User’s guide for MAPLE version 2.3.1

Data submission

In cases of individual organisms, queries must be submitted as amino acid sequences in multi-FASTA format with unique IDs in the comment lines, which must not include tab delimiters (Figure 1). To submit metagenomic data, query sequences must be translated amino acid sequences from metagenomic sequences generated by high-throughput DNA sequencers, such as the Illumina MiSeq and Illumina HiSeq systems. Moreover, the query sequence does not necessarily have to be a complete gene, but amino acid sequences longer
than 100 residues are recommended for accurate KO assignment [1-3]. The file size of the query sequences may not be over 3 million sequences due to the limitations of computational resources; an error message will be displayed when the file size exceeded this limit. Although datasets containing 3 million sequences are enough for accurate evaluation of the MCR, particularly for determining the abundance of completed modules when more sequence data are required for analysis, the user can submit several sub-datasets having less than 3 million sequences derived from the same metagenomic sample and then merge the results from all sub-datasets into one dataset by clicking the “Merge” button on the job list page. But it will take long time to merge several jobs with 3 million sequences. Actually, we can sufficiently elucidate the overall trends of the functionome harbored in the metagenome, even when using 1 million sequences [4, 5]. The user can select homology search program, conventional BLAST or GHOSTX, which is much faster than BLAST. Because it has been verified that no substantial differences in the MAPLE results between BLAST and GHOSTX occurred, GHOSTX is recommended for the metagenomic analysis with massive data [6].

After submission of a data set, a URL address for accessing the results will be displayed along with the job ID. The results will also be displayed on the current page upon completion of the job. Anywhere from a few hours to around 15 hours will be required to complete the job when GHOSTX was selected as homology search program, depending on the size and content of the query sequences; therefore, it is recommended to bookmark the URL for later access. For example, it takes approximately 30 min to complete the job using the BBH method for the Haemophilus influenzae genome (T00001) containing 1,610 protein sequences when GHOSTX was selected (BLAST: Ca. 1 h). In case of metagenomic sequences containing 3,000,000
sequence reads, it takes about 10-15 hours to complete the job using GHOSTX. An e-mail address can be specified to receive a message containing the URL from the system upon completion of the job.

Result pages

To access the tables containing the results of completion ratios for all types of KEGG modules (pathway, structural complex, functional set, and signature modules), the user clicks on the URL address displayed on the submission page; alternatively, the user is notified by e-mail and then clicks on the job ID. The user can view detailed information of the mapping results of each KEGG module by clicking on the module ID in the table. In addition, the user can access an overview of the MCR results by clicking on the “histogram (PNG)” button on the results page and accessing the mapping results of the KEGG module by placing the cursor on the module name (Figure 3).

Additional information for each module, such as taxonomy, class, and definition, is also
displayed on the results page (Figure 3). Taxonomy is defined as the biological classification based on the MCR patterns of reference organisms with the determined genomic sequences. For example, if a module contains more than four prokaryotic species (Bacteria and Archaea) belonging different phyla, the module is represented as a prokaryotic taxonomy. Similarly, if a module contains only species belonging to *Proteobacteria*, the taxonomy of the module is *Proteobacteria*. Class indicates the module type based on the MCR patterns of reference organisms, as defined in a previous study [1]. The distribution of MCRs in 1288 prokaryotic or 29 eukaryotic species (one genome per species) can be categorized into four patterns (universal, restricted, diversified, and nonprokaryotic/noneukaryotic) regardless of the module type (pathway, structural complex, signature, or functional set).

A Boolean algebra-like equation is defined by KEGG for each module, and MCRs are calculated based on this equation. Since KOs are often assigned to two or more modules, all module IDs that share each KO constructing the module are listed together with the pathway IDs containing the KO. For example, K01623, one of the members in M00167 (reductive pentose phosphate cycle), is shared in M00001 (glycolysis) and M00003 (gluconeogenesis) and is used in seven pathway maps (Figure 1). In the case of metagenomic sequences, taxonomic information for KO-assigned genes is displayed, and the details of the phylum or class level for every KO, which facilitate the classification of organisms contributing to the completion of each functional module, are listed. For example, the module for prokaryotic ribosomes (M90000), comprising 21 bacterial, 27 archaeal, and 31 common ribosomal proteins between bacteria and archaea (Figure 4), can be used to represent the taxonomic breakdown of prokaryotes in the metagenome instead of the 16S rRNA gene, whose copy number varies depending on the prokaryotic species, because most ribosomal proteins are single-copy genes in the genome, and there are only minor variations in length among each orthologous group.

The results are removed from the server one month after the job is completed; however, the user can download the MAPLE results from all results by clicking the button “MAPLE results” (Figure 1). The user can browse all MAPLE data by re-uploading previously
downloaded data from the top page. In addition, the user also can easily download the data in an Excel format for KO assignment by KAAS, calculation of the MCR, determination of the abundances of KO-assigned genes and completed modules, analysis of module significance results, and determination of the taxonomic information of the KO-assigned genes mapped to each module from the results page (Figure 5).
Comparison of results

Users can compare results not only between their own jobs but also between job(s) and KEGG annotated genomes by clicking the “MAPLE job comparison” button on the top of the page and then inputting an e-mail address to access the job list. To conduct comparative analysis, the “Comparison” button can be selected after checking the job IDs to be compared. When several IDs of jobs to be compared are checked in the job list and the “Comparison” button is pressed, the job arrangement page will be displayed. The user can add pre-analyzed individual organisms from the organism list if necessary and change the display order. The comparative results of MCR values and mapping patterns for each KEGG module are displayed side by side in parallel (Figure 6). Detailed information for each KEGG module and taxonomy of the KO-assigned genes mapped to the module are displayed, and the MCR and taxonomy results of the KO-assigned genes can be downloaded, as previously mentioned.

Comparisons between KEGG annotated genomes, excluding the user’s jobs, are also possible. The user can directly access the comparison page without submitting an e-mail address using the “MAPLE genome comparison” button on the top of the page.

Figure 6. Concept of module abundance. The module M00529 comprising 4 reaction components is defined for denitrification. In each K number set, vertically connected K numbers indicate a complex whereas horizontally located K numbers indicate alternatives. Small numbers at the lower right of the boxes show abundance of the KO (KEGG Orthology) assigned genes normalized by mean sizes of the genes categorized in each orthologous group (12C). In the case of complex (1 and 3), minimum number is defined as an abundance of complex. When there are alternatives in the reaction component (1, 2 and 3), the sum of both abundance is defined as the abundance of each reaction component. Finally, maximum abundance in the 4 reaction components is defined as the module abundance.
MAPLE results, such as MCR, $Q$-value, and module abundance, can be easily downloaded as an Excel file. Drawing histograms using these files is laborious because there are over 700 functional modules. Accordingly, we developed a program to draw histograms of...
MAPLE results (Figure 7). Users can easily create histograms by importing the Excel file containing their MAPLE results. A histogram for comparative analysis of 4 GOS data drawn by the MAPLE Graph Maker is shown in Figure 8 [6]. Now, we provide MAPLE version 2.3.1, which is a substantially faster (more than 10 times), more user friendly, and more useful system for metagenomic analysis relative to its predecessors (version 2.3.0). MAPLE Graph Maker can be downloaded from “Software download” in the top page of MAPLE and also can be obtain from http://itsl.jp/mgm/MapleGraphMaker.zip.

Reference


