1 Methods

Community-Analyzer: A platform for visualizing and comparing microbial community structure across microbiomes

Q1 Bhusan K. Kuntal, Tarini Shankar Ghosh, Sharmila S. Mande *

Bio-Sciences R&D Division, TCS Innovation Labs, Tata Research Development & Design Centre, Tata Consultancy Service Ltd., 54-B Hadapsar Industrial Estate, Pune 411 013, Maharashtra, India

ABSTRACT

A key goal in comparative metagenomics is to identify microbial group(s) which are responsible for conferring specific characteristics to a given environment. These characteristics are the result of the inter-microbial interactions between the resident microbial groups. We present a new GUI-based comparative metagenomic analysis application called Community-Analyzer which implements a correlation-based graph layout algorithm that not only facilitates a quick visualization of the differences in the analyzed microbial communities (in terms of their taxonomic composition), but also provides insights into the inherent inter-microbial interactions occurring therein. Notably, this layout algorithm also enables grouping of the metagenomes based on the probable inter-microbial interaction patterns rather than simply comparing abundance values of various taxonomic groups. In addition, the tool implements several interactive GUI-based functionalities that enable users to perform standard comparative analyses across microbiomes. For academic and non-profit users, the Community-Analyzer is currently available for download from: http://metagenomics.atc.tcs.com/Community_Analyzer/.

© 2013 Published by Elsevier Inc.

1. Introduction

The advent of high throughput sequencing technologies and the emerging field of metagenomics have facilitated rapid extraction and sequencing of the microbial genomic content present in various environments. Several available computational methods/software [1–8] facilitate analysis of the large volume of sequence data obtained from such metagenomic studies. These developments have enabled researchers in profiling the microbial groups they possess in their environment(s) of interest. Comparing metagenomic datasets is expected to help in identifying probable key agents responsible for conferring a specific phenotypic trait to a given environment. Examples of this include traits which are specific to certain disease and physiological disorders [9–12].

Currently, a number of methodologies are being used by researchers for performing comparative metagenomic analyses. These include use of Principal Component Analysis (PCA), trees generated using various distance measures and different visualization schemes (force-directed/spring-embedded layout, stacked bar/trend plots, heat maps, etc.). Such analyses have helped researchers in making meaningful inferences. For example, PCA based approach, utilized for comparing gut microbiomes of individuals from diverse nationalities, has helped in identifying three core groups of gut microbiota, referred to as 'Enterotypes' [13]. Similarly, use of tree based methods has enabled identification of the variations in microbial communities across various seasons in the Western English Channel [14,15].

A few standalone tools/pipelines are available for comparing microbial groups across various environments [3,6,16–20]. In addition, software package libraries like QIIME have been developed that facilitate comparison of microbial communities using the differential abundance patterns in the SSU-rRNA libraries obtained from the corresponding environments [21]. While the existing tools can distinguish metagenomes based on the composition of the inhabiting organisms in these environments, they do not provide insights into the reason behind the over- or under-abundance of specific groups of organisms. To address this limitation, we have developed a standalone user-interactive comparative metagenomics analysis and visualization platform, called 'Community-Analyzer'. The platform utilizes a suite of methodologies and graphical layouts in order to facilitate visualization and comparison of microbial community structures within as well as across microbiomes. Most importantly, in contrast to the above described available analysis servers/pipelines, the present tool also facilitates visualization of probable inter-microbial interaction patterns based on the co-occurrence patterns of microbial groups across environments.

Recent studies have indicated that understanding together microbial community structures and the inter-microbial interactions can provide a more comprehensive insight into the subtle differences across microbiomes [22]. The phenotypic trait(s) of any given environment can be associated with the complex inter-microbial interactions within the consortium of microbial groups residing therein. It is therefore important to identify not only the key taxonomic group(s) (specific

* Corresponding author at: TCS Innovation Labs, Tata Research Development & Design Centre, Tata Consultancy Services Ltd., 54-B Hadapsar Industrial Estate, Pune 411 013, Maharashtra, India. Fax: +91 20 6608 6399.
E-mail address: sharmila.mande@tcs.com (S.S. Mande).

group of organisms) that may be responsible for conferring a specific trait to a given environment, but also microbial groups that regulate the action of these taxonomic groups. Insights gained from such analyses will be especially useful in metagenomic studies investigating microbial communities associated with diseases and disorders. In such studies, the identified key microbial groups and the organisms regulating them may help in devising appropriate diagnostic as well as therapeutic strategies. In order to get such insights, one needs to obtain answers to the following two questions. What are the inherent similarities and differences, in terms of microbial composition, amongst the samples under study? Can we infer the probable inter-microbial interaction patterns present in different environments? We address these two key questions by introducing a new graph based layout in Community-Analyzer. The layout has been implemented along with the standard visualization methods like PCA, force-directed/spring-embedded layouts, heat maps, bar/trend plots and distance-based trees. The implementation of a combination of visualization layout along with several standard features for comparative analyses provides an innovative way of identifying similarities/differences across various microbial communities.

We exemplify Community-Analyzer's visualization and analysis capabilities using two real metagenomic datasets (as case studies). We also provide an extensive walk-through of the functionalities of the tool. End users, without any programming knowledge, can easily perform analysis using the simple interactive options provided in Community-Analyzer.

2. Materials and methods

Community-Analyzer is available as a standalone tool for Windows and Linux platforms. Fig. 1 summarizes a typical metagenomic analysis work-flow and the application of Community-Analyzer in this workflow. In a typical metagenomic sequencing project, genomic DNA (corresponding to the resident microbial groups) from an environment are extracted. Subsequently either the entire set of genomic fragments (obtained using the shotgun sequencing techniques) or fragments of specific marker genes (e.g., 16S rDNA) are sequenced. The taxonomic affiliations of each of the sequenced fragments (referred to as reads) are then obtained using suitable binning algorithm(s) [1,2,4,7,8,23]. Based on the assignments, these algorithms also provide the abundance profile of the various taxonomic groups in the given environment. These abundance profiles (referred to as taxonomic classification summaries) from multiple environments can then be provided as inputs to Community-Analyzer for visualization and comparative analysis.

2.1. Obtaining 'Community-Analyzer layout' for a set of metagenomic samples

A majority of organisms present in any environment are hitherto unknown. These organisms may belong to new species/genus/family/order/class/phylum. Thus, sequences originating from these organisms are expected to be assigned to taxa at various taxonomic levels (e.g., genus, family, order, class, phylum, etc.). Consequently, most of the binning algorithms generate taxonomic summaries (consisting of abundance profiles of taxonomic groups) at different taxonomic levels. In order to effectively compare two or more metagenomic datasets, it is thus important to evaluate the differences amongst the metagenomes at various taxonomic levels. Based on the input data (from the taxonomic summary files), Community-Analyzer generates a graphical layout (displaying microbial community structures) at user-specified taxonomic level. The graphical layout is generated to capture and display not only the inter-microbial interaction patterns across the microbial communities under study, but also the similarities/differences across the communities. The generated graphical layout, hereafter referred to as the 'Community-Analyzer layout' (Fig. 2), displays two distinct aspects of the analyzed microbial communities. First, the resident microbial (taxonomic) groups are arranged based on the correlations of the abundance patterns of various taxa across
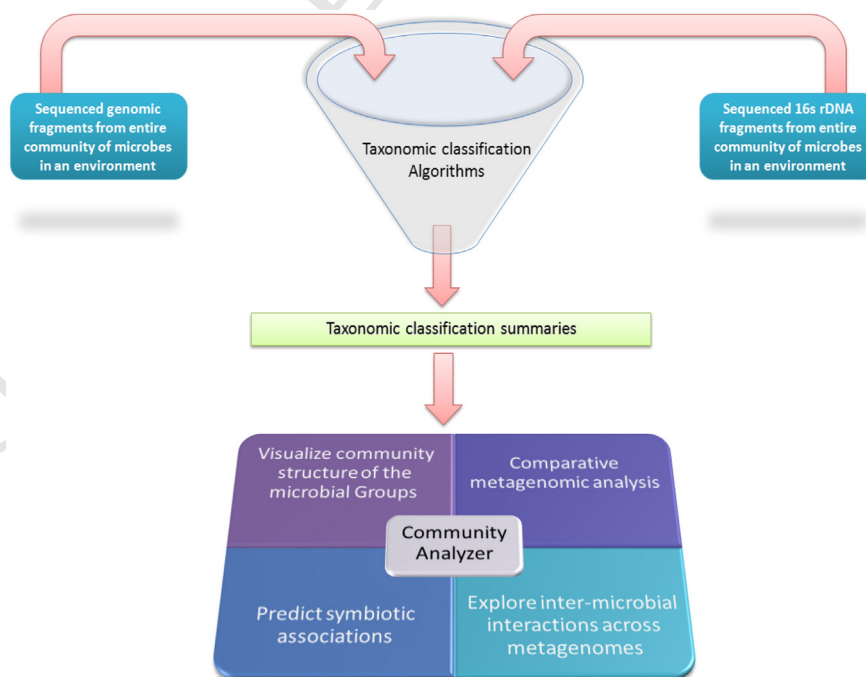
Fig. 1. Taxonomic classification summaries (obtained using representative binning algorithms) are provided as input to Community-Analyzer. In a typical metagenomic sequencing project, genomic DNA (corresponding to resident microbial groups) from an environment are extracted. This is followed by sequencing of either the entire set of genomic fragments (obtained using shotgun sequencing techniques) or the genomic fragments of specific marker genes (e.g. 16S rDNA). The taxonomic affiliations of each of these sequenced genomic fragments are then obtained using any of the available taxonomic classification methods. The obtained taxonomic classification summaries are then provided as input to the Community-Analyzer.