OXFORD

Systems biology

iMetaLab 1.0: a web platform for metaproteomics data analysis

Bo Liao[1], Zhibin Ning[1], Kai Cheng[1], Xu Zhang[1], Leyuan Li[1], Janice Mayne[1] and Daniel Figeys[1,2,*]

[1]Department of Biochemistry, Microbiology and Immunology, Ottawa Institute of Systems Biology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada and [2]Molecular Architecture of Life Program, Canadian Institute for Advanced Research, Toronto, Ontario, Canada

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wruen

Abstract

Summary: The human gut microbiota, a complex, dynamic and biodiverse community, has been increasingly shown to influence many aspects of health and disease. Metaproteomic analysis has proven to be a powerful approach to study the functionality of the microbiota. However, the processing and analyses of metaproteomic mass spectrometry data remains a daunting task in metaproteomics data analysis. We developed iMetaLab, a web based platform to provide a user-friendly and comprehensive data analysis pipeline with a focus on lowering the technical barrier for metaproteomics data analysis.

Availability and implementation: iMetaLab is freely available through at http://imetalab.ca.

Contact: dfigeys@uottawa.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Increasing evidence shows that the human gut microbiota plays important roles in human health (Clemente *et al.*, 2012). Metaproteomics provides invaluable functional insights to better understand microbial communities and their interactions with the host environment (Zhang *et al.*, 2016, 2017). However, metaproteomics data analysis is challenging due to the exceptionally large reference protein database [for example, ~9.9 million proteins compose the human gut microbiota (Li *et al.*, 2014)] and the requirement of retrieving the taxonomic information from identified peptides and proteins. Although a lot of software is available for conventional proteomics analysis, these have difficulties processing large reference databases and identifying taxonomic linkages. Therefore, metaproteomic analysis requires dedicated software that can achieve high sensitivity and accuracy of qualitative and quantitative information of peptide, protein and taxon from microbiota communities.

Previously we developed a software platform, named MetaLab, which automates metaproteomics data analysis (Cheng *et al.*, 2017). Without Metalab, processing of raw mass spectrometry (MS) data requires sequential steps of database construction, peptide

identification and quantification, taxonomy and function profile construction (May *et al.*, 2016; Timmins-Schiffman *et al.*, 2017). Instead, MetaLab inputs raw MS data, constructs sample specific databases and generates datasets including the abundances of microbial peptides and proteins, taxonomic compositions and functional profiling. Here we introduce and describe iMetaLab (http://imetalab.ca), a web-based version of MetaLab, with enhanced visualization tools to make metaproteomic analysis readily available to researchers interested in studying the microbiome. The iMetaLab web platform integrates both pre-configured cloud servers that host the MetaLab software, and a user-friendly front-end web interface that handles user requests. Below we describe the overall architecture, design principles and key features of iMetaLab.

2 Design and implementation

2.1 Overall design and architecture

iMetaLab is a completely cloud based platform accessed through a web interface. The platform consolidates MS raw data conversion, sample specific database construction, peptide identification and
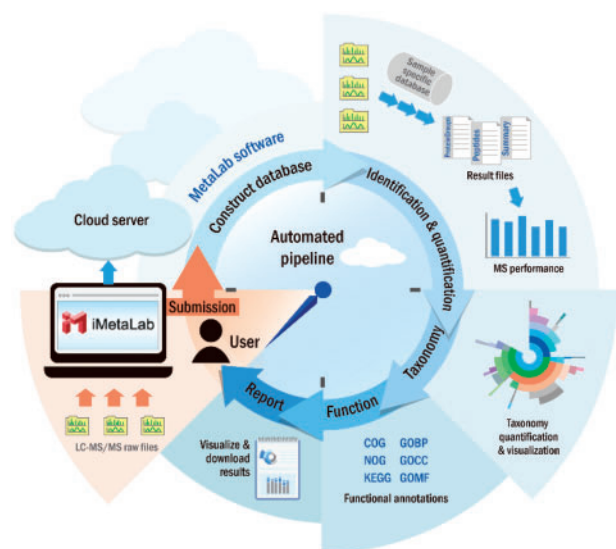
Fig. 1. The overall design and architecture of iMetaLab. iMetaLab platform consolidates MS raw data conversion, sample specific database construction, peptide identification and quantification, and taxonomy and function annotation into a single data processing pipeline. Users simply submit their MS raw files using the web interface, and iMetaLab's data processing pipeline will generate a collection of reports at the end for users to download

quantification, and taxonomy and functional profile construction into a single data processing pipeline (Fig. 1). Users simply submit MS raw files to the web interface, then through iMetaLab's data processing pipeline, data sheets containing the above information for all samples will be generated. Since iMetaLab allows the use of MetaLab without local installation, it further lowers the technical barriers for researchers and non-experts who would like to perform metaproteomic analysis. In addition, the web platform is regularly maintained and updated to the most recently released version of MetaLab. Not only can users easily submit their data processing requests, they can also easily track their submission through the web interface.

The iMetaLab platform is composed of servers that both host the MetaLab software and process client-side requests, and a front-end web interface that gathers client-side requests and displays corresponding server-side responses. The technical implementation of the platform is described in the Supplementary Section S1.

2.2 iMetaLab session and workflow

To process raw data using iMetaLab's pipeline, users first create a session using iMetaLab's web interface. For new session setup and submission, the design of the web interface is based on MetaLab's desktop graphical user interface (GUI) with modifications making it consistent with the general web surfing experience. The submission process is divided into four steps with step-by-step instruction provided in the Supplementary Section S2. Before submitting a session to the server, users can always go back to previous steps to make changes. iMetaLab considers each submission as a session giving it a unique session number.

The complete workflow of iMetaLab is composed of four major steps: sample specific database construction, peptide identification and quantification, taxonomy analysis and function annotation. Users can choose to either use the default workflow or customize their own by simply changing workflow settings when they create their sessions.

In iMetaLab, commonly used mouse gut microbiota (Xiao *et al.*, 2015) and human gut microbiota database (Li *et al.*, 2014) are

provided. To tackle large database size, iMetaLab employs a two-step searching strategy. Gene catalogues instead of the whole NCBI nr or Uniprot database are used in the two-step searching strategy. Both NCBI nr and Uniprot databases provide good sequence coverage. However, both NCBI nr and Uniprot databases are not representative, and contain a significant amount of spurious and interfering sequences. Therefore, they are oversized for the two-step searching strategy. On the other hand, gene catalogues (for both human and mouse) are close-to-complete gut microbiome databases from repetitive high-throughput genome sequencing, which are well curated and freely available. Gene catalogues outperform matched databases from genome sequencing of the same sample (Zhang *et al.*, 2017).

At step one, by default, iMetaLab first generates a delegate MS/MS spectra list from the original spectra using spectral clustering method. After searching the delegate MS/MS spectra list against the gut microbial gene catalog database, the module generates a candidate protein list, which will be used to construct a refined sample specific database. Compared to classical two-step searching approaches (Jagtap *et al.*, 2013; Zhang *et al.*, 2016), the spectra clustering strategy used in iMetaLab removes the redundant or inferior spectra, which generally account for up to 80% of the total number. This implementation significantly increases the time efficiency in the database construction step (Cheng *et al.*, 2017). Users are provided with the option to disable the spectra clustering option if they are not concerned with the processing time.

The sample specific database generated at the first step is next used for proteomic identification and quantification, which is accomplished by MaxQuant (version 1.5.3.30) (Cox *et al.*, 2008).

At the taxonomy analysis step, iMetaLab not only assigns the lowest common ancestor (LCA) to identified peptides, but also provides quantitative information of taxa, which is not yet available through other metaproteomics data analysis tools. At the last step, function annotation for each protein is assigned. The protein sequences in human and mouse gut microbiome databases were blasted against EggNOG (version 4.5.1). Users can directly obtain information about COG, NOG, KEGG, GOBP, GOCC and GOMF for each protein from iMetaLab's function annotation results.

2.3 iMetaLab workflow output

iMetaLab provides a progress tracking web interface to allow users to track their submission. The same session number used for submission is also used for tracking and accessing results. Once completed, the results become available for download. The results cover MS performance, identified unique peptides and proteins, taxonomy profile and function annotation. Both MS performance and taxonomy LCA information can be conveniently visualized using iMetaLab's within-browser quick visualization options. The taxonomy result is also available in .biom format. Both of the .csv and .biom files contain the information of phylogenetic tree, which can be used to visualize the identified taxa. The .biom file (http://biom-format.org/) is compatible with many third party tools, such as MEGAN (Huson *et al.*, 2016).

3 Conclusions and future development

iMetaLab is a cloud based platform providing a user-friendly web interface allowing general users to directly obtain peptide and taxa abundance information and protein function annotation from raw MS data. iMetaLab platform was originally released on September 13, 2017. Since then over 140 research institutions from

30 countries accessed iMetalab for metaproteomics data analysis. For iMetaLab's next design phase, we will focus on expanding its data visualization features while adding multivariate statistical and functional analysis tools to its web interface. These additions will allow researchers to more efficiently extract meaningful information from their microbiome derived, MS generated datasets using iMetaLab.

Acknowledgement

Funding

References

Cheng,K. *et al.* (2017) MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome*, 5, 157.

Clemente,J.C. *et al.* (2012) The impact of the gut microbiota on human health: an integrative view. *Cell*, 148, 1258–1270.

Cox,J. *et al.* (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, 26, 1367–1372.

Huson,D.H. *et al.* (2016) MEGAN community edition—interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.*, 12, e1004957.

Jagtap,P. *et al.* (2013) A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics*, 13, 1352–1357.

Li,J.H. *et al.* (2014) An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.*, 32, 834–841.

May,D.H. *et al.* (2016) An alignment-free "metapeptide" strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. *J. Proteome Res.*, 15, 2697–2705.

Timmins-Schiffman,E. *et al.* (2017) Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. *ISME J.*, 11, 309–314.

Xiao,L. *et al.* (2015) A catalog of the mouse gut metagenome. *Nat. Biotechnol.*, 33, 1103–1108.

Zhang,X. *et al.* (2016) MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome*, 4, 31.

Zhang,X. *et al.* (2017) Deep metaproteomics approach for the study of human microbiomes. *Anal. Chem.*, 89, 9407–9415.