

SOFTWARE ARTICLE

Open Access



# msBiodat analysis tool, big data analysis for high-throughput experiments

Pau M. Muñoz-Torres\* , Filip Rokć, Robert Belužić, Ivana Grbeša and Oliver Vugrek

\*Correspondence:  
pau.munoz@irb.hr  
Translational Medicine Group,  
Institut Rudjer Bošković, Division of  
Molecular Medicine, Bijenička Cesta  
54, 10000 Zagreb, Croatia

## Abstract

**Background:** Mass spectrometry (MS) are a group of a high-throughput techniques used to increase knowledge about biomolecules. They produce a large amount of data which is presented as a list of hundreds or thousands of proteins. Filtering those data efficiently is the first step for extracting biologically relevant information. The filtering may increase interest by merging previous data with the data obtained from public databases, resulting in an accurate list of proteins which meet the predetermined conditions.

**Results:** In this article we present msBiodat Analysis Tool, a web-based application thought to approach proteomics to the big data analysis. With this tool, researchers can easily select the most relevant information from their MS experiments using an easy-to-use web interface. An interesting feature of msBiodat analysis tool is the possibility of selecting proteins by its annotation on Gene Ontology using its Gene Id, ensembl or UniProt codes.

**Conclusion:** The msBiodat analysis tool is a web-based application that allows researchers with any programming experience to deal with efficient database querying advantages. Its versatility and user-friendly interface makes easy to perform fast and accurate data screening by using complex queries. Once the analysis is finished, the result is delivered by e-mail. msBiodat analysis tool is freely available at <http://msbiodata.irb.hr>

**Keywords:** Bioinformatics, Data analysis, Proteomics, Data mining, Mass spectrometry, High-throughput analysis

## Background

High-performance techniques, such as mass spectrometry, are extremely powerful tools used in laboratories around the world [1, 2]. Those techniques are used to capture the existing proteins in an exact point of the cell cycle, and the output results in a list of thousand of proteins. Frequently, only a few of the records in the list are interesting for researchers proposes. To find those which want to be studied, an efficient filtering method is needed.

Different approaches have been proposed to this objective. By using a spreadsheet, results can be filtered according to the resulting score given in the experiment, but it does not provide any additional information of the proteins in the list. This limitation is overcome by PIQMIe [3] which also merges the data with proteomic information. The

information resulting from PIQMIe analysis includes a concise graphical summary of the experiment, and a SQL file which can be downloaded. This file can be used to perform complex queries on the data in a relational database management system. However, it requires a previous knowledge on database querying. MaxQB Database [4] is a database focused on the quantitative analysis of proteomic data which compares the results of the experiment against others previously published. This comparison gives a clear idea of the changes occurred in the proteins concentrations of the sample due to a specific treatment, but it does not provide any type of qualitative information about the proteins in the sample. Among the commercial solutions that may help us to perform MS data interpretation, Ingenuity platform (IPA) from Qiagen [5] and ProteinCenter (<https://www.thermofisher.com/order/catalog/product/IQLAAEGABSFALUMAZB>) from Thermo Scientific are the most used. Meanwhile IPA covers different biological aspects of the data, from the genomic to its metabolomic activity, ProteinCenter is more specific for the proteomics analysis of the samples. Both services allow cross-evaluation of databases and thus simplifying the interpretation of results.

Those inconveniences can be overcome with msBiodat Analysis Tool. The service is intended to select relevant information from proteomics experiments, and can be especially useful to select proteins according to its GO [6] annotations. It provides an interesting information about protein localization, functions, or processes in which proteins are involved. The resulting information allows to identify easily which proteins can interact during a specific biological process.

## Implementation

### Web server implementation

msBiodat Analysis Tool is written in PERL 5.10.1-17, and it is hosted in a debian 6.0.10 server and runs Apache/2.2.16 web server. To manage spreadsheets, Spread-sheet::XLSX and Excel::Writer::XLSX libraries are used. Libraries used to manage mztabs files were written by the authors of this work.

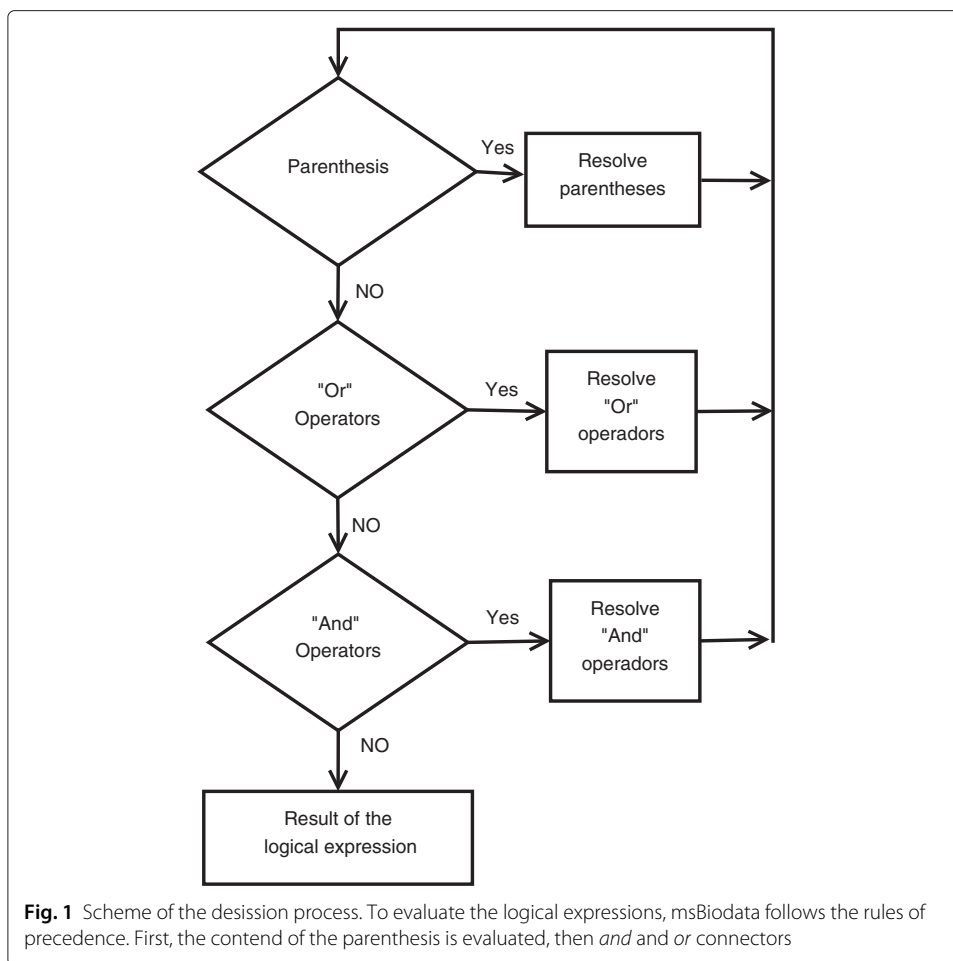
### Algorithm and decision work-flow

The work-flow followed by the algorithm to select data depends on the used query. In general, each condition in the query is solved independently. Then the complete query is evaluated according to the precedence of the connectors (*and* and *or*) and parenthesis (Fig. 1). If the evaluation of some clause depends on information not given at the spreadsheet, i.e. GO annotation, it is downloaded before resolving the clause.

## Results and discussion

A real data set will be used to show the possibilities of msBiodat. In this example, different options will be used to create complex query. To perform the example, the data from de Sousa Abreu et al [7], which can be found at [http://www.marcottelab.org/MSdata/Data\\_01/](http://www.marcottelab.org/MSdata/Data_01/), was downloaded from the MS/MS Shotgun Proteomics Data Repository (<http://data.marcottelab.org/MSdata/>). In the example is used the comparison cytosol-GFP versus MSI from the data set 1 ([http://www.marcottelab.org/MSdata/Data\\_01/z\\_MSIGFP\\_CYTO.short.annot](http://www.marcottelab.org/MSdata/Data_01/z_MSIGFP_CYTO.short.annot)).

The objective of this study was to find an association between musashi1 down-stream targets and cancer disease in humans. The high-throughput techniques used in the

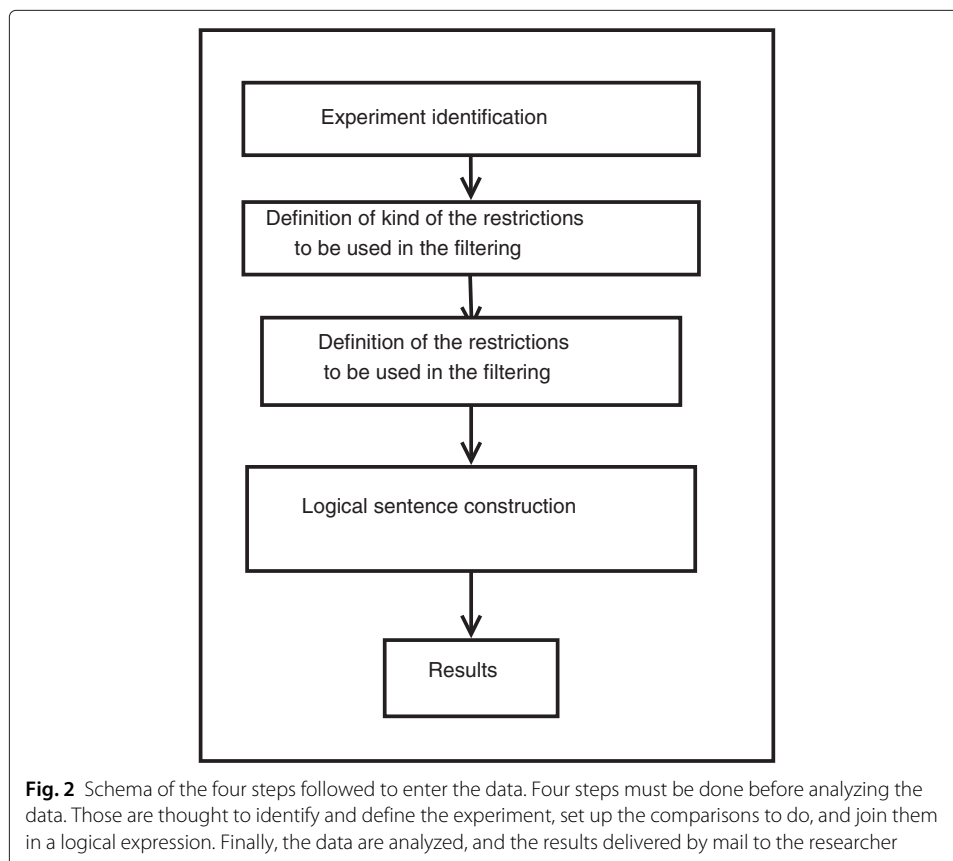


study were: microarrays (RIP-Chip), the APEX pipeline [8, 9], and a MS-based proteomic technique. After performing the experiments, the proteins that were found to have different expression in both samples were manually classified according to their GO annotation. The classification can easily be automatized using msBiodata. In the following paragraphs an overview of how to perform the analysis is explained by the example above.

The first steps of the data introduction are aimed to identify the experiment. It is necessary to fill in the following information: a project name, (to be included in the mail subject), an e-mail, and a file with the data to analyze. The file with the data can be formatted as an excel spreadsheet (XLSX) or a mzTAB file. It is also necessary to introduce where the data to be analyzed are placed. It can be a name of a sheet from an excel file, or molecule that want to be studied from a mzTAB file. In the following step, the type of statements to be used in the analysis are introduced. A statement is a condition that the selected data must meet. The web allows two types of statements which are thought to compare the content of the fields in the dataset against a value introduced by the user. Single statements are those where a unique field in the dataset is compared; it is also used for searching against GO database. On the other hand, double statements are used to evaluate the relationship between two fields of the same entry in the original data. In this example, one

comparison of each type are performed. Last two steps before analyze the data are necessary to enter the statements to perform the data selection and to link them in a logical clause (Fig. 2).

The data set was downloaded as comma-separated value file (CSV) and converted to an XLSX file using a spreadsheet software (Additional file 1). The original CSV file had 27 rows of information that were removed, one row containing the header for each field in the document, and 2431 rows of data to analyze. In this case, the analysis is performed using two statements, one of each type, which are set-up by fulfilling the elements presented in the web (pop-up menus and text boxes). The single statement is aimed to identify putative proteins involved in cancer using the possibility of querying against GO. It requires both a protein identifier and the species to be studied (*Homo sapiens*) which are introduced at the boxes reserved to this propose. The protein identifiers can be the Ensembl [10] codes (ENSP and ENST), the UniProt [11] identification code, or the Gene codes. In the dataset, the protein codes are under the label #PROTEIN ID. The clause itself is introduced by using the label Gene ontology from the menu, the comparison symbol to be used is “equal to” (=), and the keyword *proliferation* which must be written in the text box. The box labeled as modifier can be used in those cases where is needed to discriminate fields with numeric values around 0. The second clause is aimed to find those proteins which have different expression in both samples. It can be done evaluating the difference of the APEX values of both samples by calculating its coefficient. The values around the threshold can be easily discriminate by applying a decimal logarithm to



the coefficient. If the result is above 1, the evaluated protein is over-expressed at the cancer cells. As two fields from the original data are required, a double comparison must be used. The statement,  $\log_2(APEX1/APEX2) \geq 1$ , is introduced as it is explained above. This example also could be performed by using directly the field  $APEX1/APEX2$  from the data, nevertheless, using double comparison for a demonstration purpose is more appropriate. The last step is linking all together using logical connectors and parenthesis. The final query must look like *GeneOntology = proliferation and  $\log_2(APEX1/APEX2) \geq 1$* . The output resulting to this analysis will be a XLSX spreadsheet containing all the fields of the original data set with rows meeting the restrictions.

The versatility of the tool allows to increase the complexity of the queries in order to perform complex searches. An example can be a query aimed to select not only all over-expressed proteins involved in cell proliferation, but also those that are under-expressed and act as cell cycle suppressors. This query has two parts delimited by parenthesis and joined by an *or* operator which means that selected proteins must meet at least one of them. The tricky part for this query comes with the introduction of the GO terms. It is programmed to interpret all the terms written in the same text box as if they were linked by the *or* operator. To ensure that the selected proteins by the second part of the query will be involved in the cell cycle suppression, two GO statements must be included: *cycle*, and *suppression*. In this case, the final query is: *((Gene Ontology = proliferation and  $\log_2(APEX1/APEX2) \geq 1$ ) or (Gene Ontology = cycle and Gene Ontology = suppression and  $\log_2(APEX1/APEX2) < 1$ )).*

In the end, a special type of search will also be introduced: it will be used to find elements of a given field in a different one. This is performed by using *in* operator, as a way to compare the fields in a double clause. The result is that all items in the first field will also appear in the second field. In this case, no more elements are required.

## Conclusion

A tool to select data from the results of a mass spectrometry experiments is presented. The aim of the tool is to help researchers at the last steps of the high-throughput experiments by making data interpretation easy. It is thought as an easy-to-use and versatile web application which allows researchers to perform fast data selection by combining information from different sources. msBiodat Analysis Tool is especially useful to select proteins according its annotations, providing interesting information about processes where they are involved. Once the analysis is finished, the results are sent by e-mail to the researcher.

If a comparison against others tools is performed, the simplicity of msBiodat analysis tool presents some clear advantages. PIQMIe is the most similar tool to one presented in this article. It is thought as a data management, analysis and visualization tool, and returns both, a summary of the analyzed data, and an SQLite document. This document allows researchers to obtain interesting information by combining its data using a relational database management system. Nevertheless, querying those systems requires a previous knowledge. Our tool overcomes this inconvenient by implementing a guided path that researchers can follow to obtain the same information. ProteinCenter from thermo Scientific, and IPA ingenuity from Quiagen are powerful tools to investigate the relationship between the different proteins in a sample obtained from MS experiments. They allow to look inside putative pathways according the possible relations between proteins. To

achieve this goal, the software query into different databases, and score each element in the sample according to the service criteria. As they are commercial products, a user license is required. In contrast, msBiodata is not able to build a pathway with the proteins in the sample, but it can be freely used by all the scientific community. MaxQB is a database for proteomics projects. It allows to compare projects and cells lines, and visualize the differences in the protein expression levels. It can be used to motorize the response of the cell in front of different treatments. Again, it is far from the objectives of msBiodata Analysis tool, but the combination of both tools can be interesting for researchers.

### Availability and requirements

Project name: msBiodat

Project home page: <http://msbiodata.irb.hr>

Operating system: Linux

Programming language: Perl

Other requirements: Perl + dependencies

License: GNU GPL

Any restrictions to use by non-academics: none

The code of the web page is available under request.

### Additional file

**Additional file 1:** Excel spreadsheets. XLSX file containing the data from Sousa Abreu et al. which is used in the example of the article. (XLSX 611 kb)

### Acknowledgments

Not applicable.

### Funding

This work was supported by funds received from the European Union's Seventh Framework Programme for Research and Technological Development under grant agreement No 316289 - InnoMol, FP7-REGPOT-2012-2013-1.

### Availability of data and materials

The data from a Sousa Abreu et al is public available at the url [http://www.marcottelab.org/MSdata/Data\\_01/z\\_MSIGFP\\_CYTO.short.annot](http://www.marcottelab.org/MSdata/Data_01/z_MSIGFP_CYTO.short.annot). The consent of the authors of the study to use the data was also given. The data is also available as supplementary materials.

### Authors' contributions

FR and RB carried out the video production. All the authors suggested the functionalities of the software and carried out the beta testing of the web page. PM implemented the software and wrote the draft of the manuscript. All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

### Consent for publication

Not applicable.

### Ethics approval and consent to participate

Not applicable

Received: 8 April 2016 Accepted: 25 July 2016

Published online: 19 August 2016

### References

1. Chandramouli K, Qian PY. Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum Genomics Proteomics*. 2009;2009. doi:10.4061/2009/239204.
2. Domon B, Aebersold R. Mass spectrometry and protein analysis. *Science*. 2006;312(5771):212–7. doi:10.1126/science.1124619.

3. Kuzniar A, Kanaar R. Piqmie: a web server for semi-quantitative proteomics data management and analysis. *Nucleic Acids Res.* 2014;42(Web Server issue):100–6. doi:10.1093/nar/gku478.
4. Schaab C, Geiger T, Stoehr G, Cox J, Mann M. Analysis of high accuracy, quantitative proteomics data in the maxqdb database. *Mol Cell Proteomics.* 2012;11(3):. doi:10.1074/mcp.M111.014068. <http://www.mcponline.org/content/11/3/M111.014068.full.pdf+html>.
5. Felciano RM, Bavari S, Richards DR, Billaud JN, Warren T, Panchal R, Krämer A. Predictive systems biology approach to broad-spectrum, host-directed drug target discovery in infectious diseases. *Pac Symp Biocomput.* 2013;17–28.
6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet.* 2000;25(1):25–9. doi:10.1038/75556.
7. de Sousa Abreu R, Sanchez-Diaz PC, Vogel C, Burns SC, Ko D, Burton TL, Vo DT, Chennasamudaram S, Le SY, Shapiro BA, Penalva LOF. Genomic analyses of musashi1 downstream targets show a strong association with cancer-related processes. *J Biol Chem.* 2009;284(18):12125–12135. doi:10.1074/jbc.M809605200.
8. Vogel C, Marcotte EM. Calculating absolute and relative protein abundance from mass spectrometry-based protein expression data. *Nat Protoc.* 2008;3(9):1444–1451. doi:10.1038/nprot.2008.132.
9. Dasari S, Nagalla S, Searle B, Turner MA. Methods and systems for identification of macromolecules. Google Patents. 2005. Patent App. PCT/US2005/005, WO, 524. <http://www.google.com/patents/WO2005088302A1?cl=en>.
10. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kähäri AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Aken BL, Birney E, Harrow J, Kinsella R, Muffato M, Ruffier M, Searle SMJ, Spudich G, Trevanion SJ, Yates A, Zerbino DR, Flicek P. Ensembl 2015. *Nucleic Acids Res.* 2015;43(Database issue):662–9. doi:10.1093/nar/gku1010.
11. The Uniprot Consortium. Uniprot: a hub for protein information. *Nucleic Acids Res.* 2015;43(Database issue):204–12. doi:10.1093/nar/gku989.

Submit your next manuscript to BioMed Central  
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

