Fachbereich Informatik

Wilhelm-Schickard-Institut

Cognitive Systems

Lars Rosenbaum

Sand 1

72076 Tübingen

Germany

Phone: +49 7071 29-70789

Fax: +49 7071 29-5091

lars.rosenbaum@uni-tuebingen.de

www.cogsys.cs.uni-tuebingen.de

To the editors of

Journal of Chromatography B.

March 2, 2014

**Revised manuscript (former title: “Straightforward interpretation of metabolomics, proteomics, transcriptomics, and genomics data by comprehensive visualization and pathway enrichment using a single software tool”), submitted under manuscript number JCB-13-943**

Dear Dr. Theodoridis,

we have carefully read all reviewer comments and addressed each point appropriately. We are convinced that the suggestions made by the reviewers substantially improved the quality of the manuscript. In the following, we would like to discuss all reviewer comments in detail. The reviewer comments are shown in red and our answers in black below.

**Guest editor comments:**

*1) Please consider proof reading of the manuscript by a metabolomics expert in order to improve weak points in descriptions and terminology (as suggested by referee 1)****.***

As suggested by Reviewer 1 the manuscript was reviewed by an expert in the field of metabolomics. The corrections specifically included the revision of sentences where an inappropriate terminology was used in the previous version of the manuscript. Furthermore, we included detailed descriptions of the algorithms (e.g., for pathway enrichment analysis based on cross-omics data) implemented in the InCroMAP software.

*2) Please provide some additional results (preferably from metabolomics studies) and comparison with the performance of other software tools.*

In order to emphasize the increasing importance of cross-omics studies, which particularly involve the profiling of the metabolome, we cited two pioneering studies in this field, which have recently been conducted by Gruden *et al.* and Armiour *et al.*, respectively. Furthermore, Figure 2 is now based on real data from a study focused on the analysis of metabolic alterations in mouse liver tumors. The corresponding manuscript has recently been accepted for publication by the International Journal of Cancer, which is briefly described in the revised manuscript. For the sake of completeness, we also included the final version of this manuscript into the submission. Furthermore, we added a table in which we compare the features offered by InCroMAP with the features implemented in existing software tools for the visualization of metabolomics data.

*3) The way that data is imported could be described more eloquently.*

The subsection 2.1 in which we describe the import of heterogeneous types of omics data into the software was revised accordingly.

*4) Comments from referee 2 on targeted metabolomics data should be given consideration.*

The reader is referred to the revision notes for Reviewer 2.

**Reviewer 1 comments:**

*1) Also the implementation of the metabolomics part is very simple: only differentially expressed metabolites are considered. First of all, this terminology is strange: metabolites are not espressed but formed.*

We agree with Reviewer 1 that the algorithm used for pathway enrichment analysis is simple in the sense that only differentially formed metabolites are considered. If more sophisticated algorithms (e.g., single sample profiling or quantitative enrichment analysis) that account for metabolite concentrations are of interest, metabolite set enrichments can also be calculated externally by using the MSEA (<http://www.msea.ca>) or IMPaLA ([http://impala.molgen.mpg.de](http://impala.molgen.mpg.de/)) software. The pathways of interest may then be automatically loaded from KEGG or imported into InCroMAP in the BioPAX format. Please note that we do not intend to replicate these features in our software. Nevertheless, we set a high value on ensuring the compatibility to related tools in order to overcome potential limitations of our software.

The issue concerning the terminology was addressed by writing “differing” instead of “differentially expressed”, which should be more appropriate.

*2) Secondly, which kind of metabolomics is considered here? The authors talk about pathways but if one measures in blood, then the concept of pathway is very difficult.*

We apologize that this information was not included in the former version of the manuscript. We now clearly state in the introduction that the current version of InCroMAP focuses on the implementation of targeted metabolomics in systems biology data evaluation. In the revised version of the manuscript we also included an illustrative application of the tool, where metabolomic alterations in mouse liver tumors were inferred, based on NMR-based metabolomics data as well as mRNA, miRNA, protein and DNA methylation data.

We entirely agree that the investigation of pathways in cell or tissue lysates is much easier. However, in particular in human studies in most cases the investigation of metabolites in body fluids reflecting metabolic pathways is the only way and therefore usually performed. Consequently numerous metabolomics studies have been published investigating for example human plasma to study metabolic pathways or alterations therein. In addition, since decades traditional plasma metabolite parameters have been analyzed to study pathways, like glucose/lactate to study glycolysis, plasma fatty acid profiles by gas chromatography, or acyl carnitine pattern in blood for diagnostic purposes to detect enzymatic defects in the catabolic fatty acid pathway.

*3) The authors are clearly not from the field of metabolomics (they do not mention the proper and much used tools for data (pre)m processing in metabolomics. They do not refer to existing tools or software for metabolite enrichment analysis.*

We apologize that the bioinformatic focus of our manuscript led to the conclusion that the authors are not from the metabolomics field. We can dispel these concerns about lack of metabolomics expertise of the authors. The senior author (RL), for example, published in 2013 six articles presenting results from metabolomics / lipidomics studies (sum of impact factors: 38.6) and one review in J. Chrom. A.

Furthermore, we improved the clarity of the description of the data pre-processing step in metabolomics in the method section and cited exemplarily our recent publication about data preprocessing (Kenar E., et al.: Automated label-free quantification of metabolites from LC-MS data. *Mol. Cell. Proteomics*, 13: 348-359 2014). Existing software tools for metabolite enrichment analysis are now included in subsection 2.2.

**Reviewer 2 comments:**

**Major Revisions**

*1) The multi-level integration is clearly an advantage of the proposed tool. But this could be demonstrated by a powerful application to show which new biological insights can be provided by integrating more than two levels. Since the first software version was published in 2012, are there experimental studies using this tool? This should be cited. Also unpublished collaboration work could be shortly explained.*

*Since the main purpose of the manuscript is the integration of metabolomics data, an example application from the viewpoint of metabolomics would be important.*

In the EU-funded project MARCAR ([http://www.imi-marcar.eu](http://www.imi-marcar.eu/)), we perform extensive cross-omics profiling based on liver samples from rodent in-vivo studies. Most notably, one of our manuscripts entitled „Metabolic programs orchestrated by the activated Ha-ras and β-catenin oncoproteins in mouse liver tumors” was recently accepted for publication by Int. J. Cancer. In this study the InCroMAP software was extensively applied to the pathway-based, integrated analysis of transcriptomics, proteomics and metabolomics data obtained from mouse liver tumors differing in their mutation status. As proposed by Reviewer 2 this work is now cited and shortly explained in the revised version of the manuscript (p. 4, subsection 2.3). Additionally, we updated Figure 2, which is now based on real data from the mouse tumor study. However, since we would like to remain the focus on the presentation of the novel features of the InCroMAP software (e.g., enhanced support for metabolomics data), an in-depth discussion of the gained biological insights is beyond the scope of this article.

*2) How does the proposed tool outperform existing tools?*

*The authors avoid a detailed comparison with existing tools like MassTRIX and Paintomics by using the argument that those tools consider only two omics levels. Even though the proposed tool has the advantage of integrating several omics levels, typical practical applications consider only two levels, e.g., metabolomics and transcriptomics. A more detailed comparison of the proposed software tool with existing tools would therefore be necessary.*

We agree with Reviewer 2 that a more detailed comparison with existing tools should be part of the manuscript. Thus, besides a detailed comparison to the enrichment methods of Paintomics and IMPaLA in subsection 2.2, we compiled a table that summarizes the features of InCroMAP in comparison with the existing tools MassTRIX, IMPaLA, and Paintomics. A short discussion of the differences between the tools is included in section 3.

Paintomics is most similar compared to InCroMAP. However, as web-service it does not support an interactive analysis of the data, which allows for browsing through the different pathways. MassTRIX can annotate metabolic features (mass and retention time) with candidate identifiers, but it does not have an integrated enrichment analysis and cannot perform a gradual coloring based on the abundance profiles. IMPaLA has the most advanced enrichment analysis, but does not support probe-level data and does not provide any means of visualization. Furthermore, InCroMAP provides unique features for data analysis and visualization which are not offered by any published software. Most notably, a structured, global view of the changes in cellular metabolism can be generated by using the metabolic overview feature provided by InCroMAP (shown in Figure 2B). While InCroMAP facilitates the integrated analysis of metabolomics, mRNA, miRNA, protein and DNA methylation data, other pathway analysis tools are limited to either solely metabolomics data (e.g., MSEA, MBRole, MPEA) or facilitate the integration with only one other platform (e.g., IMPaLA, Paintomics, MassTRIX).

*3) The authors emphasize their strong effort in accepting and merging different kinds of gene or metabolite identifiers. But it remains unclear which of those tasks are new compared to similar software.*

We thank Reviewer 2 for making us aware of this important point which was insufficiently addressed in the old version of the manuscript. In this context we would like to point out that InCroMAP automatically recognizes gene identifiers from diverse databases (e.g., EntrezGene IDs, HGNC gene symbols, etc.) as well as probeset identifiers from the most common microarray platforms (e.g., Affymetrix, Agilent, etc.). While different types of identifiers are also accepted by the IMPaLA software, which also facilitates the calculation of integrated pathway enrichments, IMPaLA does not support probeset IDs and enable the mapping of these identifiers to gene symbols. Furthermore, InCroMAP is currently the only pathway enrichment analysis tool which is capable of integrating miRNA expression data by modeling the impact on pathways based on confirmed miRNA-mRNA interactions. As for gene identifiers, InCroMAP can automatically recognize identifiers of several different databases. While we do not support automatic annotation of metabolite features, like MassTRIX, we tried to support as much identifiers as possible by merging the metabolite identifier information of four large compound databases: HMDB, LIPIDMAPS, KEGG, and PubChem. More information on the generation of the mapping database can now be found in subsection 2.2.

*4) To identify pathways affected by an experiment, a "special pathway enrichment" algorithms is used, but a detailed description of how metabolomics data are included in the algorithm is missing as well as a comparison to pathway enrichment of other tools. Also, what does it mean "pathway enrichment is limited to targeted metabolomics data"? Is the visualisation in general limited to targeted data?*

We thank Reviewer 2 for this critical remark. To the best of our knowledge only two other programs (IMPaLA/Paintomics) exist which are capable of the integrating both transcriptomics and metabolomics data for the purpose of pathway enrichment analysis. In short, IMPaLA calculates the enrichment p-values independently for each platform using either a hypergeometric test or Wilcoxon enrichment analysis. Then, as the experiments are considered independent, the joint p-value is computed based on the product of the p-values calculated for the individual platforms. Paintomics also uses a hypergeometric test and calculates one final p-value. However, the details remain unclear from the publication. In contrast to IMPaLA, InCroMAP performs only one statistical test in which all genes and metabolites showing significant changes are combined to one list. Analogously, the gene and metabolite universe are pooled in the hypergeometric test statistic. As suggested by Reviewer 2, the conceptual differences between the InCroMAP, Paintomics, and IMPaLA algorithms for cross-platform enrichments are now explained comprehensively in the revised version of the manuscript (subsection 2.2).

In general InCroMAP is not limited to targeted metabolomics data. However, assigning multiple candidate identifiers to metabolic features using mass and retention time usually results in many-to-many mappings because a metabolic feature can be mapped to several candidate identifiers and vice versa. These many-to-many mappings are problematic for the overrepresentation analysis used by InCroMAP. The visualization itself is also applicable for nontargeted data. The revised manuscript explains the problem of nontargeted data and the resulting many-to-many mappings in more detail in subsection 2.2.

*5) The presented tool requires input data in form of significant values (p-value/ log fold change). This means the complete data pre-processing remains by the user. A big user friendly improvement could be done by integrating some parts of data processing or even accepting raw data as input, as partly done by other tools.*

We thank Reviewer 2 for this suggestion. However, we would like to point out that the use of tabular processed data corresponding to fold changes and/or p-values is one of the central design concepts underlying the InCroMAP software. All core developers agreed on this concept as firstly, the central aim of InCroMAP is not the replication of features (e.g., for the pre-processing of raw data) offered by existing tools and secondly, as the ease-of-use is ensured by a largely uniform input format which is suitable for the import of heterogeneous omics datasets. Thus, beyond elementary convenience functions (e.g., for the detection and conversion of IDs) no support in data preprocessing is provided.

**Minor Revisions**

*1) What is the meaning of the acronym InCroMAP? I could not find.*

As suggested by Reviewer 2 the meaning of the acronym InCroMAP (Integrated analysis of Cross-platform MicroArray and Pathway data) is now explained in the introduction.

*2) The title "Straightforward interpretation… " might be misleading since there is no data interpretation in the manuscript.*

We agree with Reviewer 2 that the title should indicate more clearly that the focus of this work lays on the software and not on its application to biological data. Thus, the title of the manuscript was changed to “Integrated enrichment analysis and pathway-centered visualization of metabolomics, proteomics, transcriptomics and genomics data by using the InCroMAP software”.

*3) Figure 2B shows a pathway, but metabolite data (colored circles) seems not to be included in the visualization.*

Since the focus of the manuscript is on metabolomics data visualization, we agree with Reviewer 2 that it is more appropriate to show a metabolic pathway in Figure 2B. The Figure was now updated and depicts the KEGG pathway “Glycolysis / Gluconeogenesis”. Since Reviewer 1 also suggested to mention experimental studies using InCroMAP in this article, the pathway was overlaid with real data from a study of the intermediary metabolism of Ras-mutated mouse liver tumors. As stated previously the corresponding manuscript was recently accepted by the International Journal of Cancer and was added to the submission for the sake of completeness.

**Reviewer 3 comments:**

**Major Revisions**

*1) I´m missing more detailed explanations on the computation, which is important in the field of such tools. E.g. good background set for genes, proteins and metabolites are needed for calculation of enrichment and p-values. Furthermore, the authors do not explain in detail how combined enrichment works and how values are calculated.*

We fully agree with Reviewer 3 that a more detailed description of the algorithm used for the pathway enrichment is required. Thus, we added a comprehensive explanation of the adopted methods to the methods section (see subsection 2.2). Furthermore, as suggested by Reviewer 2 we also explain how the algorithm implemented in InCroMAP differs from the approaches used in related software.

*2) I would also like to know how they implemented the metabolite database for mapping between the different identifiers. This is important, because a lot of people are struggling with harmonizing different metabolomics databases.*

We thank Reviewer 3 for this remark. Indeed, the generation of an internal metabolite identifier database for the mapping between different identifiers was possibly one of the milestones during the development of the extended version of InCroMAP. The most crucial step was deciding for an internal identifier. We chose InChiKeys as they are a compact, almost certainly unique identifier for a metabolite, which does not change as frequently as some database identifiers (HMDB, LIPIDMAPS, etc.). We downloaded compound information from PubChem, LIPID MAPS, KEGG, and HMDB. We merged the data based on the InChIKeys with self-made python scripts and kept only entries that contained identifiers for at least two different databases. We are aware of the fact that the generated database almost certainly contains errors because two different databases might assign a different InChIKey for a metabolite. Furthermore, some database entries from HMDB or LIPIDMAPS do not have InChiKeys, e.g. a class of compounds. It might be possible to map this information to our generated identifier database using compound taxonomies. However, we did not implement this feature in the current version, mainly because it might introduce more errors in the generated database. We outlined the points discussed in this paragraph in subsection 2.1.

**Minor Revisions**

*1) Page4 line 2-3. Sentence "A main goal..." can be improved.*

In response to the suggestion made by Reviewer 3 the corresponding sentence was revised accordingly.

Sincerely yours,

Lars Rosenbaum

(for the authors)