Journal of Biotechnology 167 (2013) 156–165

Contents lists available at SciVerse ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

MetaSAMS—A novel software platform for taxonomic classification, functional annotation and comparative analysis of metagenome datasets

Martha Zakrzewski[a,1], Thomas Bekel[b,1], Christina Ander[a,c], Alfred Pühler[d], Oliver Rupp[a], Jens Stoye[a,c], Andreas Schlüter[d,*,2], Alexander Goesmann[a,2]

[a] Institute for Bioinformatics (IfB), Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany
[b] Evonik-Degussa GmbH, R&D Bioproducts, Halle/Westfalen, Germany
[c] Genome Informatics, Faculty of Technology, Bielefeld University, Bielefeld, Germany
[d] Institute for Genome Research and Systems Biology, Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany

ARTICLE INFO

Article history:
Received 19 April 2012
Received in revised form 21 July 2012
Accepted 21 September 2012
Available online 29 September 2012

Keywords:
Metagenomics
Taxonomic classification
Functional annotation
Biogas

ABSTRACT

Metagenomics aims at exploring microbial communities concerning their composition and functioning. Application of high-throughput sequencing technologies for the analysis of environmental DNA-preparations can generate large sets of metagenome sequence data which have to be analyzed by means of bioinformatics tools to unveil the taxonomic composition of the analyzed community as well as the repertoire of genes and gene functions. A bioinformatics software platform is required that allows the automated taxonomic and functional analysis and interpretation of metagenome datasets without manual effort. To address current demands in metagenome data analyses, the novel platform MetaSAMS was developed. MetaSAMS automatically accomplishes the tasks necessary for analyzing the composition and functional repertoire of a given microbial community from metagenome sequence data by implementing two software pipelines: (i) the first pipeline consists of three different classifiers performing the taxonomic profiling of metagenome sequences and (ii) the second functional pipeline accomplishes region predictions on assembled contigs and assigns functional information to predicted coding sequences. Moreover, MetaSAMS provides tools for statistical and comparative analyses based on the taxonomic and functional annotations. The capabilities of MetaSAMS are demonstrated for two metagenome datasets obtained from a biogas-producing microbial community of a production-scale biogas plant. The MetaSAMS web interface is available at https://metasams.cebitec.uni-bielefeld.de.

1. Introduction

With the advent of new high-throughput technologies, ultrafast DNA sequencing has become a standard method for many scientific applications that produces large amounts of data in a short time and at continuously decreasing costs. In metagenomic studies, DNA is isolated directly from the environment and subsequently sequenced. In contrast to classic whole genome sequencing of cultivated organisms, metagenome sequencing also provides access to the non-culturable fraction of microbial communities. In general, five tasks are addressed within metagenome studies: (1) taxonomic profiling, (2) prediction of metabolic functions, (3) identification of full-length genes and gene variants for biotechnological applications, (4) binning of metagenome sequence data and (5) comparative analyses. In the following, currently available bioinformatic tools for completing these tasks will be summarized.

Different tools are available for taxonomic characterizations of metagenome data such as the RDP classifier (Wang et al., 2007), MEGAN (Huson et al., 2007), Sort-ITEMS (Monzoorul Haque et al., 2009) and CARMA3 (Gerlach and Stoye, 2011). The RDP classifier assigns 16S rRNA gene sequences to a taxonomic category. Since only a limited amount of metagenomic sequences encodes 16S rRNA genes, approaches based on the prediction of environmental gene tags (EGTs) are applied in addition. MEGAN classifies metagenomic reads based on the results of a sequence homology search such as BLAST (Altschul et al., 1997) against a reference database. Sequences with more than one database hit are classified employing an algorithm called Lowest Common Ancestor (LCA) assignment. Sort-ITEMS uses a reciprocal BLAST search and the LCA approach to assign reads to a taxon. CARMA3 extends the reciprocal BLAST approach. It aims at the deduction of both a taxonomic as well as a functional profile from metagenome data based on

[*] Corresponding author at: Center for Biotechnology (CeBiTec), Universität Bielefeld, D-33594 Bielefeld, Germany. Tel.: +49 0521 106 8757.
E-mail address: ASchluet@cebitec.Uni-Bielefeld.DE (A. Schlüter).
[1] T.B. and M.Z. contributed equally to this work.
[2] A.S and A.G. contributed equally to this work.

0168-1656/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jbiotec.2012.09.013

the classification of EGTs according to protein sequence or protein family data.

A further challenge in metagenome analysis is the functional characterization of a metagenome. BLAST searches against gene and protein databases are widely used for the functional interpretation of metagenomic reads. BLAST and Hidden Markov Model (HMM) based searches on metagenome data are limited by the short length of the reads produced by high-throughput sequencing technologies. These short reads often only encode protein fragments or domains. However, for biotechnological applications, full-length genes or gene clusters and their variants are of interest.

An additional task is the assignment of sequences to phylogenetic groups using binning approaches such as those described by TACOA (Diaz et al., 2009) or PhyloPythiaS (Patil et al., 2011). Finally, comparative tools for the analyses of shared or unique taxa or functionalities between different metagenomes are necessary. Because of decreasing costs, sequencing of several metagenomes from different environments and under different conditions became feasible (Tringe et al., 2005; Turnbaugh et al., 2009). Consequently, comparative analyses and visualizations are required to describe relationships between multiple metagenomes.

Due to the vast amount of data generated by next-generation sequencing technologies, availability of metagenome annotation systems is important that automate these tasks and enable the integration of novel tools. A variety of metagenome annotation systems are currently used, for example MEGAN (Huson et al., 2007), MG-RAST (Meyer et al., 2008), and IMG/M (Markowitz et al., 2008) that address some of the tasks outlined above. MEGAN is available as a standalone tool, which generates functional and taxonomic profiles for a given metagenome based on the assignment of sequences to entries of the NCBI taxonomy database. The program focuses on the visual exploration of metagenomes. It does not include any means to perform the actual BLAST searches, which are mandatory for the subsequent LCA-assignments. Therefore, this platform is inappropriate for users without access to the necessary compute infrastructure for performing such compute-intensive calculations. In contrast to MEGAN, the MG-RAST system is a platform providing the complete analysis of single and multiple metagenomes, including both the generation of taxonomic profiles based on the detection and classification of 16S rDNA fragments as well as functional analyses by means of the SEED framework (Overbeek et al., 2005). The IMG/M repository is a data management and analysis system for microbial metagenomes. It provides tools for analyzing the functional capabilities of microbial communities based on their metagenome sequences.

The SAMS (Bekel et al., 2009) platform was implemented for the analysis of whole genome shotgun sequences, expressed sequence tags and sequences obtained by ultrafast sequencing. It accomplishes extensive bioinformatics analysis and visualizes obtained results. However, SAMS features shortcomings concerning the processing of high amounts of sequence reads generated by high-throughput sequencing. Therefore, MetaSAMS was implemented as a metagenome platform by extending the data schema and tools of SAMS to allow the storage, processing and analyses of metagenomes. In this work, MetaSAMS is introduced which provides an automated taxonomic and functional annotation pipeline that accomplishes the major tasks in metagenome data analysis.

2. Methods

2.1. System design and implementation

The design of MetaSAMS was derived from the Sequence Analysis and Management System (SAMS) (Bekel et al., 2009), which was originally developed for quality control in whole genome shotgun projects and for the automated analysis of Sanger EST and cDNA data. The system design is based on a three tier architecture that embeds the database layer, the business logic layer, and the presentation layer. The data are stored using a relational database management system (RDBMS). Access to the data is implemented by using the O2DBI software (Linke, B., unpublished) that provides an automatic object relational mapping. The business layer enables access to SAMS projects by the generalized project management system (GPMS), which is commonly used in various bioinformatics software packages at the CeBiTec (Bielefeld University). SAMS is available through a web interface which is based on CGI scripts running on an Apache server.

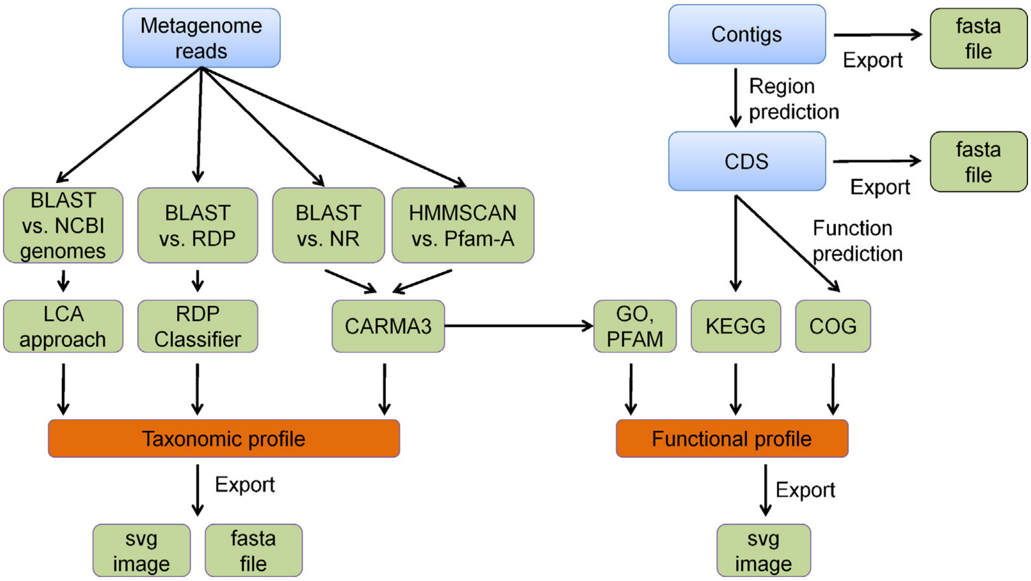

Fig. 1. A schematic overview of the MetaSAMS analysis workflow. The imported metagenome data are processed in two pipelines leading to taxonomic and functional profiles. The taxonomic profile is produced based on the results of an LCA approach of multiple BLAST hits, RDP classifier and CARMA3. The latter tool together with the functional annotation of identified coding sequences (CDS) contributes to the functional profile. Taxonomic and functional profiles are available in csv format, while the corresponding charts can be exported in svg format.