Journal of Chromatography B, 966 (2014) 77–82

Contents lists available at ScienceDirect

Journal of Chromatography B

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

# Integrated enrichment analysis and pathway-centered visualization of metabolomics, proteomics, transcriptomics, and genomics data by using the InCroMAP software☆

Johannes Eichner[a,1], Lars Rosenbaum[a,*,1], Clemens Wrzodek[a], Hans-Ulrich Häring[b,c,e], Andreas Zell[a], Rainer Lehmann[b,d,e,**]

[a] Center for Bioinformatics, University of Tübingen, Tübingen, Germany
[b] Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Centre Munich at the University of Tübingen, Tübingen, Germany
[c] Division of Endocrinology, Diabetology, Vascular Medicine, Nephrology and Clinical Chemistry, Department of Internal Medicine IV, University Hospital Tübingen, Tübingen, Germany
[d] Division of Clinical Chemistry and Pathobiochemistry, Department of Internal Medicine IV, University Hospital Tübingen, Tübingen, Germany
[e] German Center for Diabetes Research (DZD), Germany

ARTICLE INFO

*Article history:*
Received 30 September 2013
Accepted 15 April 2014
Available online 25 April 2014

*Keywords:*
Metabolomics
Transcriptomics
Proteomics
Genomics
Enrichment analysis
Pathway visualization

ABSTRACT

In systems biology, the combination of multiple types of omics data, such as metabolomics, proteomics, transcriptomics, and genomics, yields more information on a biological process than the analysis of a single type of data. Thus, data from different omics platforms is usually combined in one experimental setup to obtain insight into a biological process or a disease state. Particularly high accuracy metabolomics data from modern mass spectrometry instruments is currently more and more integrated into biological studies. Reflecting this trend, we extended InCroMAP, a data integration, analysis and visualization tool for genomics, transcriptomics, and proteomics data. Now, the tool is able to perform an integrated enrichment analysis and pathway-based visualization of multi-omics data and thus, it is suitable for the evaluation of comprehensive systems biology studies.



## 1. Introduction

Today, high-throughput methods for the analysis of biological systems, such as microarrays, next generation sequencing, as well as mass spectrometric and NMR approaches generate a wealth of omics-scale data. To develop hypotheses about a biological process or a disease state, a variety of omics platforms for measuring different genomic, proteomic, and metabolomic features are combined in one experimental setup. Probably the most frequently analyzed genomic feature is messenger RNA (mRNA), which can be quantified on a genome-wide level by gene expression chips (microarrays) or next generation sequencing techniques. Other important (epi)genomic features include microRNA (miRNA), single-nucleotide polymorphisms (SNPs), and epigenetic information, such as the promoter methylation status (DNAm). Proteomic features can be derived from abundance profiling of particular protein patterns and posttranslational modifications, such as acetylation and phosphorylation. Furthermore, mass spectrometric and NMR-based metabolomics platforms allow to investigate complex metabolite patterns with high accuracy. In recent years, increasing attention has been drawn to the integration of metabolomics with molecular profiling on the transcriptional and protein level [1,2]. The sensible and integrated visualization of omics data at different levels of abstraction is crucial to obtain biological insight without being overwhelmed by the intrinsic complexity of the data [3]. A key concept for detecting alterations in cell

☆ This paper is part of the special issue "Metabolomics II" by G. Theodoridis.

* Corresponding author.
** Corresponding author at: Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Centre Munich at the University of Tübingen, Tübingen, Germany. Tel.: +49 7071 29 8 31 93.
*E-mail addresses:* johannes.eichner@uni-tuebingen.de (J. Eichner), rainer.lehmann@med.uni-tuebingen.de (R. Lehmann).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jchromb.2014.04.030
1570-0232/

signalling or metabolism in a biological system are pathway-based visualizations.

A plethora of tools were developed for the inspection of data from individual platforms (see [3] for examples). Furthermore, solutions for the combined visualization of transcriptomics (mRNA) and metabolomics data exist [4,5]. Both Paintomics [4] and MassTRIX [5] are web-service based tools that are capable of visualizing mRNA microarray and identified metabolomics data in KEGG pathways [6]. MassTRIX is able to handle unidentified metabolic data by comparing each mass against theoretical adducts stored in metabolomics databases. An example of a tool that can handle several heterogeneous types of omics data is the commercial Ingenuity Pathway Analysis software (www.ingenuity.com). However, Ingenuity does not provide an integrated visualization of multi-level omics data. In summary, the comprehensive analysis of multi-omics data is presently limited because current high-level analysis tools are not able to perform an integrated analysis of data from multiple types of omics platforms, are focused on certain specialized platforms, or not freely available. Thus, novel analysis tools and appropriate visualizations are required, since complex interactions between multiple layers of a biological system can only be inferred by the integration of omics data across multiple platforms.

In this contribution, we present an extended version of the tool InCroMAP (Integrated analysis of Cross-platform MicroArray and Pathway data) [7,8]. InCroMAP is a stand-alone Java software originally developed for the enrichment analysis and pathway-based visualizations of genomic and proteomic data, where multiple biological layers were monitored in the same set of samples. The application was specifically designed to provide a high ease of use for investigators of all kinds of analytical disciplines, i.e. detailed experiences in bioinformatics are not necessary. Consequently, all information required, for example, for the mapping between different metabolite identifiers or the annotation of miRNAs with mRNA targets, is either directly included in the tool or dynamically downloaded in the background. Previous versions of InCroMAP focused on genomics and transcriptomics data and were designed for the combined analysis of a wide variety of omics platform types, where each feature can be linked to a gene or genomic interval. Up to the present, metabolomics data was not supported. In the extended version of InCroMAP presented here, we support the enrichment analysis and pathway-based visualization of annotated metabolomics data and thereby complemented the tool for comprehensive systems biology data evaluation. For convenience, InCroMAP supports several commonly used metabolite identifiers, such as specific database identifiers (e.g., HMDB [9]), common synonyms, and InChIKeys.

InCroMAP is freely available under the LGPL3 license at http://www.cogsys.cs.uni-tuebingen.de/software/InCroMAP, including a comprehensive users guide and several example data files to test the capabilities of the tool.

## 2. Methods

Before systems biology data from heterogeneous platforms can be integrated and visualized with high-level data analysis tools like InCroMAP, the data has to be preprocessed in a platform-dependent manner (see Fig. 1). A typical workflow for metabolomics data from nontargeted mass spectrometric measurements includes the following steps. First, metabolite features, which are characterized by mass and relative intensity, are extracted from the raw data files by applying software tools, like our recently developed FeatureFinder software [10] or commercial tools implemented in the software of the instrument. Ideally, the feature induced by a metabolite should contain all signals that were generated by the metabolite (including isotopic peaks) [10]. Second, the extracted features have to be aligned and linked between the different samples. The aforementioned preprocessing steps can be performed with sophisticated open-source software tools like OpenMS [11]. The features are then subjected to quality controls and low-level statistical data analysis, which includes normalization and calculation of measures for differential abundance between conditions (e.g., *p*-values, fold changes, or log ratios). Mostly, these tasks are performed with commercial statistics software, open-source omics analysis applications like Mayday [12], or directly with the statistical programming language R (www.r-project.org). In a final preprocessing step, the metabolite features are annotated with candidate identifiers using information provided by metabolomics databases, such as PubChem [13], HMDB [9], or LIPID MAPS [14]. Similar preprocessing workflows exist for other platforms, like microarrays, next generation sequencing, and proteomics. The processed data can then be imported and analyzed with InCroMAP.

In a typical use case of InCroMAP (see Fig. 1), the user first imports his preprocessed multi-level omics data, given in tabular format. Then, the differing metabolites, genes and proteins are determined for each platform, based on appropriate cutoffs. Next, relevant pathways related to the biological background of the experiments are inferred. For the detection of relevant pathways, InCroMAP employs a special pathway enrichment algorithm, which integrates differing metabolites, genes, and proteins across multiple platforms. The resulting pathways can then be selected for further visual inspection from a table in which each pathway is associated with a significance value. Alternatively, the metabolic overview function of InCroMAP can be used to generate an interactive global map of cellular metabolism, in which each subordinate metabolic pathway is colored according to the significance of its enrichment. The results of the enrichment analysis can be exported in tabular format and the pathway-based visualizations can be easily stored as JPEG images.

### 2.1. Data import

InCroMAP software uses an intuitive, largely uniform input format for the import of heterogeneous types of processed omics data. The software accepts tabular input files (e.g., CSV files exported from MS Excel) which consist of platform-dependent metadata columns containing appropriate identifiers (e.g., Affymetrix, EntrezGene, HMDB IDs, etc.) and data columns corresponding to the fold-changes or *p*-values calculated for a particular sample group.

InCroMAP is able to handle various types of identifiers for genomics, transcriptomics, proteomics, and metabolomics data. It supports the recognition of identifiers used by the most common oligonucleotide microarray manufacturers (e.g., Affymetrix, Agilent, etc.). Additionally, generic formats facilitate the import of processed data, provided that each measurement can be either associated with a certain gene, genomic region, protein, or metabolite. For miRNA data the systematic name (i.e., miRBase ID) has to be provided. The miRNA transcripts can then be automatically connected to canonical pathways based on confirmed or predicted interactions with potential target mRNAs. Protein data is simply linked to the corresponding genes. Since diverse proteins corresponding to the same gene may have been measured, a gene identifier as well as an arbitrary identifier for the protein modification has to be provided. Metabolomics data may be annotated by a database identifier (e.g., KEGG, HMDB, PubChem, LIPID MAPS), an InChIKey, or a common synonym. Internally, all metabolites are mapped to InChIKeys. Pathways of interest can either be automatically downloaded from KEGG or imported from other sources (e.g., Reactome, BioCarta, etc.) in BioPAX format [15].