Research

Analyzing Environmental Sequences with MEGAN

Steven Hwang[1]

[1]Johns Hopkins University, Advanced Academic Program: Bioinformatics program

Received on Dec. 2012

ABSTRACT

Motivation: The main goal of the emerging field of metagenomics is to better understand entire communities of microorganisms through sequencing and analysis of environmental samples. The analysis is aimed to identify microbes, their functions, relationships with other microbes, and evolutionary information. MEGAN (MEtaGenome Analyzer) was developed as a tool to answer these questions by analyzing sequencing read datasets.

1 INTRODUCTION

Two main approaches to metagenomic sequencing are (1) amplicon sequencing of a specific locus, usually 16S rRNA, and (2) random shotgun sequencing. This is followed by the taxonomic, functional, and comparative analysis of the sequencing reads. While the metagenomic data provides information to which genes are present in a sample, metatranscriptomic and metaproteomic data are equally important since it reveals the metabolic activities of the sampled community at a specific place and time. These activities could then be examined for responses to environmental factors and changes. MG-RAST is a readily useable tool to analyze new meta-omics datasets, but since it is a web server, researchers may have concerns about uploading unpublished data. Huson, et. al. developed MEGAN in 2007, the first stand-alone tool to examine the taxonomical content of an environmental sample by analyzing the sequenced reads.

MEGAN can be used on metagenomic data (DNA reads), metatranscriptomic data (RNA reads), metaproteomic data (peptide sequences), and 16S rRNA data. MEGAN's main function is to parse and analyze a comparison between sequenced reads against a referenced database, the most common database being NCBI's BLAST. MEGAN can also parse SAM formatted files as well as files from the Ribosomal Database Project (RDP) <http://rdp.cme.msu.edu/> and the SILVA rRNA database project <http://www.arb-silva.de/> websites. The first version of MEGAN only dealt with the taxonomic content of a single dataset, but MEGAN2 expanded to allow for comparative taxonomic analysis of multiple datasets. The user can open and analyze multiple datasets simultaneously because MEGAN provides a comparative representation of the different classifications.MEGAN3 was developed to include functional analysis of metagenomic data using GO ontology, which was replaced with SEED and KEGG in MEGAN4. Not only does MEGAN provide data analysis at different ranks of the NCBI taxonomy, but it also provides detailed analysis of individual reads. The goal of this paper is to explore MEGAN's functionalities, particularly the SEED and KEGG classification features, by reproducing the comparative analysis performed by Huson, et. al. (Sept 2011).

2 METHODS

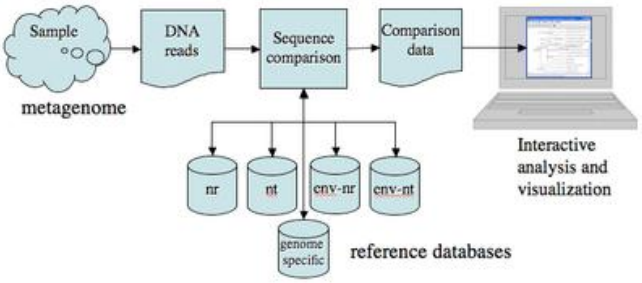

Fig. 1. The MEGAN workflow
Source: http://ab.inf.uni-tuebingen.de/software/megan/

2.1 Preparing input file for MEGAN

In order to generate an input file for MEGAN, a preprocessing step involves comparing sequenced reads with a database of reference DNA or protein sequences. Although MEGAN is not limited to a specific reference database, BLASTX comparison against the NR database requires about 10,000 CPU hours for every gigabase of sequence. If the user is only interested in a taxonomic analysis of the metagenomic data, a faster taxonomic classifier, such as Naïve Bayes Classification (NBC) tool, can be used. MEGAN is easy to use compared to other metagenomic tools since MEGAN essentially only requires a BLAST output file as input.

This paper will utilize already published datasets from a coastal ocean mesocosm study of phytoplankton. One dataset contains metagenomic data, one metatranscriptomic dataset, two metaproteomic datasets, and one 16S rRNA dataset. Read sequences from these datasets were compared against the NCBI-NR database except for the 16S rRNA dataset, which was compared against the SILVA rRNA database.

*To whom correspondence should be addressed.

2.2 MEGAN analysis: LCA algorithm

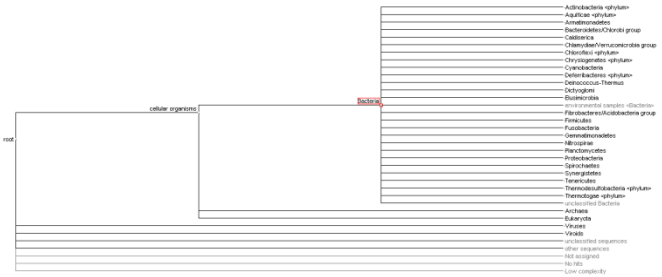

Fig. 2. MEGAN reads in the NCBI Taxonomy database, and users can navigate through the tree by collapsing/uncollapsing nodes and using the search feature.

When MEGAN starts, the current NCBI taxonomy is read in, and the top ranks of the taxonomy are displayed (Figure 2). Users are able to use MEGAN to interactively explore the 670,000+ taxa that make up NCBI taxonomy database by uncollapsing the nodes (Tree → Uncollapse) to show the next level of the taxonomy. The user can also show all nodes in the subtree below a selected node (Tree → Uncollapse subtree).

To begin analysis on the BLAST file generated from the comparison search against the BLAST database (File → Import from BLAST), we will need a RMA (read-match archive) file, which contains all reads and comparison matches in compressed form, or a RMAZ file that has these files stored separately. MEGAN will automatically parse the BLAST file and use the LCA (lowest common ancestor) algorithm to perform a taxonomic classification on the input file, and/or carry out a functional analysis using SEED.

All previous versions of MEGAN used to parse for key words in the header of a reference sequence that matches a taxon in the NCBI taxonomy to assign matches. However, MEGAN4 utilizes the new LCA algorithm to generate a species profile by assigning the sequenced reads to taxa in NCBI taxonomy database in the following manner:

(1) Compare read sequences to reference database (e.g. NCBI's NR and NT databases) using BLAST or another sequence comparison tool.

(2) Determine all matches (hits) by reads. For each read, determine a set of taxa that the read has a hit in.

(3) Using the set of taxa determined in (2), find the lowest node in the NCBI taxonomy that includes this set, and set the read to this taxon node.

Every single read is assigned to a taxon node: if the read alignment is very specific, then it will be assigned to a single node, whereas the less specific a read is (e.g. less taxa a read has a hit in), the higher up the read is assigned in the NCBI taxonomy. If the set of hits for a read are encompassed by two different taxa and one taxon is the ancestor of the other in the NCBI taxonomy, MEGAN will assign the read to the descendant since it is more specific. Thresholds can be set in MEGAN to filter reads based on the bit score of hits (Options → Change LCA Parameters → Min Score) or to filter displayed taxa based on hits by a minimal number of reads (Options → Change LCA Parameters → Min Support), which would prevent reads with a set containing conflicting hits (e.g. hits for taxa on different subtrees) from being assigned to the root node of the NCBI taxonomy. The default threshold requires a minimum of five reads to hit a taxon in order to MEGAN to determine it present in the environmental sample. If a read is assigned to a taxon that is later deemed not present due to not surpassing the threshold, the read will be assigned to a node higher in the NCBI taxonomy that has met the threshold requirement. Another threshold can be set on the score of the best hit to filter reads with low scoring reads. At the end, the result of the LCA algorithm is displayed as a rooted tree with nodes representing different taxa. By default, the sizes of the nodes are scaled based on the number of reads assigned to that node/taxon.

2.3 MEGAN analysis: SEED analysis

With MEGAN4, users can now use SEED classification for functional analysis. SEED assigns genes to functional roles, which are grouped into subsystems based on similarities to each other. MEGAN represents these classifications as a rooted tree, where the internal nodes represent the different subsystems and the leaves (end nodes) represent the functional roles. Since different subsystems may share the same functional role, some leaves may be the same. MEGAN assigns each read to a functional based on the gene within the read that had the highest score from the BLAST comparison. MEGAN displays the number of reads assigned to each functional role. For comparative analysis, the user can map multiple datasets onto the SEED hierarchy, which also allows for generation of distance data matrices based on SEED content.

2.4 MEGAN analysis: KEGG analysis

Along with SEED, MEGAN4 integrates the Kyoto Encyclopedia for Genes and Genomes (KEGG) database for pathway analysis. MEGAN assigns each read a KEGG orthology (KO) accession number based on the best hit (of the read sequences from the BLAST comparison) which has a known KO accession number. MEGAN calculates the number of hits for each KEGG pathway, and the user can see the reads that were groups into that particular pathway. Since MEGAN allows for comparative analysis of multiple datasets, MEGAN will use different colors to show what pathways are found in a particular dataset. The user can also see the KEGG pathway by selecting the node (Options → KEGG pathway).

2.5 Comparative analysis

MEGAN shows a comparison between different datasets through a tree where the nodes are displayed as visual representations of the hits by reads, such as pie charts, bar charts, and a heat map (Figure 3), that show the number of reads assigned to that specific node/taxon. Once all datasets are read into MEGAN individually, the compare function (Options → Compare →

Select All → Apply) automatically compiles all the datasets into one integrated view as seen in Figure 3. Once the datasets are merged into a single window, the user can perform a functional analysis with SEED or carry out a KEGG analysis.

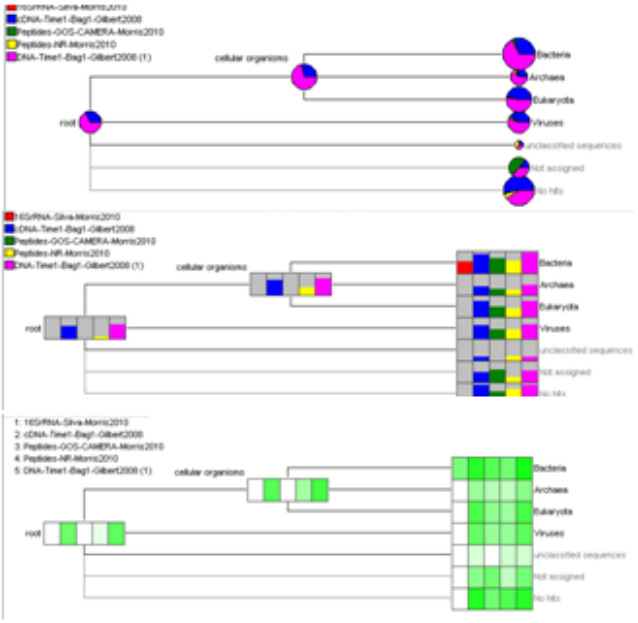

Fig. 3. Different options for node visualization. Note the size of the pie charts correlate with the number of total read sequences assigned to that particular node.

Not only can MEGAN facilitate visual comparison of the metagenomes, but it can also carry out a computational comparison of the datasets. The user can calculate a distance matrix for the datasets based on data from a taxonomic, SEED, or KEGG classification. The distances are calculated from the number of reads assigned to the selected node through a number of different methods (e.g. Goodall's ecological index, UniFrac, Euclidean). This distance matrix can then be visualized as a multidimensional scaling (MDS) plot or split network.

2.6 Other features of MEGAN

As mentioned previously, MEGAN allows users to view the sequence alignment of all reads mapped to a particular taxon. The "Alignment Viewer" (Window → Alignment Viewer) shows the multiple sequences aligned to the reference sequence, and there are multiple options provided by MEGAN to manipulate the data. For example, the user can sort the sequences by name (Options → Sort by Names) and starting position, (Options → Sort by Start), and color code base positions that align/misalign to the reference sequence (Options → Color Matches/Mismatches). MEGAN can also show insertions in read sequences (Layout → Show Insertions), nucleotides and amino acids in alignment (Layout → Show Nucleotides/Amino Acids), and the reference and consensus sequence (Layout → Show Reference/Consensus). Overall, the menu options are very intuitive and easy to use.

MEGAN also has a search option (Edit → Find) that users can use to find taxa, genes, or other strings within the NCBI taxonomy and KEGG & SEED classifications. This search feature can be used to determine first occurrence, next occurrence, and the total number of occurrences of the search key word.

3 RESULTS

MEGAN4 was used to process all five meta-omics datasets in the following manner:

3.1 Taxonomic comparison of multiple datasets

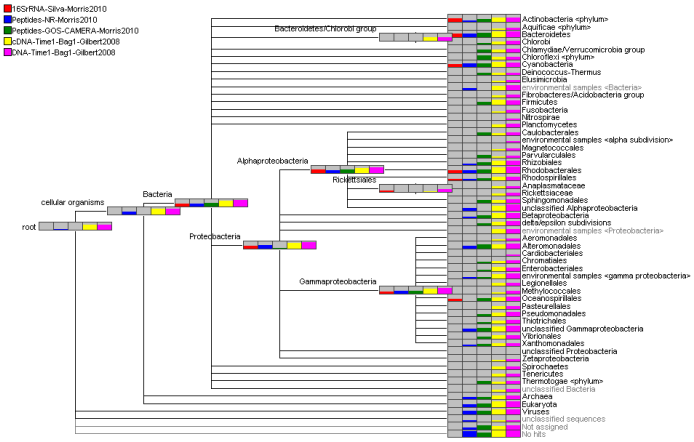

Fig. 4. Taxonomic analysis of a 16S rRNA dataset (red), two metaproteomic datasets (blue, green), one metatranscriptomic dataset (yellow), and one metagenomic dataset (pink). Each node represents a taxon, and the different bar charts at each node represent the different datasets, each bar chart representing the number of reads assigned to that taxon by the LCA algorithm.

A taxonomic comparative analysis was performed on the datasets, and, as shown in Figure 4, a legend of the datasets is provided in the upper left corner. The NCBI Taxonomy was expanded down to the rank of Order and Phylum, but we could have gone further down the NCBI Taxonomy if desired by repeatedly expanding nodes that we are interested in. Users can see the number of reads assigned to a particular node by selecting the taxon of interest as shown in Figure 5.

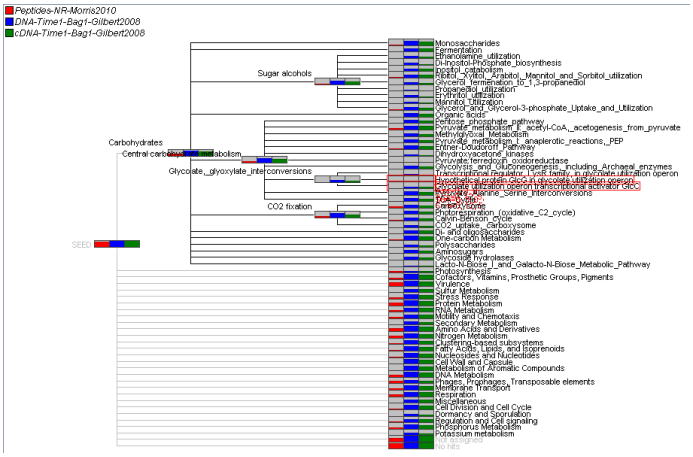

Fig. 5. Statistics for each node can be displayed by selecting the taxon of interest or hovering over with the cursor.

3.2 SEED comparison of multiple datasets

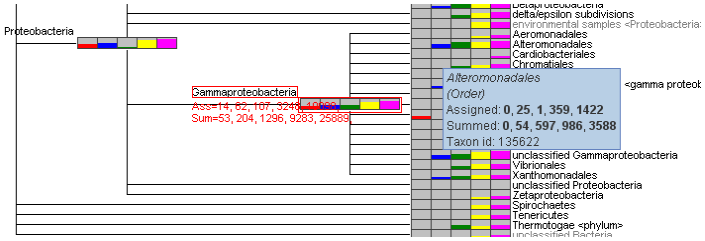

Fig. 6. SEED classification of metaproteomic (red), metatranmetagenomic (blue), and metatranscriptomic (green) data.

Figure 6 shows a SEED comparison of a metaproteomic dataset (Peptides-NR), metagenomic dataset (DNA-Time1-Bag1), and metatranscriptomic dataset (cDNA-Time1-Bag1). For each read, MEGAN4 considers all matches with bit scores over the minimum threshold (default threshold = 35 bits). The highest scoring match which has a functional role in SEED's classification is where the read will be assigned. Not only does each subsystem in SEED has one or more functional roles, but the same functional role, and thus the same read sequence, can be part of different subsystems. As explained, the bar charts represent the number of reads assigned to the particular nodes, which represent different subsystems. Compared to the other datasets, the bar charts representing the metaproteomic data does not show nearly as much representation as the other two datasets simply because the amount of read sequences is much lower.

3.3 KEGG comparison of multiple datasets

Like the SEED analysis, comparative analysis with KEGG is intuitive to perform. MEGAN first loads a file containing the KEGG classifications, followed by a file containing NCBI's RefSeq accession numbers mapped to KEGG's KO accession numbers. When the KEGG analysis is performed, all possible matches are consider for each read sequence. The highest scoring match that has a KEGG group assignment is where the read will be assigned to. There are one or more KEGG groups in each pathway, and it is not uncommon for KEGG groups to be part of more than one pathway. Figure 4 shows the KEGG functional analysis of a metaproteomic, metatranscriptomic, and metagenomic dataset. The user can view the high level analysis (Figure 7a) or use MEGAN to see the finer details provided by KEGG analysis. MEGAN4 contains a complete set of KEGG pathway files which is easily accessible through the user interface. For example, if the user if interested seeing whether a particular environmental sample contains polymerase enzymes (Figure 7b) or components of the cell cycle pathway(Figure 7c), MEGAN has a nice visual breakdown of the different functional pathways and components of interest. Each rectangle represents subunits of the different polymerases or components of the cell cycle pathway, and each color represents the number of read sequences corresponding with a particular dataset that matches to that component/pathway. As with SEED functional analysis and NCBI Taxonomy performed with MEGAN, users can use MEGAN to extract read sequences that have been assigned to a particular node for further analysis. Like SEED analysis, a read sequence can be assigned to multiple pathways.

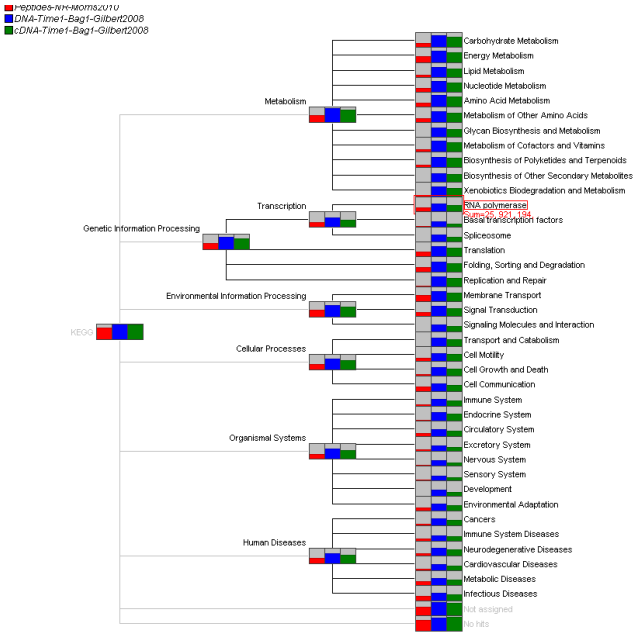

Fig. 7a. Higher-level representation of a KEGG functional analysis of metaproteomic (red), metagenomic (blue), and metatranscriptomic data (green).

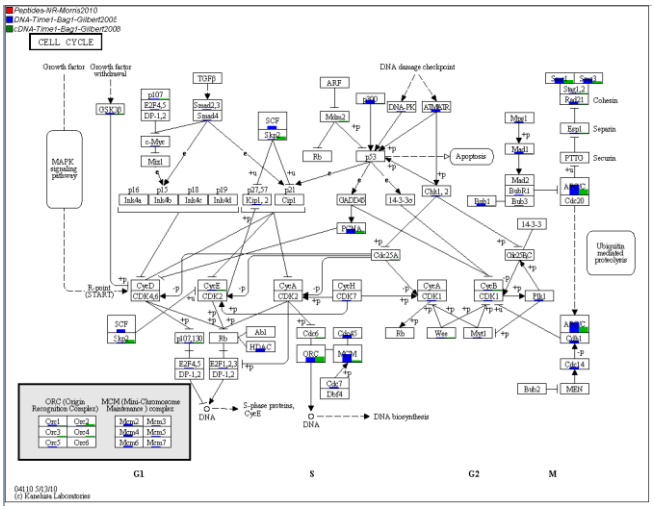

Fig. 7b. A MEGAN4 KEGG functional comparison analysis of metaproteomic (red), metagenomic (blue), and metatranscriptomic data (green), displaying the polymerase enzymes. Note the metaproteomic dataset most likely only contains bacterial microorganisms.

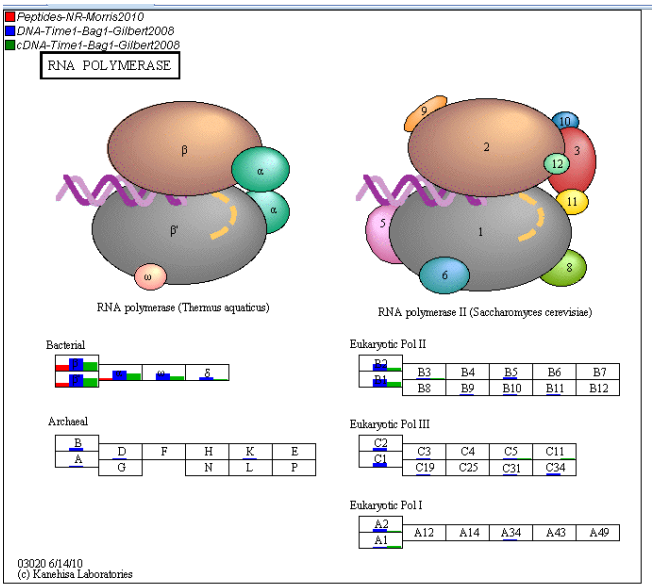

Fig. 7c. A MEGAN4 KEGG functional analysis of metaproteomic (red), metagenomic (blue), and metatranscriptomic data (green), displaying the cell cycle pathway. Note the metaproteomic dataset most likely only contains bacterial microorganisms.

DISCUSSION

The goal of the first version of MEGAN was to simply provide a taxonomical analysis of a single dataset against a referenced database. It has now developed to perform SEED and KEGG classification, including multiple datasets for comparative analysis. MEGAN's integrative features are intuitive, perform quickly, and, most importantly, its performance is comparable to MG-RAST[1]. There is a drawback to MEGAN concerning metaproteomic data: because the LCA algorithm relies on a reference database, current protein databases only contain a small fraction of information compared to the diversity of proteins; therefore, MEGAN is susceptible to reporting false negatives when analyzing metaproteomic read sequences. The bottleneck of using MEGAN is the preprocessing of comparing sequencing reads to a reference database. Alignment-free approaches for the most part are much quicker than BLAST-based analyses. One feature that I noticed to be missing or disabled was the ability to extract sequences from the detailed functional analysis (Figure 7b & 7c). In summary, MEGAN4 is an intuitive but powerful tool that utilizes SEED and KEGG classification for comparison of taxonomic profiles derived from different meta-omic data. With further development and integration of other tools, MEGAN will continue to be an invaluable tool in the growing meta-omics field.

ACKNOWLEDGEMENTS

Many thanks to Dr. Joshua Orvis for his guidance this semester. Many topics were covered the last four months, and, without his teaching and exceptional interaction with the class, it would not have been successful.

The Windows installation of MEGAN4 was used in this analysis, but it available for Linux, which requires several Java libraries (BrowserLauncher, Jama, MRJAdapter, batik, colt, h2, jcommon, and postgresql) which are found in the jars folder of the MEGAN installation directory.

All datasets use in this MEGAN4 analysis can be downloaded here:
<http://ab.inf.uni-tuebingen.de/software/megan4/megan4paper/welcome/>

REFERENCES

Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007: 17: 377–386

Huson DH, Mitra S, Ruscheweyh HJ, et al. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 2011; 21(9):1552–60.

Huson DH, Mitra S,. Introduction to the Analysis of Environmental Sequences: Metagenomics with MEGAN. Methods in Molecular Biology. 2012; 856: 415-429.

Mitra S, Stärk M, Huson DH. Analysis of 16S rRNA environmental sequences using MEGAN. BMC Genomics. 2011; 12(Suppl 3):S17. doi: 10.1186/1471-2164-12-S3-S17.