Title

Using the Metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes

Source

Elizabeth M. Glass, Argonne National Laboratory and University of Chicago;

Jared Wilkening, , Argonne National Laboratory and University of Chicago;

Andreas Wilke, Argonne National Laboratory and University of Chicago;

Dionysios Antonopoulos, Argonne National Laboratory and University of Chicago;

Folker Meyer, Argonne National Laboratory and University of Chicago;

Introduction

Shotgun metagenomics creates millions of fragments of short DNA reads, meaningless unless analyzed appropriately. The Metagenomics RAST server (MG-RAST) is a web-based, open source system that offers a unique suite of tools for analyzing these data sets. After dereplication and quality control, fragments are mapped against a comprehensive nonredundant database (NR). A phylogenetic reconstruction and a metabolic reconstruction are computed from the set of hits against the NR. The resulting data is made available for browsing, download, and—most important—comparison against a comprehensive collection of public metagenomes. A submitted metagenome is visible only to the user, unless the user makes it public or shares with other registered users. Public metagenomes are available to all.

Related Information

MG-RAST can be accessed at http://metagenomics.anl.gov.

Materials

Data

- Shotgun sequence data (e.g., 454 or Solexa)

Equipment

- Computer (Internet-connected)

Method

Registration and Data Submission

1. **Registration** is required in order to submit data. A registration link can be found on the MG-RAST home page and requires that a first and last name, preferred login name, email, and country be provided (Figure 1). If data sets are to be shared with select users, then a group and name can be requested here.

   Upon submission of an account request, a confirmation email will be sent to the email address provided in the registration form. Upon confirmation, login and submission sequence files for analysis to MG-RAST is possible.

2. Submitting a metagenome to MG-RAST. First, **uploading** the sequence file to the MG-RAST server (Figure 2). A variety of sequencing technologies are currently supported. MG-RAST is able to handle 454 GS-20, GS-FLX, TITANIUM, and Solexa 76bpx2 reads. Processing starts once the job has been uploaded (Figure 2). A fasta file containing just the nucleotide sequences can be uploaded. In this case the file name should end in .fa, .fasta, .fas, .fsa, or .fna.

   - If the sequence file is larger than 30 MB, the file should be compressed (tar and gzip); the compressed file should end in '.tgz'. For other options, contact user support (email: mg-rast@mcs.anl.gov).

   - If the project has resulted in more than one distinct metagenome data set, include all sequence files in one compressed archive file.

   - For each FASTA file, a corresponding file of data quality information* in a single compressed file may be included. To do so, both files must be compressed into a single archive, and then uploaded. The quality file name should end in .qual; and the archive name should end in .tgz. Files encoded as html, pdf, rtf, doc, docx, embl, gff3, or gtf are not supported.

     Generally, systems and tools use quality files to improve the accuracy of assembly and of the consensus sequence. Currently MG-RAST is not using quality files for generating analysis, but it archives them for future use.

     ** The format of the .qual file is similar to that of the corresponding FASTA file. For each read there should appear a header line identical (except possibly for the "description" field) to that in the FASTA file. One or more lines giving the qualities for each base follow this. Quality values should be integers between 0 and 99 and should be separated by spaces*

   The second and third steps in job submission are to **provide a job name and metadata** describing the sample. Metadata includes information outlined in the Genome Standards Consortium's (GSC) Minimum Information about a Metagenome Sequence (MIMS - http://gensc.org/gc_wiki/index.php/MIGS/MIMS). Besides filling in this information online, the metadata can also be uploaded. Metagenomes submitted to MG-RAST remain private unless the user selects to publish or share with selected registered users.

   The "Upload summary" tab contains information regarding the status of the upload. Clicking on "Finish the upload" will submit the job in the queue. Email is sent regarding job completion or any problems with the data or submission.

3. After the upload is complete, the job is submitted to the queue. Once processing is completed, email notification is sent. The status of each job can be viewed in "Manage Data" from the MG-RAST home page or from the link on the Upload Summary. The time required to process a job (typically, less than a week) depends on the number and size of other jobs in the queue as well as on the size of the submitted job.

Managing Jobs

4. The "Manage Uploaded Data" link allows for tracking the progress of jobs submitted to MG-RAST. Jobs are displayed in a table with sortable and searchable columns.

    - View details of one job: An overview of the progress is shown in the table as a series of colored boxes (e.g., grey = not started, blue = queued for computation, yellow = in progress, green = successfully completed, red = failed).
    - Access completed job or results: There are two points to access data.
        - From the MG-RAST home page, completed jobs are displayed in a list behind the public data in the green tab labeled "Private Metagenomes." This tab is not displayed if the user is not logged in.
        - From the job details page (accessed via the jobs table in "manage your uploaded data"), view the annotated data or download.
        - Link to the metagenome viewer (discussed below). This link is available only upon completion of processing.
        - Download results in GenBank format.
    - Share an annotated (meta)genome with one or more other users:
        - A user can share a job with others by clicking the link and adding the email addresses of registered users to those the user would like to grant access.
        - To share with many people (e.g., a class, research group), the user requests a new group by emailing mg-rast @ mcs.anl.gov. Group memberships may be viewed from the account management page, which is accessed by clicking on the icon of a pair of people at the far right of the green menu bar. This is also where users can change their password, if needed.
    - Delete job: Under "view details" link in the jobs table, the green menu bar in the header of the page provides an option labeled with the job number. The only action to choose from this menu is "Delete this job."

Viewing and Analyzing Results

5. After selecting a private or public genome (or from the Jobs Overview, select "Browse annotated metagenome in Seed-Viewer"), the Metagenome Overview is displayed (Figure 3). This page shows a summary of information regarding the submitted data. The first table shows the input description and lists how many sequences along with maximum, minimum, and average sequence lengths. The second table provides a broad phylogenetic overview that lists how many sequences are classified as Archaea, Bacteria, Eukaryota, or Other on the basis of protein and rRNA sequence analysis. The lengths and %G+C content of the sequences are displayed in histograms. A paragraph of automatically generated text summarizes the data is also provided on this page.

    How are Overview statistics calculated?

- Total number of sequences: The total number of submitted sequences a given metagenome. It is possible and indeed probable that some sequences cannot be matched to anything in our database.

- Total sequence size: The sum of the lengths (bp) of all submitted sequences.

- Average sequence length: The total sequence length divided by the number of sequences

- Longest sequence length: The length (bp) of the longest sequence submitted.

- Shortest sequence length: The length of the shortest sequence submitted.

6. **Reconstructions of the community and community metabolism** can be obtained via the **Sequence profile** page. This page is available via the menu entry "Metagenome" → "Sequence profile" (Figure 4).

   One of the most important aspects of the MG-RAST user interface is its ability to support projects with widely different parameters. This is achieved by allowing the dynamic adjustment of parameters for sequence matching. MG-RAST has a fast interface to modify 4 different parameters to determine the quality of the results. Figure 5 shows the dialog box for parameter adjustment. After selecting a parameter set appropriate to the task, force an update of the display "re-compute results." (Please note this will affect only the display on the selected page.)

7. The system computes a reconstruction of the **set of organisms present** in the sample and makes it available via the "Phylogenetic profile" option. This is available via the menu entry "Metagenome" → "Sequence profile." The page shows the results using a default setting. Since every sample is different, the web interface allows the user to vary parameters and observe the effects of these changes on the profile. A 97% similarity is recommended when using the **ribosomal RNA (rRNA) databases** as well as a **50 bp overlap**; for protein databases, an e-value cut-off of 10^{-5}. If more stringent parameter settings are required, requiring a minimum alignment length and percent identity is a useful addition or alternative to the e-value cut-off. Adjusting these parameters varies the resulting sets of predicted organisms in the profile.

   A note on the rRNA databases: currently MG-RAST uses RDP (http://rdp.cme.msu.edu/index.jsp), Silva (http://www.arb-silva.de/) and GREENGENES (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi). These comprehensive databases each differ somewhat in domain coverage (Archaea, Bacteria, Eukaryota) as well as in taxonomic classifications. MG-RAST does not recommend one database over another but provides the user choices.

8. The reconstruction of the **community metabolism** follows the same model as the organism mapping. Navigate to "Metagenome" → "Sequence profile" and select "Metabolic profile". Again, the ability to adjust parameters for the underlying sequence comparison is available.

   Profile results are presented in two ways: pie chart and table. Phylogeny and Metabolism are hierarchical, and the pie charts reflect that configuration. By clicking on a section of the pie chart, an additional chart appears detailing the breakdown of that group. The numbers shown in the chart and table are actual counts.

9. In addition to browsing the predicted metabolic profile, searching for particular

functional roles is possible. The search page also provides basic statistics on the number of fragments that corresponded with given genes and those that have been assigned a functional role. By typing keywords into the text box, the list is narrowed down.

10. MG-RAST enables various comparative analyses.

- ***Compare Metagenome to Other Metagenomes – Heat Maps***

MG-RAST heat maps allow for the comparison of the metabolism or the phylogeny of a metagenome with one more other metagenomes (Figure 8). For metabolic reconstructions the Subsystem dataset is available. For phylogeny, the ribosomal databases and the SEED protein database are all options. Parameters are also changeable (e-value, percent identity, and minimum alignment length) and allow analyses to be refined to suit the sequence characteristics of a sample. A minimal alignment length of 50 bp is recommended for use with all rRNA databases. The Heat Maps show the relative abundance, which is calculated using the number of sequences in a subsystem/tax class as a fraction of the total number of sequences in a subsystem/dataset. This allows for correction based on the sample size.

- ***Compare Metagenome to Organism – Recruitment Plot***

The recruitment plot enables comparison of metabolic reconstructions of the metagenome to that of a bacterial genome (Figure 9). Choosing an organism predicted in the sample, the user can compare the metabolic coverage. Like most of the comparative tools in MG-RAST, the parameters of the calculated Metabolic Reconstruction (including e-value, percent identity, and minimum alignment length) can be modified.

- ***Compare Metagenome – KEGG Map***

Samples can also be viewed and analyzed using KEGG maps (Figure 10). Mapping of functional roles to KEGG maps is done using functional assignments from analysis against the SEED. These maps are hierarchical, just like the Subsystems, which allow browsing of the sample on various levels or compare it with other metagenomes.

A note on KEGG and SEED: KEGG (Kyoto Encyclopedia of Genes and Genomes) connects known information on molecular interaction networks, such as pathways and complexes, information about genes and proteins generated by genome projects, and information about biochemical compounds and reactions. The SEED is based on expert creation and annotation of subsystems. Subsystems are a set of functional roles that an annotator has decided should be thought of as related. Frequently, subsystems represent the collection of functional roles that make up a metabolic pathway, a complex, or a class of proteins. The SEED also contains a library of protein families based on the collection of subsystems, as well as correspondences between genes in closely related strains.

Troubleshooting

Problem: Upload fails.

[Step 2]

Solutions:

1. File format is incorrect:

Must be nucleotide sequences in FASTA format

2. File is large (>30MB): User needs to create a .tgz file.

Example: Create the file metagenome.tar.gz from two fna files.

tar -cvzf metagenome.tar.gz seqfile1.fna seqfile2.fna

Problem: User does not see newly submitted job in the job list.

[Step 4]

Solution:

1. Newly submitted jobs can take several minutes to be displayed in the user's job queue. Please try again after 10 minutes.

2. If after 10 minutes the job does not show up, please contact support at mg-rast@mcs.anl.gov for assistance in determining whether the job has been submitted successfully.

Problem: Results of metagenome profile analysis are sparse.

[Step 6]

Solution:

1. Change selection parameters (e.g., e-value, percent identity). Just like using BLAST, increasing e-value or decreasing percent identity and/or alignment length will reduce the stringency for inclusion of sequence fragment hits to the results.

Discussion

MG-RAST makes computationally intensive analyses possible for the scientific community. With the advent of this server, a small academic or industrial team can now obtain high-quality results for dozens on data sets within days or a week and with very little effort or investment. With built-in support for multiple data sources and a back end that houses abstract data types, the MG-RAST server is stable, extensible, and freely available to all researchers.

Besides high-quality annotations and enormous compute power, MG-RAST provides automated analyses of phylogenetic context, identification of genes and protein families, and subsystem and metabolic reconstructions. The analyses are made available to the user in a short turnaround time, significantly increasing productivity.

Acknowledgments: This work was supported in part by the U.S. Department of Energy, under Contract DE-AC02-06CH11357.The well-annotated proteins, resulting from the SEED large-scale Subsystem Download annotation effort, enable MG-RAST.

References:

MG-RAST technology has been used in the following:

1. Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., Furlan, M., Desnues, C., Haynes, M., Li, L. *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629-632.
2. Dinsdale, E.A., Pantos, O., Smriga, S., Edwards, R.A., Angly, F., Wegley, L., Hatay, M., Hall, D., Brown, E., Haynes, M. *et al.* (2008) Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE*, **3**, e1584.
3. Edwards, R.A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D.M., Saar, M.O., Alexander, S., Alexander, E.C., Jr. and Rohwer, F. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, **7**, 57.

Figure Legends

Figure 1. Registration fields for obtaining a MG-RAST account.

Figure 2. Uploading a metagenome for analysis. The first step in creating a job in MG-RAST is to upload a sequence file.

Figure 3. Metagenome overview. (1) Overall statistics of the sample (see below as to how these are calculated). (2) Links to Profile and Comparative Analysis tools. (3) Statistical summary in paragraph form. (4) Summary table of taxonomic distribution based on best protein similarity to SEED and 16S-based similarity to RDP. (5) Graphical representations of sequence length and GC distributions. (6) Outline of the metagenome description and MIGS data the user submitted along with the sequence file.

Figure 4. Navigation to the Sequence Profile option. Select Metagenome and then Sequence Profile from the menu bar.

Figure 5. Adjusting sequence-matching parameters. Shown are parameters suitable for matching to the Silva ribosomal database.

Figure 6. Sequence Profiles. This figure shows a Metabolic Profile for one of the public metagenomes in MG-RAST. In addition to interactive displays, profiles have parameters that can be modified by the user. (1) Changeable parameters: select the type of profile to view as well as change the parameters to identify similarity between the sample and that of the proteins in subsystems or the RNA databases. (2) Information about both profile types and recommendations. (3) Once modifications to the profile are made, click on re-compute results. This will show the new profile based on the selections. (4) Summary of profile restrictions. (5) Summary graph of the number of sequences that were classified given the parameters chosen. (6) Results pie chart. Clicking on a group will create a second pie chart breaking down the distribution in that group. This can be iterated until the final sets of subsystems or organism names are reached. (7) Tabular view of results. Each column is searchable and sortable, and the table can be downloaded.

Figure 7. Searching by protein function in MG-RAST.

Figure 8. Heat Map comparison. Shown is a simple example of two metagnome samples being compared with regard to their metabolic profiles. (1) Changeable parameters: select the type of profile, as well as change the parameters to identify similarity between your sample and that of the proteins in subsystems or the RNA databases. (2) Browse or search for metagenomes to compare with a given metagenome. Make sure to add them to the

comparison by using the left and right arrow keys. (3) Once modifications have been made, click on re-compute results. This will show the new comparison based on the selections. (4) Venn diagram of your comparisons. Mousing over the dots provides information of what is in the union or intersections. Choosing what groups of organisms or subsystems are unique or similar to one another in the table below is possible by using the drop down menu next to the Venn diagram. (5) Summary of the metagenomes chosen. (6) Download the table. (7) Taxonomy and subsystems are hierarchical. Users can select what hierarchical level in which to view the results. All results tables are searchable and sortable.

Figure 9. Recruitment Plot and fragment details. Compare the metabolism of a metagenome with the metabolic reconstructions from bacterial genomes. Using the organisms predicted to be in a sample, view the metagenome coverage of a given bacterium. (1) Changeable parameters (e-value, percent identity, and minimum alignment length). (2) After changing the parameters, recompute the results. (3) E-value color legend. (4) Linear view of the bacterial chromosome with metagenome fragment hits above and below, colored by e-value. (5) E-value histogram provides the distribution of hits to the genome. Best hits are shown and used in the count. (6) Summary of the results. (7) Viewing the BLAST hits against the genome. (8) Downloadable results table or FASTA sequences. (9) BLAST results and alignments. Tables are searchable and sortable.

Figure 10. KEGG maps and comparisons. View the metabolic distribution using KEGG maps and hierarchy. Compare the metagenome with up to four genomes or metagenomes. Shown is the highest metabolic level in KEGG for this metagenome and two organisms that have a large number of fragments similar to given genomes. (1) Select a metabolic category. (2) Choose genome or metagenomes to compare with. (3) Click on select to view selections. (4) Pathway distribution for the metagenome. (5) After selecting two genomes to compare with, a table of results is shown. Each category is selectable to get a more refined view of the pathway or process.