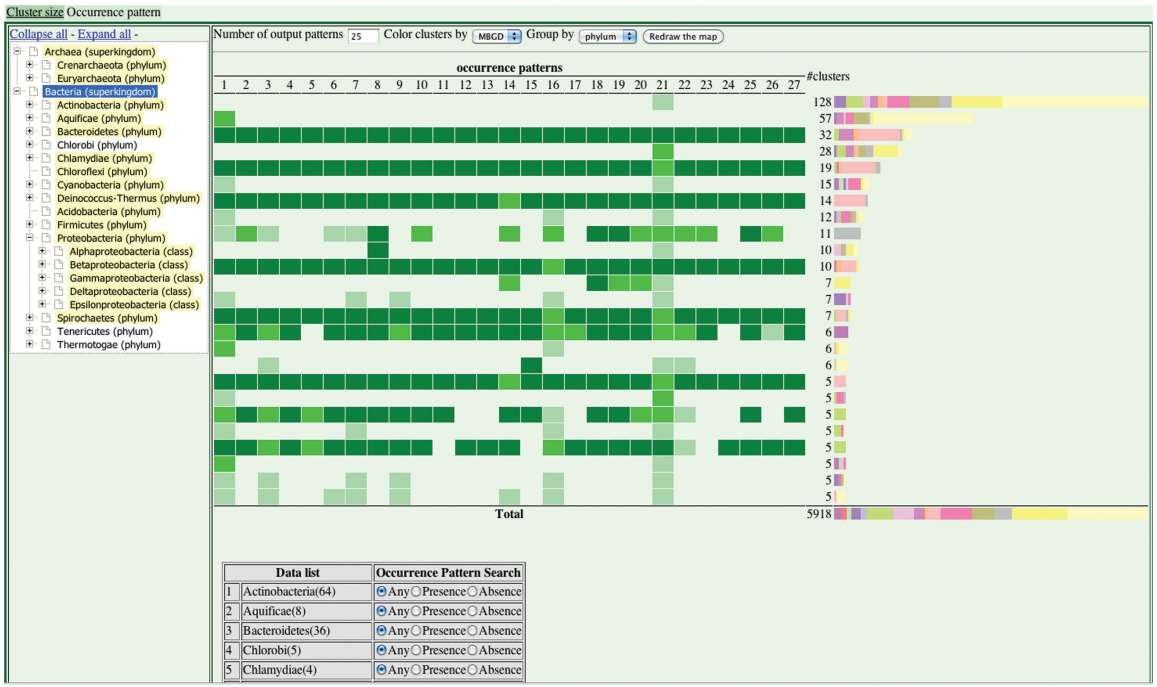

Figure 1. Occurrence pattern display with the reduced pattern representation. Displayed is a histogram of occurrence patterns (with color in histogram bar representing functional category) in the entire orthologous cluster table of the superkingdom Bacteria (highlighted in the left-hand taxonomy tree). Here, occurrence patterns are compressed in the phylum level (as selected in 'Group by' menu), meaning that each element in an occurrence pattern corresponds to a phylum; the corresponding phylum name in each element is shown in the table below the histogram, where users can specify a condition of presence or absence of each taxon to filter the pattern. The colors in the occurrence patterns represents: dark green, $r = 1$ (1); green, $0.5 \leq r < 1$ (+); light green, $0 < r < 0.5$ (–); none, $r = 0$ (0); where r is the ratio of the genomes that have the orthologs in that taxon and the characters in parentheses (1,+,–,0) are the quantized characters for reduced pattern.

CoreAligner adopts a relaxed conservation criterion in exchange for considering syntenic conservation, which allows CoreAligner to collect generally more genes than the universal core genes that is defined as a set of genes conserved in all the genomes in the given taxonomic group (data not shown).

For each ortholog table of a taxonomic group whose rank is family or below, we applied the CoreAligner program to define the core structure, which can be seen by clicking the CoreAligner tab in the ortholog cluster overview page (Figure 2). This display represents a kind of a gene-by-gene genome alignment, where each column corresponds to the extracted core orthologous groups that are ordered left to right according to the consensus order determined by CoreAligner; genes in each genome are represented by dots and neighborhood relations between genes are represented by lines connecting them. Users can scale up/down the diagram and see the information of each orthologous group by clicking the diagram.

EXTENDED ORTHOLOG TABLE GENERATED BY INCREMENTAL PROCEDURE

As already mentioned, the default ortholog table in MBGD is created using the default set of organisms that contains a representative genome from each genus. MBGD also holds the extended ortholog table that covers all the organisms in the database, which is created by adding genes of unselected genomes into appropriate ortholog group in the default ortholog table (14). We now use a new incremental procedure, named MergeTree (I. Uchiyama, unpublished program), to create the extended ortholog table. In contrast to the previous simpler version that only assigns each unclassified gene to an orthologous group, MergeTree is an incremental version of the DomClust program that performs hierarchical clustering to create orthologous groups at the domain level; that is, MergeTree classifies each unclassified gene into appropriate ortholog group at the domain level (i.e. it splits genes into domain if required) and constructs hierarchical clustering trees containing newly assigned genes. Thus the extended ortholog table is now represented in essentially the same form with other ortholog tables generated by the DomClust program.

CONCLUSION AND FUTURE DIRECTIONS

From the perspective of our database construction, continuous increase in the size of the database poses two problems: efficient manipulation of data (for computers) and effective representation of data (for humans). To cope with these problems, MBGD has adopted the strategy to consider a reduced subset of genomes, i.e. it allows users to choose an appropriate subset of genomes to compare (7), where taxonomic information can be used to choose a reduced but still useful subset. Though it can work for

Figure 2. Core genome alignment display showing the core structure of the family Bacillaceae. In this diagram, each column corresponds to an extracted 'core' orthologous group and each row corresponds to genome. A dot represents an orthologous gene and a line represents a neighborhood relation between genes in each genome. Red line indicates that there is an inversion and green line indicates that there are some inserted genes between them. Note that core alignments are available only for taxonomic groups whose ranks are family or below that contain sufficiently diverse genomes, which are highlighted yellow in the taxonomy tree viewer in the left-hand panel.

various kinds of actual genome projects, it may not so efficient to explore the entire microbial diversity. Here, we advanced our approach by preparing pre-calculated ortholog tables for several major taxonomic groups, which allows users not only to see a particular taxonomic group that they are just interested in, but also to see and compare many taxonomic groups that they might be interested in. Taxonomic information was also used to make a reduced representation of phylogenetic patterns, which can summarize large orthologous table efficiently.

Given the large diversity of the microbial world, the number of microbial genome data should still continue to increase. We will continue to improve the way of handling large and diverse genomic data along with this direction. Since an ortholog database is important reference for processing and annotating a large number of newly determined genomic and metagenomic data, quality of ortholog grouping is also important. Thus we are also trying to improve quality of ortholog classification to create a refined reference set of orthologous groups.

ACKNOWLEDGEMENTS

Computational resources were provided by the Data Integration and Analysis Facility, National Institute for Basic Biology.

FUNDING

National Bioscience Database Center, Japan Science Technology Agency; and Grant-in-Aid for Publication of Scientific Research Results from Japan Society for the Promotion of Science (248044). Funding for open access charge: National Bioscience Database Center, Japan Science Technology Agency.

Conflict of interest statement. None declared.

REFERENCES

1. Uchiyama,I. (2006) Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. *Nucleic Acids Res.*, **34**, 647–658.

2. Davidsen,T., Beck,E., Ganapathy,A., Montgomery,R., Zafar,N., Yang,Q., Madupu,R., Goetz,P., Galinsky,K., White,O. *et al.* (2010) The comprehensive microbial resource. *Nucleic Acids Res.*, **38**, D340–D345.
3. Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Friedland,G.D., Huang,K.H., Keller,K., Novichkov,P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.
4. Markowitz,V.M., Chen,I.M., Palaniappan,K., Chu,K., Szeto,E., Grechkin,Y., Ratner,A., Jacob,B., Huang,J., Williams,P. *et al.* (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.*, **40**, D115–D122.
5. Powell,S., Szklarczyk,D., Trachana,K., Roth,A., Kuhn,M., Muller,J., Arnold,R., Rattei,T., Letunic,I., Doerks,T. *et al.* (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.
6. Altenhoff,A.M., Schneider,A., Gonnet,G.H. and Dessimoz,C. (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.*, **39**, D289–D294.
7. Uchiyama,I. (2003) MBGD: microbial genome database for comparative analysis. *Nucleic Acids Res.*, **31**, 58–62.
8. Uchiyama,I. (2008) Multiple genome alignment for identifying the core structure among moderately related microbial genomes. *BMC Genomics*, **9**, 515.
9. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
10. Kosuge,T., Abe,T., Okido,T., Tanaka,N., Hirahata,M., Maruyama,Y., Mashima,J., Tomiki,A., Kurokawa,M., Himeno,R. *et al.* (2006) Exploration and grading of possible genes from 183 bacterial strains by a common protocol to identification of new genes: Gene Trek in Prokaryote Space (GTPS). *DNA Res.*, **13**, 245–254.
11. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
12. Uchiyama,I., Higuchi,T. and Kobayashi,I. (2006) CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes. *BMC Bioinformatics*, **7**, 472.
13. Uchiyama,I. (2007) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res.*, **35**, D343–D346.
14. Uchiyama,I., Higuchi,T. and Kawai,M. (2010) MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Res.*, **38**, D361–D365.
15. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
16. Gray,G.S. and Fitch,W.M. (1983) Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. *Mol. Biol. Evol.*, **1**, 57–66.