Biochimica et Biophysica Acta 1844 (2014) 52–62

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

A practical data processing workflow for multi-OMICS projects ☆

Michael Kohl [a,*], Dominik A. Megger [a], Martin Trippler [b], Hagen Meckel [a], Maike Ahrens [a], Thilo Bracht [a], Frank Weber [c], Andreas-Claudius Hoffmann [d], Hideo A. Baba [e], Barbara Sitek [a], Jörg F. Schlaak [b], Helmut E. Meyer [a], Christian Stephan [a,f,1], Martin Eisenacher [a,1]

[a] Medizinisches Proteom-Center, Ruhr-Universitaet Bochum, Universitaetsstrasse 150, D-44801 Bochum, Germany
[b] Department of Gastroenterology and Hepatology, University Hospital of Essen, Hufelandstrasse 55, D-45122 Essen, Germany
[c] Department of General, Visceral and Transplantation Surgery, University Hospital of Essen, Hufelandstrasse 55, D-45122 Essen, Germany
[d] Department of Medicine (Cancer Research), Molecular Oncology Risk-Profile Evaluation, University Hospital of Essen, Hufelandstrasse 55, D-45122 Essen, Germany
[e] Department of Pathology and Neuropathology, University Hospital of Essen, Hufelandstrasse 55, D-45122 Essen, Germany
[f] Kairos GmbH, Universitaetsstrasse 136, D-44799 Bochum, Germany

ARTICLE INFO

Article history:
Received 3 December 2012
Received in revised form 15 February 2013
Accepted 20 February 2013
Available online 15 March 2013

Keywords:
Multi-OMICS
Quantitative Proteomics
Quantitative Transcriptomics
Data processing workflow
Regression analysis
Biomarker

ABSTRACT

Multi-OMICS approaches aim on the integration of quantitative data obtained for different biological molecules in order to understand their interrelation and the functioning of larger systems. This paper deals with several data integration and data processing issues that frequently occur within this context. To this end, the data processing workflow within the PROFILE project is presented, a multi-OMICS project that aims on identification of novel biomarkers and the development of new therapeutic targets for seven important liver diseases. Furthermore, a software called CrossPlatformCommander is sketched, which facilitates several steps of the proposed workflow in a semi-automatic manner. Application of the software is presented for the detection of novel biomarkers, their ranking and annotation with existing knowledge using the example of corresponding Transcriptomics and Proteomics data sets obtained from patients suffering from hepatocellular carcinoma. Additionally, a linear regression analysis of Transcriptomics vs. Proteomics data is presented and its performance assessed. It was shown, that for capturing profound relations between Transcriptomics and Proteomics data, a simple linear regression analysis is not sufficient and implementation and evaluation of alternative statistical approaches are needed. Additionally, the integration of multivariate variable selection and classification approaches is intended for further development of the software. Although this paper focuses only on the combination of data obtained from quantitative Proteomics and Transcriptomics experiments, several approaches and data integration steps are also applicable for other OMICS technologies. Keeping specific restrictions in mind the suggested workflow (or at least parts of it) may be used as a template for similar projects that make use of different high throughput techniques. This article is part of a Special Issue entitled: Computational Proteomics in the Post-Identification Era. Guest Editors: Martin Eisenacher and Christian Stephan.

1. Introduction

Multi-OMICS approaches aim on an integration of different biological entities to understand their interrelation and the functioning of larger systems [1,2]. Besides their pure gain of knowledge such holistic approaches also become more and more socially important, because they may help to solve the problems of an aging population and are key technologies for personalized and preventive medicine [3,4]. Multi-OMICS approaches may also be used to identify new biomarkers for (early) disease diagnostics or they may improve the sensitivity and specificity of existing ones [5,6]. Addition of spatial and time aspects enables mathematical modeling (i.e. Systems Biology) [1], which may lead to novel insights into the mechanisms of pathogenesis and may support the development of new therapeutics [7,8].

Systems biology models benefit from inclusion of all relevant parts of the system under consideration [9,10]. On the cellular level, this means

Abbreviations: ASCII, American Standard Code for Information Interchange; BC-FC, Box–Cox-transformed fold changes; BG, bio-molecule group; CRAN, The Comprehensive R Archive Network; D_{eucl}, Euclidean distance; FC, fold change; GUI, graphical user interface; HCC, hepatocellular carcinoma; HGNC, HUGO Gene Nomenclature Committee; KEGG, Kyoto Encyclopedia of Genes and Genomes; LD, liver disease; MeSH, Medical Subject Headings; RAID, redundant array of independent disks; ^{X}PlatCom, CrossPlatformCommander; OL_{DLCT}, DIGE-LC-MS-Transcriptomics overlap

☆ This article is part of a Special Issue entitled: Computational Proteomics in the Post-Identification Era. Guest Editors: Martin Eisenacher and Christian Stephan.

* Corresponding author at: Medizinisches Proteom-Center, Building ZKF E.141, Ruhr-Universitaet Bochum, D-44801 Bochum, Germany. Tel.: +49 234 32 29288; fax: +49 234 32 14554.
E-mail address: michael.kohl@rub.de (M. Kohl).

[1] These authors share last authorship.

1570-9639
http://dx.doi.org/10.1016/j.bbapap.2013.02.029

that at least the most important 'players' (i.e. bio-molecules) and their interaction have to be investigated. Such important bio-molecules comprise the genes, which serve as information repository, defining the constraints of possible cell behavior. Parts of this information (regulated amongst others by methylation and miRNA) are retrieved by transcription forming mRNA molecules. This enables the production of proteins, the effector molecules that modulate the cellular state and behavior. The presence of proteins influences the amount of metabolites, i.e. small molecules with various functions (e.g. signaling, energy transfer and many more).

Therefore, an adequate description of the cell as a system requires the combination of several molecular biological disciplines that measure the entirety of bio-molecules: Genomics, Transcriptomics, Proteomics and Metabolomics. Fortunately, all these so-called "OMICS" approaches have developed methods that allow quantitative measurements of the targeted molecules [10,11]. Therefore, main prerequisites are fulfilled in order to give a detailed representation of the dynamics of a cell as a system. In the long term, these data can be used for development of mathematical models that may answer questions that cannot be investigated by experimentation alone.

Multi-OMICS approaches entail several issues already concerning data acquisition: Biological processes may encompass spatial and temporal scales of several orders of magnitude. Though having important consequences for cellular functioning, the amount of some relevant cellular components can be quite low. In Proteomics for example, absolute quantitative measurement of such low abundant proteins especially within complex protein samples, is hampered by technical constraints or requires application of specialized and often expensive methods. Furthermore in Proteomics and – less important – in Transcriptomics usually a mixture of cells in different states has to be analyzed to fulfill "minimum amount" requirements.

Data processing and result interpretation are also demanding due to different data structures and formats [12], the high amount of data generated and the need to annotate the experimental findings with existing knowledge obtained from the scientific community.

This paper deals with several data integration and data processing issues that frequently occur within the context of multi-OMICS projects. To this end, the data processing workflow within the PROFILE project is presented, a multi-OMICS project that aims on identification of novel biomarkers and the development of new therapeutic targets for seven important liver diseases. Although this paper focuses on combination of data obtained from quantitative Proteomics and Transcriptomics experiments only, several approaches and data integration steps are also applicable for other OMICS technologies. Therefore, this workflow (or at least parts of it) may serve as a template for similar projects. Furthermore, a software called CrossPlatformCommander (abbreviated as ^{X}PlatCom) is sketched, which facilitates several steps of the proposed workflow in a semi-automatic manner.

2. Material and methods

2.1. Data processing within the PROFILE project

Fig. 1 shows a sketch of the multi-OMICS data processing workflow within the PROFILE project. Inputs are both the data generated with Proteomics and Transcriptomics technologies as well as the patient and sample characteristics. Output of the workflow is high quality biomarker candidates ranked with respect to different statistical criteria. They are also annotated and enriched with existing knowledge and are structured for easy "manual inspection" by the experimenter.

In the following, the most important steps of the PROFILE workflow are addressed.

2.1.1. Storing patient/sample characteristics and data

Data processing within large projects like PROFILE starts with collecting and storing different kinds of information obtained by teams that work in different locations (here: surgeons and pathologists in clinics, staff of Transcriptomics and Proteomics measurement units). In order to ensure a consistent and structured collection of patient/sample characteristics, database-driven biobanking software (CentraXX by Kairos GmbH, Bochum, Germany) and a central network-accessible data repository are mandatory.

Furthermore, high-throughput techniques of multi-OMICS projects usually yield high amounts of data and therefore strongly benefit from modern centralized data storage systems: Such systems at least support handling of large file sizes. Furthermore, they provide an adequate fail-safe backup and archiving solution, e.g. a RAID system that is set up in order to ensure data integrity even in case of hardware failures. For further information regarding storage solutions for Proteomics facilities please refer to the article "The Amino Acid's Backup Bone — Storage solutions for Proteomics facilities" in this issue.

2.1.2. Quality Control (QC) and creation of patient/sample groups

Reliable data analysis strongly depends on well characterized samples and on data that passes the respective technical workflow without major problems. To this end, statistical procedures are used at the very beginning to assess high quality standards for the experimental data and for building homogenous groups of patients/samples for inclusion into further comparisons.

We use explorative (e.g. box plots) and multivariate methods (e.g. hierarchical clustering and principal component analysis) carried out using the R software environment and the package arrayQualityMetrics [13] in order to prove the expected global group structure (e.g. clearly distinguishable groups of healthy and diseased patients, healthy and abnormal tissue). Details of quality control are beyond the scope of this paper because we focus on the comparison and interpretation of data that pass this QC filter. Such kind of preliminary detection and avoidance of technical and biological outliers is indispensable in order to both minimize costs and time efforts and strongly enhance the reliability of the findings.

After QC, pseudonymized patient/sample characteristics such as gender, age, and ethnicity are used for maximizing group homogeneity. Though R scripts for QC and patient/sample matching are available, the implementation within ^{X}PlatCom is an outstanding task.

2.1.3. Data preparation for ProLiC (^{X}PlatCom module)

2.1.3.1. Data import/conversion. The above mentioned OMICS techniques have been developed within independent scientific communities and the technical equipment is developed by different vendors. In summary, this leads to different file formats: Currently, patient and sample characteristics are imported from CentraXX and ^{X}PlatCom extracts protein and peptide identification information from ProteinScape 1.3 (Bruker Daltonics, Bremen, Germany), a Proteomics laboratory information management system. Furthermore, ^{X}PlatCom reads in .csv files exported by DeCyder (GE Healthcare, Munich, Germany), a software for processing 2D DIGE gels and by Progenesis (Nonlinear Dynamics, Newcastle upon Tyne, UK), a software for LC–MS label-free quantitative Proteomics. Progenesis result files contain so-called features. In this context, a feature denotes a set of isotopes in the "retention time to m/z" map, which in their entirety originate from a peptide ion. The corresponding peptide abundance is calculated by summing up the peak volumes of this isotope set.

^{X}PlatCom processes Transcriptomics data given in the Affymetrix CEL data file format.

2.1.3.2. Internal data structure. To be able to store essentially different features of transcripts and proteins, ^{X}PlatCom internally uses a generic data structure (using the Java programming language). This structure consists of different hierarchically organized classes. The Java class *biomolecule* stores feature information shared by transcripts, proteins and other bio-molecules (e.g. measured expression levels or sequence

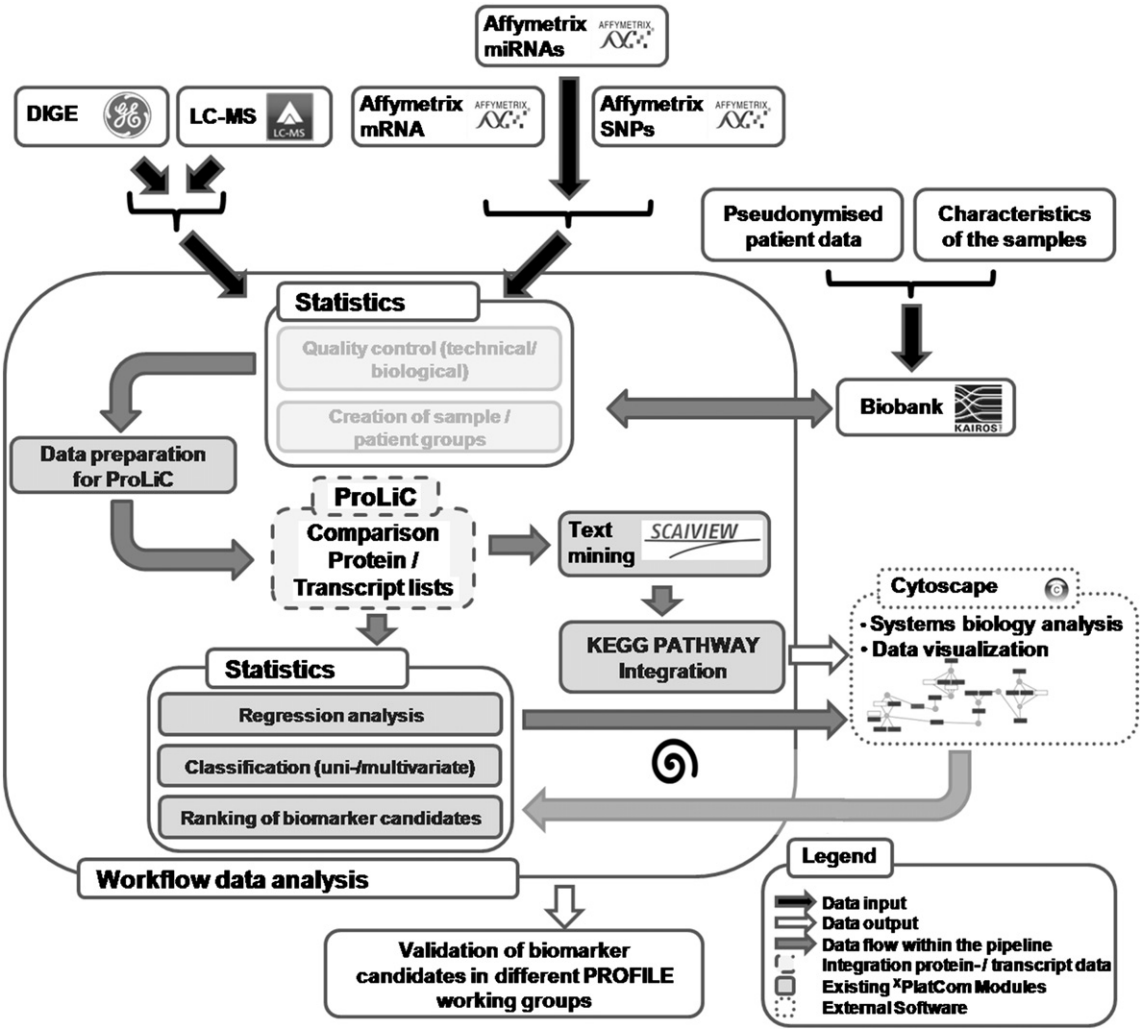


Fig. 1. Sketch of the data processing workflow performed within the PROFILE project. Transparent parts of the sketch indicate parts, which will be implemented in the near future. Note, that the "spiral" symbol denotes an iterative application of analysis procedures.

information). Derived classes like *protein* or *transcript* inherit the characteristics of the super-class and implement further features (e.g. a set of protein-related peptides or transcript-related probe sets along with the quantitation results).

2.1.3.3. Determination of common identifiers. Different -OMICS platforms also use different nomenclatures for labeling experimental data, i.e. different accession numbers or identifiers. Enabling their relation is therefore of high relevance within multi-OMICS projects. Transcript and protein identifiers often have a long-lasting history. Different scientific communities have assigned different identifiers to the same biological entity in databases and these identifiers may have changed over time. Therefore, an important data preparation step within PROFILE is the mapping of protein and probe set identifiers to one representative identifier, the "representative gene name". For Affymetrix probe sets the NetAffx Query tool [14] is used to acquire the relationship between a probe set identifier and a gene name. For protein accessions gene names are extracted by the Mascot (Matrix Science Ltd., London, UK) search engine from the UniProtKB/Swiss-Prot database used.

Then, non-primary gene name aliases and synonyms are replaced by the "representative gene name". This procedure relies on the information given by the gene_info_gz source hosted at the NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/; Entrez Gene [15]). This file links the so-called "current official gene symbol", which origins from a species-specific nomenclature committee, to a set of known gene name aliases or previous gene names.

2.1.3.4. Pre-selection. If proteins share the same peptide, its quantification is a mix of multiple protein quantifications. Preliminary data preparation thus also allows determining unique peptides with respect to the content of a certain protein database. Therefore, the Unique Peptide Finder software [16] is used to generate a theoretically digested peptide database from proteins given in the FASTA format. XPlatCom applies pre-selection in order to eliminate proteins that do not contain a minimum number of peptide identifications and/or a minimum number of unique peptides.

It is possible to select only significantly regulated Progenesis features or probe sets for the calculation of aggregated regulation values. However, this thresholding procedure is not applied in the analysis presented in the Results section.

2.1.3.5. Data aggregation. Results from LC–MS label free quantification may comprise different quantified features belonging to the same peptide sequence.

Regarding 2D-DIGE experiments, proteins may occur in several quantified gel spots. Furthermore, multiple probe sets of the Affymetrix U219 array plate may have been assigned to the same representative gene name. XPlatCom calculates aggregated values (possible aggregations:

Table 1
Effect of determining common identifiers, pre-selection, and aggregation of peptides/probe sets on data sets obtained from both Proteomics (LC–MS, DIGE) and Transcriptomics experiments.

DIGE	
Number of proteins	120
Number of proteins after pre-selection	114
Number of peptides after pre-selection	2771
Number of substituted gene name aliases	17
LC–MS (progenesis)	
Number of progenesis features with identifications	22,084
Number of proteins	3293
Number of proteins after pre-selection	2835
Number of peptides (aggregated, after pre-selection)	14,713
Number of substituted gene name aliases	342
Transcriptomics experiment	
Number of probe sets	48,801
Number of representative gene names	19,908
Number of substituted gene name aliases	310

arithmetic mean, the median or the maximum values) for these peptides, proteins and transcript expressions, respectively.

On demand, ^{X}PlatCom can add missing protein sequence information from the ExPASy WWW server (http://www.expasy.ch/; [17]), if Swiss-Prot accessions are available. Therefore, ^{X}PlatCom takes advantage of the package ExPASy, which is part of the Biopython project [18]. For execution of the Python scripts from within Java we use Jython (version 2.5.2; www.jython.org).

As final step of data preparation, the output is stored in a tab delimited text file ready for input into the Protein List Comparator software (ProLiC, see following section), which handles the next stage of the PROFILE data processing workflow.

2.1.4. Protein List Comparator (ProLiC)

As indicated by the name, the original purpose of ProLiC was the comparison of protein lists obtained from different Proteomics experiments. However, the software was extended in order to enable comparison of Transcriptomics and Proteomics bio-molecules.

ProLiC starts with calculating overlaps and complements for an arbitrary number of protein lists. It can apply three different algorithms, depending on the information available for a given protein list: accession numbers, sequences, peptides.

For the first algorithm, proteins are identical, if they have equal accession numbers. However, protein accessions may differ due to different databases used for identification or due to different versions of the same database. Therefore, matching of protein identifications with respect to accession numbers alone is usually not satisfying.

The second algorithm permits a more reliable comparison of protein lists from different sources by considering protein sequence information. ProLiC supports application of both the Needleman–Wunsch [19] (global alignment) and the Smith–Waterman [20] (local alignment) algorithms. "Percent identity" values are calculated for each pairwise alignment according to the algorithms given by Raghava and Barton [21]. Two protein sequences are considered as "sufficiently identical" if the calculated percent identity value passes a user defined threshold.

The third algorithm relies directly on peptide sequences that are identified by MS experiments. Here proteins pass the identity criteria if they share the same (or a subset of) peptides. Because the peptide–spectrum-matches that explain a certain protein are the actual available information obtained from the Proteomics experiment, this possibility is considered as the most accurate utilization of the available information. This technique is applied for the exemplary analysis in this article.

As the final result of ProLiC "sufficiently identical" proteins across all given protein lists are determined ("protein groups") and thus overlaps and complements are known and can be visualized in a Venn diagram.

Currently, ProLiC assigns a transcript to one or more existing protein groups, if its representative gene name is identical to at least one protein within this group/these groups. This procedure leads to a data structure called "bio-molecule group" (BG). A member of a bio-molecule group is either a transcript or a protein sharing some characteristics with any other member of the BG. However, there must be also at least one difference to the other BG group members; this can be for example the experiment type (e.g. DIGE or LC–MS) in which the bio-molecule was found. ProLiC saves computed overlap and complement lists containing bio-molecule groups in a tab separated ASCII file format.

2.1.5. Statistics (^{X}PlatCom module)

Overlap or complement lists generated by ProLiC are further processed to support their inspection and assessment. The statistical techniques currently implemented have been chosen with respect to the requirements of the PROFILE project (i.e. biomarker identification and drug target discovery).

2.1.5.1. Regression analysis (^{X}PlatCom module). Currently, the presented workflow comprises integration of Transcriptomics and Proteomics data. In order to evaluate the correlation between both -OMICS approaches, a linear regression analysis is carried out on both a global scale (with respect to all calculated transcript and protein regulations of a list obtained from ProLiC) and on the bio-molecule group level (considering the protein and transcript regulations of each sample within a bio-molecule group). Only bio-molecule groups are selected for regression analysis, where corresponding transcript and protein regulations exist in more than two samples (otherwise the correlation coefficient r or the coefficient of determination R^2 is 1.0).

As shown previously, neither mRNA nor protein abundances are normally distributed [22]. Therefore, a transformation procedure should be applied in order to ensure a correct estimation of significance of the correlation coefficient. We followed Nie et al. [23] and use the Box–Cox transformation [24] with an automated estimation for the parameter λ in order to ensure normality distribution. Then, these

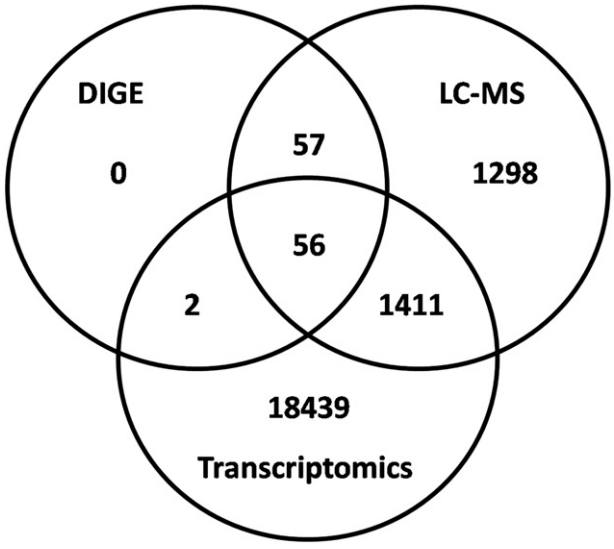


Fig. 2. A Venn diagram of overlaps and complements of the PROFILE HCC LC–MS, DIGE, and Transcriptomics bio-molecule groups. Note, that number of bio-molecule groups obtained for these technology platforms PRINCIPALLY differ: in DIGE only significantly different spots are picked and available for further data processing; in LC–MS not all proteins are identified (due to fragmentation undersampling, small dynamic range, limit of detection). In contrary, Transcriptomics virtually supports measurement of all mRNA molecules in a sample. Note further, that no thresholding procedure with respect to p-values of the Student's t-test was applied in this study.

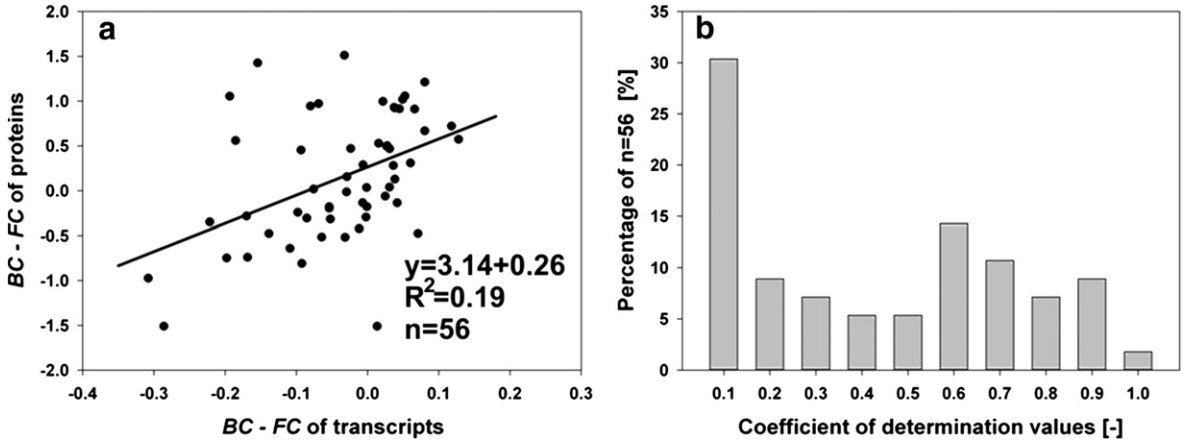


Fig. 3. Correlation between the Box–Cox-transformed transcript and protein fold changes (*BC–FC*) values given for OL_{DLCT}: a) global linear regression analysis and b) distribution of the coefficient of determination values R^2 calculated for each bio-molecule group.

values are used for linear regression analysis with the statistics environment R. The software takes advantage of the R packages car ([25]; CRAN: car) and MASS ([26]; CRAN: MASS).

2.1.5.2. Thresholding and ranking of biomarker candidates (^{X}PlatCom module). Current high throughput techniques yield huge amounts of regulated bio-molecule data. Research standards usually require validation of the findings using either an independent data set or/and application of an independent method. Because such validation procedures are often both expensive and time-consuming a selection of the most promising biomarker candidates is often mandatory. The usual way to select bio-molecules is to set thresholds for regulation (fold change, *FC*) and *p* value of a statistical test. In PROFILE, *FC* denotes the regulation of a transcript or a protein between two analysis groups and *p* is the result of a paired two − tailed Student's t-test. Default ^{X}PlatCom thresholds are $FC > 2$ or <0.5 and $p < 0.05$. Additionally, ^{X}PlatCom ranks the transcripts and proteins of a BG list using both *FC* and *p*. Therefore, the so-called Euclidean distance d_{eucl} is calculated by:

$$d_{eucl} = \sqrt{(\log_{10}(p))^2 + (\log_2(FC))^2}$$

Logarithms to bases 2 and 10 are intentionally used to equalize the different scales of *FC* and *p*. The higher d_{eucl} the larger is the distance to the origin of the coordinate system. Bio-molecules with the highest d_{eucl} values are the best candidates for subsequent validation.

2.1.6. Integration of existing knowledge from literature and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (^{X}PlatCom module)

Integration of existing knowledge is an important task for the interpretation of results. However, both the huge amount of available information and the diversity of terminologies are challenging: Most biomedical information is available in an unstructured format (e.g. articles in scientific journals or textual entries in databases). A simple search within large literature databases that assess information published within the area of life sciences often yields on the one hand many false positive matches and on the other hand may ignore important information.

In order to cope with these issues, strategies have already been developed for better retrieval of relevant information using so-called text mining algorithms. SCAIVIEW (Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Sankt Augustin, Germany) is such a software for text mining in the biomedical area. SCAIVIEW uses ProMiner [27] for an advanced terminology and name entity recognition. In the context of PROFILE several SCAIVIEW queries have been performed resulting in gene names already associated with specific disease types.

SCAIVIEW result files also contain "gene name" <-> "KEGG pathway" relationships. ^{X}PlatCom extracts this existing knowledge into a table, displaying a KEGG pathway identifier together with the representative gene names from the ProLiC BGs occurring in that pathway.

2.1.7. Cytoscape

The results of the ^{X}PlatCom workflow, i.e. bio-molecule groups along with related information (aggregated expression values, coefficient of determination values (R^2) etc.) can be loaded into the Cytoscape software [28–30]. Cytoscape can be used for the production of pathway visualizations (with colors corresponding e.g. to regression coefficient, rank, d_{eucl}) and for further analysis: For example with the Cytoscape plug-ins MCODE [31] and BinGO [32] it is possible to perform an enrichment analysis and to identify bio-molecules that share the same cellular function. Further extension of ^{X}PlatCom includes a re-import of the Cytoscape results (as indicated by the "spiral" symbol in Fig. 1) and annotation of the related bio-molecules with this data. This allows iterative re-analysis, e.g. a regression analysis of bio-molecules related to a certain functional category.

2.2. Experimental setup

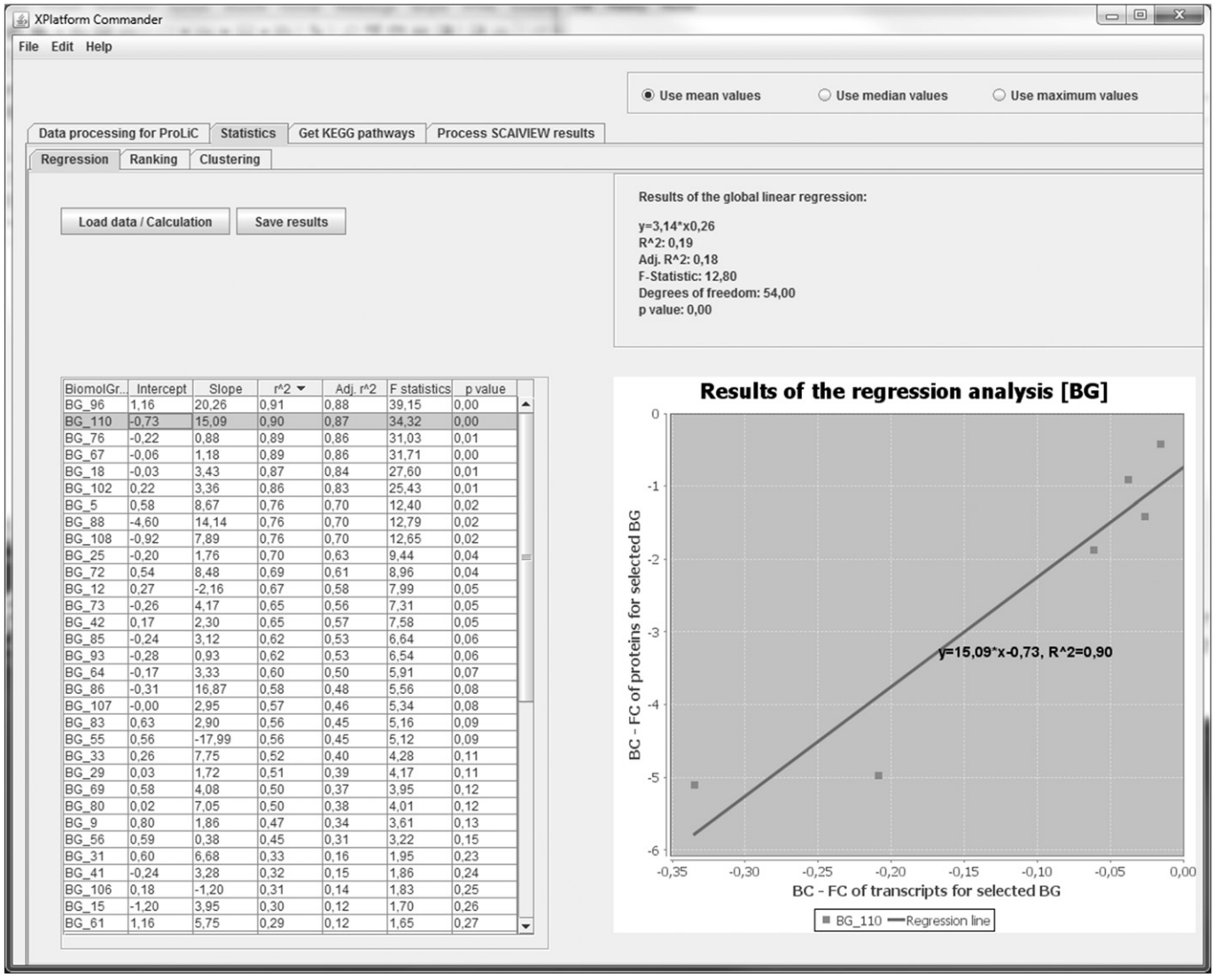
In the Results section of this article, performance of the software is demonstrated in the context of hepatocellular carcinoma (HCC), the fifth most common cancer worldwide [33].

Experimental procedures were carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans.

Transcriptomics and Proteomics (2D-DIGE and Label free LC–MS) experiments were carried out with samples obtained from six patients suffering from hepatocellular carcinoma. Both tumor and normal liver tissue (from surgical margin) samples were collected from each patient.

2.2.1. Transcriptomics experiments

From human liver needle biopsies (5–8 mg tissue) total RNA was isolated automatically on a QIAcube using the miRNeasy Micro Kit according to the protocols of the manufacturer (Qiagen, Hilden, Germany). Prior to microarray analysis RNA samples were quantified on a NanoDrop 1000 (Peqlab, Erlangen, Germany). RNA integrity was determined on the Experion Automated Electrophoresis System using the Experion RNA StdSend Analysis Kit (Bio-Rad, Munich, Germany). After passing those quality controls RNA samples were preprocessed with the Affymetrix3' IVT Express Kit. Samples were hybridized on Human Genome U219 16-Array Plates using the AffymetrixGeneTitan MC Instrument and the GeneTitan Hybridization, Wash, and Stain Kit (Affymetrix, Santa Clara, CA, USA). The GeneTitan MC Instrument was



BiomolGr...	Intercept	Slope	r^2	Adj. r^2	F statistics	p value
BG_96	1,16	20,26	0,91	0,88	39,15	0,00
BG_110	-0,73	15,09	0,90	0,87	34,32	0,00
BG_76	-0,22	0,88	0,89	0,86	31,03	0,01
BG_67	-0,06	1,18	0,89	0,86	31,71	0,00
BG_18	-0,03	3,43	0,87	0,84	27,60	0,01
BG_102	0,22	3,36	0,86	0,83	25,43	0,01
BG_5	0,58	8,67	0,76	0,70	12,40	0,02
BG_88	-4,60	14,14	0,76	0,70	12,79	0,02
BG_108	-0,92	7,89	0,76	0,70	12,65	0,02
BG_25	-0,20	1,76	0,70	0,63	9,44	0,04
BG_72	0,54	8,48	0,69	0,61	8,96	0,04
BG_12	0,27	-2,16	0,67	0,58	7,99	0,05
BG_73	-0,26	4,17	0,65	0,56	7,31	0,05
BG_42	0,17	2,30	0,65	0,57	7,58	0,05
BG_85	-0,24	3,12	0,62	0,53	6,64	0,06
BG_93	-0,28	0,93	0,62	0,53	6,54	0,06
BG_64	-0,17	3,33	0,60	0,50	5,91	0,07
BG_86	-0,31	16,87	0,58	0,48	5,56	0,08
BG_107	-0,00	2,95	0,57	0,46	5,34	0,08
BG_83	0,63	2,90	0,56	0,45	5,16	0,09
BG_55	0,56	-17,99	0,56	0,45	5,12	0,09
BG_33	0,26	7,75	0,52	0,40	4,28	0,11
BG_29	0,03	1,72	0,51	0,39	4,17	0,11
BG_69	0,58	4,08	0,50	0,37	3,95	0,12
BG_80	0,02	7,05	0,50	0,38	4,01	0,12
BG_9	0,80	1,86	0,47	0,34	3,61	0,13
BG_56	0,59	0,38	0,45	0,31	3,22	0,15
BG_31	0,60	6,68	0,33	0,16	1,95	0,23
BG_41	-0,24	3,28	0,32	0,15	1,86	0,24
BG_106	0,18	-1,20	0,31	0,14	1,83	0,25
BG_15	-1,20	3,95	0,30	0,12	1,70	0,26
BG_61	1,16	5,75	0,29	0,12	1,65	0,27

Fig. 4. Graphical user interface of the XPlatCom software. The screenshot shows the regression page with the results of regression analyses for the bio-molecule groups of OL_{DLCT}. The table lists regression results on the BG level (sortable). The diagram shows the Box–Cox transformed values along with the regression line for the BG selected in the table. Additionally, the results of the global regression analysis are given in the upper right panel.

controlled by the AffymetrixGeneChip Command Console Software (AGCC, version 3.2.4) by which also the CEL files were processed for further downstream analysis. First round GeneChip data quality control was carried out with the Affymetrix Expression Console Software (version 1.2.1).

2.2.2. Proteomics experiments

After protein purification (centrifugation at 15000 *g* for 5 min) a 2D-DIGE minimal labeling experiment was performed. ImageQuant ™ (GE Healthcare, Munich, Germany) is used for image processing. Relative protein quantitation is carried out with the DeCyder 2D™ software (GE Healthcare, Munich, Germany). Selected differently expressed protein spots were identified using MALDI-TOF-MS (UltraFlex™ II instrument (Bruker Daltonics, Bremen, Germany)) or nano-HPLC-ESI-MS/MS (Bruker Daltonics HCT plus, Bruker Daltonics, Bremen, Germany). For LC–MS analysis, 5 μg of each protein sample was loaded on a 4–20% SDS-PAGE gel (Anamed, Groß-Bieberau, Germany) and allowed to run into the gel for about 1 cm (15 min at 50 V). After Coomassie-staining, in-gel trypsin digestion was performed following standard procedures. Quantitative label-free analyses were performed on an Ultimate 3000 RSLCnano system (Dionex, Idstein, Germany) online coupled to a LTQ Orbitrap Velos instrument (Thermo Scientific, Bremen, Germany). For each analysis 15 μl of sample was injected, corresponding to an amount of 350 ng tryptic digested proteins. For the ion-intensity-based label-free quantification the Progenesis LC–MS™ software (version 4.0.4265.42984, Nonlinear Dynamics Ltd., Newcastle upon Tyne, UK) was used. A detailed specification of the experimental procedure is given in a subsequent paper.

2.3. Technical details of the CrossPlatformCommander software

XPlatCom is an application written in the Java programming language (Java Standard Edition 6, Update 13, Oracle Corporation, Redwood City, CA, USA). It possesses interfaces for the interaction with Phyton, Perl (www.perl.org) and for the integration of R scripts via Rserve 0.6-8 ([34], CRAN: Rserve). The XPlatCom software consists of several connected modules. Integration of new functionality is simplified due to this modular concept. Each module can be used as a separate command line tool or from within a graphical user interface (GUI, Fig. 4).

The command line tools can be included into user written (shell) scripts facilitating batch processing of XPlatCom. The results of EACH

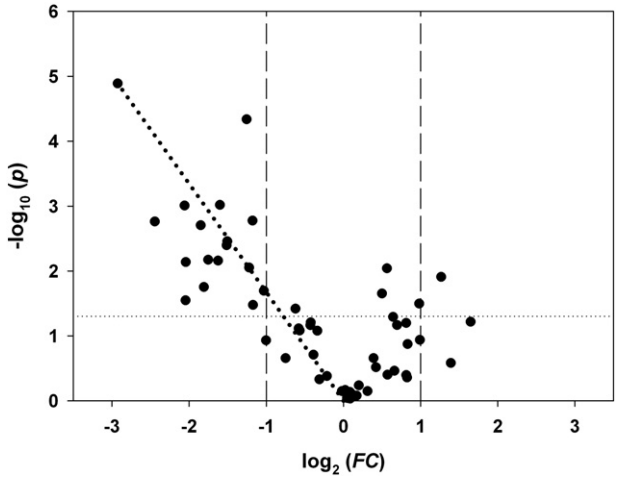


Fig. 5. Volcano plot for the proteins in OL_{DLCT} according to *FC* and Student's t test *p*-value. The arrow indicates the protein with the largest Euclidean distance value (bold dotted line to the origin of the coordinate system, i.e. $d_{eucl} = 5.7$). The thin lines mark the applied thresholds for the fold change (dashed lines correspond to *FC*s of 2 and 0.5, respectively) and for the *p*-value (dotted line corresponds to a *p*-value of 0.05).

module can be stored as ASCII formatted files. These files can then be used as input files for the subsequent ^{X}PlatCom module.

^{X}PlatCom is currently in beta status. A detailed description of the software and related information (installation, data formats, example dataset etc.) will be included in the software package at release date at http://www.medizinisches-proteom-center.de/software.

3. Results

Here, performance of the ^{X}PlatCom software is demonstrated in the context of hepatocellular carcinoma (HCC) one of the liver diseases in focus of the PROFILE project. Due to an ongoing patent registration procedure neither probe sets, nor protein accessions, nor representative gene names can be reported. Instead, the principle outcome of some steps of the data processing workflow within the PROFILE project is presented.

3.1. Data preparation for ProLiC

Considering the LC–MS experiment there are in the mean 6.7 features assigned to a protein. After pre-selection and aggregation of peptides, a protein contains on average 5.2 peptides. In the DIGE experiment, there is a very high number of peptides (24.3) assigned on average to each protein.

86% of 3293 (LC–MS) and 95% of 120 (DIGE) of the proteins pass the protein pre-selection procedure (Table 1).

In the Transcriptomics experiment on average 2.5 probe sets are aggregated to each representative gene name. For the Proteomics experiments, there is a high proportion of gene names that are considered as aliases (non-primary gene names) and substituted, i.e. 12% in LC–MS and 15% in DIGE. For details see Table 1.

3.2. Application of ProLiC

ProLiC is used to calculate the overlaps and the complements for the three full-length bio-molecule lists of LC–MS, DIGE and Transcriptomics experiments (see Fig. 2).

In the remaining sections, the outcome of further data analysis steps is – except as stated otherwise – shown for the overall DIGE–LC–MS–Transcriptomics overlap (56 bio-molecule groups). In the following, this bio-molecule group list is referred as OL_{DLCT}.

3.3. Results of the linear regression analysis

Global correlation calculated for the OL_{DLCT} list is weak ($R^2 = 0.19$, Fig. 3a). Regarding the overlap of the LC–MS and the Transcriptomics experiments (1411 bio-molecule groups in Fig. 2) there is almost no correlation on the global scale ($R^2 = 0.07$; data not shown).

However, within the bio-molecule groups, some of the 56 bio-molecule groups in OL_{DLCT} show a strong correlation between the transcript and protein regulations (Fig. 3b). In 11% of the bio-molecule groups the coefficient of determination is greater than or equal to 0.8 (for all these coefficients of determination the *p*-value of a corresponding F statistics is <0.05). A similar number of well correlated bio-molecule groups are found in the overlap of the LC–MS and the Transcriptomics experiments (8% of the R^2 values ≥ 0.8; data not shown).

3.4. Thresholding and ranking of biomarker candidates

^{X}PlatCom is used to rank bio-molecules of a given BG list. Fig. 5 shows the Volcano plot for proteins in OL_{DLCT}, which are measured by the LC–MS experiment. 18 proteins (32%) exhibit *FC* and *p*-values better than the applied thresholds (i.e. $FC > 2$ or <0.5 and p-value < 0.05). Most of these proteins are down-regulated between disease vs. healthy groups.

3.5. Integration of existing knowledge from the literature

On November 7th, 2012 several SCAIVIEW queries regarding HCC and other diseases were conducted (see Table 2). 41 representative gene names (73%) of the 56 representative gene names in OL_{DLCT} have been also found in the SCAIVIEW HCC query. There are no gene names that were found in the liver disease query, but not for HCC. 13 representative gene names (23%) have been found in the neoplasm query, but not in the HCC query (Fig. 6).

3.6. Integration of KEGG pathway information

The SCAIVIEW result file obtained from the overall gene/protein query (query number 4 of Table 2) is used to assign the representative gene names of OL_{DLCT} with KEGG pathways: 59 KEGG pathways are identified that contain representative gene names of OL_{DLCT} (some representative gene names may occur in more than one pathway). Most of these pathways contain only a small number of bio-molecules (Fig. 7). However, 18 out of 59 (30.5%) KEGG pathways are related with 5 or more bio-molecules.

Table 2
Results of the conducted SCAIVIEW disease queries. The disease queried, the number of related documents, and the number of representative gene names that were reported by SCAIVIEW are given as well. Query number 4 (without MeSH Disease term) is for mapping representative gene names to KEGG identifiers.

Query number	Query details	Disease	Number of documents found	Number of repr. gene names found
1	[Human Genes/Proteins] AND [MeSH Disease: "Carcinoma, Hepatocellular"]	HCC	26,582	4803
2	[Human Genes/Proteins] AND [MeSH Disease: "Liver Diseases"]	Liver disease	6115	1802
3	[Human Genes/Proteins] AND [MeSH Disease: "Neoplasms"]	Neoplasms	356,893	12,394
4	[Human Genes/Proteins]	–	3,734,116	18,714

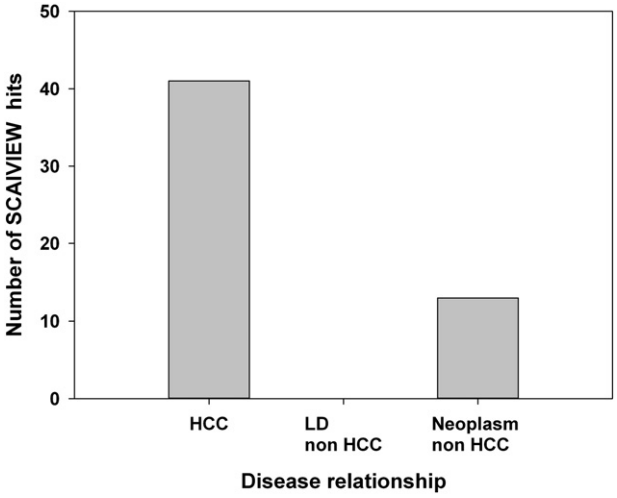


Fig. 6. Number of SCAIVIEW hits (representative gene names) found in the results of HCC, liver disease (LD) and neoplasm queries. Only representative gene names of the bio-molecule groups in OL_{DLCT} are considered.

4. Discussion

A workflow, which includes several techniques aiming on integration and processing of data generated in the context of multi-OMICs projects, was presented. To this end, ^{X}PlatCom was developed, which enables execution of several steps of this workflow in a semi-automated manner. This approach was applied to data of both Proteomics and Transcriptomics experiments generated within an ongoing liver disease project (PROFILE).

In this section, the performance of the data processing workflow is discussed in comparison with similar published approaches that refer on integration of Proteomics and Transcriptomics data. Furthermore, application of the suggested workflow is shown in relation to biomarker discovery, a common use case within the PROFILE project. Possible future extensions of the data processing workflow are sketched and remaining issues concerning data integration are addressed as well.

4.1. Performance of the data processing workflow

4.1.1. Data integration

Data of both the Proteomics and the Transcriptomics experiments are linked by representative gene names. Assignment of these common identifiers for both gene and protein entities is considered as a fundamental step for a combined analysis of multi-OMICs data [1].

Results show that in case of the Transcriptomics data the proportion of substituted aliases is very low (1.6%). However, a relevant part of the gene names, which are returned by Mascot (extracted from the used FASTA database) and assigned to proteins of the LC–MS and DIGE experiments are either gene name aliases or previous names (12% in the LC–MS and 15% in the DIGE data set). Therefore, the conversion of these names to the current official gene symbol provided by the NCBI is considered as crucial data preparation step for ensuring an optimal data comparison. There are approximately 50% of the Proteomics BGs that can be matched to transcripts (via representative gene names). This roughly corresponds to values given in the literature. Waters et al. cross-reference data obtained from two mRNA microarray platforms with data obtained from a LC–MS Proteomics experiment and achieve a 40% overlap [1].

However, the suggested procedure still includes some uncertainty. NCBI gene names arise from different sources. Amongst others, the gene_info_gz of NCBI Entrez Gene uses gene names obtained from species-specific nomenclature committees. This is in case of human gene names the HUGO Gene Nomenclature Committee (HGNC), that currently (2011) provides almost 30,000 approved gene names [35]. Nevertheless, the "current official gene symbols" are not unique between species. Therefore, it is intended to extend the ^{X}PlatCom software with a mapping opportunity to NCBI GeneIDs, because these identifiers are unique across all taxa.

4.1.2. Regression analysis

A major feature of the suggested data processing workflow is the calculation of correlation between transcripts and proteins. Globally, only 19% of the variance in the protein regulations of the OL_{DLCT} list can be explained by the transcript regulations. However, a weak global correlation between the protein and the messenger RNA expression ratio has frequently been reported [23,36]. Waters et al. review several quantitative studies comparing mRNA and protein abundance and report very low correlation in particular for studies that consider mammals [1]. Lower eukaryotes, like for example yeast, usually show a closer coupling of gene and protein regulations. Waters attributes this to a stronger relationship between genes that control cell cycle and metabolism in lower eukaryotes.

The results of our study are in general agreement with several similar mammalian studies. It has been estimated that differential expression of mRNA can determine the corresponding protein expressions by only 20%–40% [10,37,38]. Chen et al. [39] compare gene and protein expressions of 76 lung adenocarcinomas and 9 control samples (86 samples in total). Global correlation for this data set gives a Spearman correlation coefficient of −0.025. Another study uses a set of 60 human cancer cell lines and performs a comparative study with protein and mRNA expression patterns. On the mRNA level both cDNA microarrays and Affymetrix oligonucleotide chips are utilized. Global regression analysis over all 60 cell lines yields coefficient of correlation values of 0.27 (cDNA vs. protein) and 0.16 (oligonucleotide vs. protein), respectively. Tian et al. compared kinetic changes of both mRNA and protein levels for mouse (*Mus musculus*) liver samples depending on drug response [38]. R^2 obtained from global regression analysis is 0.29.

Several reasons may account for the low agreement of regulation on the Transcriptomics and Proteomics level. In has been frequently suggested that post-transcriptional regulatory mechanisms are very important [1,40]. Waters et al. [1] estimate that these mechanisms account to more than fifty percent to the discordance between mRNA and protein abundance profiles.

Regarding the mRNA this includes export from the nucleus and splicing of pre-mRNA. The life cycle of mRNA is strongly regulated and the half-lives-of mRNA individuals may vary by several orders of magnitude [41]. Furthermore, mechanisms that control translation contribute largely to the variation of mRNA-protein correlations [42,43].

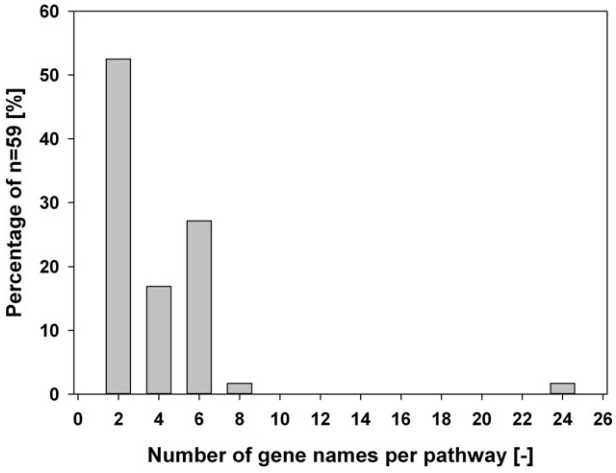


Fig. 7. Histogram for the number of gene names in OL_{DLCT} assigned to a KEGG pathway identifier.

Regarding the proteins, their half-lives are controlled by multiple highly specific and regulated mechanisms including for example protein folding and ubiquitin-mediated protein degradation [44]. Considering the complex interrelationship between gene and protein expression levels, occurrence of rather low coefficient of determination values is not surprising.

As a result, the correlation between Transcriptomics and Proteomics data is most likely far from linearity. Additionally, protein data is often biased towards high abundant proteins [23]. Therefore, capturing existing correlations within such data sets strongly depends on the ability of the statistical methods applied to cope with these issues and incorporating of adequate methods is a crucial task for a comparison of Transcriptomics and Proteomics experiments. Such techniques comprise data transformation and normalization, compensating incomplete proteomic data and statistical methods that are sensitive for non-linear relationships (for review, see [10,23]).

4.1.2.1. Data transformation and normalization. ^{X}PlatCom currently includes Box–Cox transformation in order to ensure normality. However, normalization with respect to the length of both transcripts and proteins was suggested for the improvement of correlation [23]. Implementation of this technique is currently under development.

4.1.2.2. Handling of incomplete proteomic data and the detection of non-linear correlations. Several methods have been suggested for estimation of missing values in a given Proteomics dataset with respect to the available measurements. This includes for example a '*k* nearest neighbor algorithm', the 'row-mean method', which was applied to a neuroblastoma DIGE dataset [45,46] and a 'zero inflated Poisson regression model', which was adapted for the prediction of missing protein abundance values [47]. In order to cope with the non-linear correlation between Transcriptomics and Proteomics data, a 'stochastic gradient boosted trees' approach can be used that outperforms the results of simple linear regression analysis [48].

Improvement of the regression module is considered a major task of further development. To this end, the performance of several of the above mentioned techniques will be evaluated with respect to the HCC data set.

4.1.2.3. Context related regression analysis. Another important aspect regarding the correlation of Proteomics and Transcriptomics data is considering the context of pathways and biological functioning [49,50]. It has been frequently observed that correlation of Proteomics and Transcriptomics data is related to subcellular localization and functional categories [40,50–52]. For example, it has been shown that correlations related to the localization categories 'Nucleus' and 'Cell periphery' are significantly higher than the global correlation [40]. Beyer et al. [50] observed strong correlation between mRNA and protein abundance for the functional categories 'metabolism', 'energy', and 'protein synthesis'. Furthermore, molecular machines (e.g. the ribosome, [53]) and protein/gene pairs with structural functioning [54] show a high mRNA–protein correlation.

4.2. Choice of biomarker candidates

Discovery of novel biomarkers is a complex task. There are several important characteristics that should be considered while selecting new biomarker candidates. For some important characteristics, relevant information is provided by the ^{X}PlatCom software and selection of promising biomarker candidates is strongly facilitated.

First, promising biomarker candidates show significant regulations with respect to the conditions under consideration. Additionally, differentially expressed bio-molecules should pass the t-test criteria. To this end, the ranking according to calculated d_{eucl} values, which takes both measures into account, is suitable for selection of such biomarker candidates. However, volcano plots reveal the contribution of each measure to the d_{eucl} value. This enables selection of biomarker candidates with preference of either fold change or *p*-value.

Second, accounting for known relations, which are obtained from the literature, is also important. Findings that agree with the results that have been reported for the considered disease confirm the reliability of the experimental setup. However, the most important findings are new discoveries, which are promising novel biomarker candidates. SCAIVIEW result files can be utilized for both purposes. Optionally, gene names are indicated that have been related with several disease categories of interest. In case of the HCC study, *liver diseases* or neoplasms is selected as 'co-related' categories. The most promising candidates are most likely these bio-molecules that are reported in neither SCAIVIEW result list, because they have not been associated with the hepatocellular carcinoma, other liver diseases or another neoplasm before.

Third, multivariate variable selection and classification are an important means for the development of assays suitable of disease recognition. Bio-molecules are interacting with many other bio-molecules. There are many factors that influence the expression level of a transcript or a protein. Because diseases usually alter the functioning of at least the affected cells or even the 'steady state' of the individual as a whole, measurement of a single biological entity is most likely not sufficient to clearly detect a disease or to differentiate between disease stages. Therefore, considering a panel of transcript and protein expression measurements will most likely improve both the specificity and sensitivity of a biomarker.

Therefore, it is intended to include several multivariate variable selection and classification procedures into the PROFILE data processing workflow. Implementation of these techniques as a module of ^{X}PlatCom is an on-going process. For example, an R package named PAA is developed at our institute (http://www.medizinisches-proteom-center.de/software) for Protein microarray analysis (using a random forest classifier approach [55]). This R package includes a multivariate wrapper method along with a backwards elimination algorithm that iterates through different feature (e.g. protein) subsets. Then, the classifier is applied to a current subset. PAA will be adopted for multi-OMICS data analysis.

4.3. Further extension of the data processing workflow

Further development of the ^{X}PlatCom software concerns integration of further OMICS platforms (miRNA, DNA methylation and SNP experiments). Significantly different miRNA expressions could be related to significant down-regulations of the mRNA belonging to the gene repressed by that miRNA, if known. Otherwise, data may be utilized to hypothesize, which gene is repressed by a miRNA, if that is still unknown. A similar analysis may be performed to show, whether a methylation pattern corresponds to inactivation of neighboring genes.

Concerning measurement of both protein and gene expression levels, the data collection scheme of the PROFILE project includes only DIGE/Label free and DNA microarray technologies.

However, there are different approaches available for both OMICS platforms. Regarding quantitative Proteomics a large methodological branch introduces differential mass tags into the protein or peptide that do not alter their biochemical characteristics. Such techniques are frequently grouped into 'metabolic labeling' (e.g. SILAC) and 'chemical labeling' (e.g. ITRAQ, ICPL, ICAT etc.) strategies (for review see [56–58]). In comparison with label free both metabolic and chemical labeling approaches show a higher precision. Additionally, labeling enables the absolute quantitation of protein changes even within complex protein samples (AQUA method, [59]).

On the other hand, label free methods are inexpensive. Furthermore, these techniques require no additional steps for data acquisition and they can be applied to data sets obtained from standard LC–MS/MS

runs carried out for protein identification. In summary, there is no "one method fits all" opportunity in the field of quantitative Proteomics.

On the contrary, in Transcriptomics a new developed technique named RNA sequencing (RNA-seq) has drawn great attention [60,61]. There are considerable advances in comparison with the well-established hybridization-based microarray techniques, e.g. a higher dynamic range, low background noise, a low amount of necessary RNA for measurement, etc. [61]. Therefore, this technology may substitute DNA microarrays as standard for gene expression profiling studies.

Basically, the XPlatCom software is a generic tool for the processing of quantitative Proteomics and Transcriptomics data. Therefore, each of the above mentioned techniques is supported.

However, the data output formats provided by different vendors and different scientific communities strongly vary and the programming of converters is a necessity for importing data from e.g. ITRAQ or RNA-seq experiments into XPlatCom. A huge diversity of data formats is a common task in bioinformatics. In both Proteomics and Transcriptomics the relevance of this issue has been recognized. Several initiatives have been founded that aim on the developing of standard formats. Amongst others, the Proteomics Standards Initiative (PSI) develops MzQuantML [62], which is a standard format that handles results obtained from quantitative experiments that measure proteins and peptides by mass spectrometry. The format supports both label free and labeling techniques. Final release of MzQuantML is announced for early 2013 (Gerhard Mayer, editor of the PSI Proteomics Informatics group, personal communication).

Amongst other things, the Functional Genomics Data Society (FGED, [63]) aims on development of standard data formats and required minimum information standards for Transcriptomics experiments. FGED provides two standard formats for microarray data (MAGE-ML [64] and MAGE-TAB [65]). To the best of our knowledge there is no standard format available for RNA-seq data.

In summary, standardization strongly facilitates sharing of data and the usage of available software tools. In order to provide analysis of data obtained from a broad range of Transcriptomics and Proteomics techniques, further development of the XPlatCom software includes consideration of available standard data formats.

5. Conclusions

A workflow for processing and integrating multi-OMICS data was presented using the example of corresponding Transcriptomics and Proteomics data sets obtained from patients suffering from hepatocellular carcinoma. The software CrossPlatformCommander (XPlatCom), which facilitates the execution of several tasks of such a workflow, was also presented. Application of the software was shown for the detection of novel biomarkers, their ranking and annotation with existing knowledge.

However, for capturing correlations between Transcriptomics and Proteomics data, a simple linear regression analysis is not sufficient. In the future, a search for bio-molecule sub-groups (e.g. built with respect to similar molecular function or cellular localization) with potentially better correlations may reveal meaningful trends in co-analyzed Transcriptomics and Proteomics data. Implementation and evaluation of alternative statistical approaches, which are adapted to cope with the non-linearity of such relationships, are also promising. Additionally, another important task is the integration of the multivariate variable selection and classification approach.

Keeping specific restrictions in mind the suggested workflow may be used as a template for similar projects even when considering different high throughput techniques.

Acknowledgements

This work was funded by BioNRW.PROFILE, Förderkennzeichen 005-1006-0050. PROFILE is co-funded by the European Union (European Regional Development Fund - Investing in your future) and the German federal state North Rhine-Westphalia (NRW). This work was also funded by P.U.R.E. (Protein Unit for Research in Europe), a project of North Rhine–Westphalia (Germany).

References

[1] K.M. Waters, J.G. Pounds, B.D. Thrall, Data merging for integrated microarray and proteomic analysis, Brief. Funct. Genomics Proteomics 5 (2006) 261–272.
[2] A.R. Joyce, B.O. Palsson, The model organism as a system: integrating 'omics' data sets, Nat. Rev. Mol. Cell Biol. 7 (2006) 198–210.
[3] L. Hood, R. Balling, C. Auffray, Revolutionizing medicine in the 21st century through systems approaches, Biotechnol. J. 7 (2012) 992–1001.
[4] A.D. Weston, L. Hood, Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine, J. Proteome Res. 3 (2004) 179–196.
[5] N. Spielmann, D.T. Wong, Saliva: diagnostics and therapeutic perspectives, Oral Dis. 17 (2011) 345–354.
[6] M. Katoh, Bioinformatics for cancer management in the post-genome era, Technol. Cancer Res. Treat. 5 (2006) 169–175.
[7] L.S. Sefcik, J.L. Wilson, J.A. Papin, E.A. Botchwey, Harnessing systems biology approaches to engineer functional microvascular networks, Tissue Eng. Part B Rev. 16 (2010) 361–370.
[8] F. Mac Gabhann, B.H. Annex, A.S. Popel, Gene therapy from the perspective of systems biology, Curr. Opin. Mol. Ther. 12 (2010) 570–577.
[9] T. Ideker, T. Galitski, L. Hood, A new approach to decoding life: systems biology, Annu. Rev. Genomics Hum. Genet. 2 (2001) 343–372.
[10] W.W. Zhang, F. Li, L. Nie, Integrating multiple 'omics' analysis for microbial biology: application and methodologies, Microbiology 156 (2010) 287–301.
[11] M.M. Koek, R.H. Jellema, J. van der Greef, A.C. Tas, T. Hankemeier, Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives, Metabolomics 7 (2011) 307–328.
[12] J. Quackenbush, Data standards for 'omic' science, Nat. Biotechnol. 22 (2004) 613–614.
[13] A. Kauffmann, R. Gentleman, W. Huber, arrayQualityMetrics — a bioconductor package for quality assessment of microarray data, Bioinformatics 25 (2009) 415–416.
[14] G. Liu, A.E. Loraine, R. Shigeta, M. Cline, J. Cheng, V. Valmeekam, S. Sun, D. Kulp, M.A. Siani-Rose, NetAffx: Affymetrix probesets and annotations, Nucleic Acids Res. 31 (2003) 82–86.
[15] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova, Entrez Gene: gene-centered information at NCBI, Nucleic Acids Res. 39 (2011) D52–D57.
[16] M. Kohl, G. Redlich, M. Eisenacher, A. Schnabel, H.E. Meyer, Automated calculation of unique peptide sequences for unambiguous identification of highly homologous proteins by mass spectrometry, J. Proteomics Bioinform. 1 (2008) 006–010.
[17] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R.D. Appel, A. Bairoch, ExPASy: the proteomics server for in-depth protein knowledge and analysis, Nucleic Acids Res. 31 (2003) 3784–3788.
[18] P.J.A. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M.J.L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics, Bioinformatics 25 (2009) 1422–1423.
[19] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.
[20] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.
[21] G.P.S. Raghava, G.J. Barton, Quantification of the variation in percentage identity for protein sequence alignments, BMC Bioinforma. 7 (2006) 415.
[22] B. Futcher, G.I. Latter, P. Monardo, C.S. McLaughlin, J.I. Garrels, A sampling of the yeast proteome, Mol. Cell. Biol. 19 (1999) 7357–7368.
[23] L. Nie, G. Wu, D.E. Culley, J.C. Scholten, W. Zhang, Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications, Crit. Rev. Biotechnol. 27 (2007) 63–75.
[24] G.E.P. Box, D.R. Cox, An analysis of transformations, J. R. Stat. Soc. B 26 (1964) 211–252.
[25] J. Fox, S. Weisberg, An R Companion to Applied Regression, 2nd ed. SAGE Publications, Thousand Oaks, Calif., 2011
[26] W.N. Venables, B.D. Ripley, Modern Applied Statistics With S, 4th ed. Springer, New York, 2002.
[27] D. Hanisch, K. Fundel, H.T. Mevissen, R. Zimmer, J. Fluck, ProMiner: rule-based protein and gene entity recognition, BMC Bioinforma. 6 (2005) S14.
[28] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.
[29] M.S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A.R. Pico, A. Vailaya, P.L. Wang, A. Adler, B.R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G.J. Warner, T. Ideker, G.D. Bader, Integration of biological networks and gene expression data using Cytoscape, Nat. Protoc. 2 (2007) 2366–2382.
[30] M. Kohl, S. Wiese, B. Warscheid, Cytoscape: software for visualization and analysis of biological networks, M. Hamacher, M. Eisenacher, C. Stephan (Eds.), Methods Mol. Biol. 696 (2011) 291–303.
[31] G.D. Bader, C.W. Hogue, An automated method for finding molecular complexes in large protein interaction networks, BMC Bioinforma. 4 (2003) 2.

[32] S. Maere, K. Heymans, M. Kuiper, BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks, Bioinformatics 21 (2005) 3448–3449.
[33] H.B. El-Serag, L. Rudolph, Hepatocellular carcinoma: epidemiology and molecular carcinogenesis, Gastroenterology 132 (2007) 2557–2576.
[34] S. Urbanek, Rserve — a fast way to provide R functionality to applications, in: F.L.K. Hornik, A. Zeileis (Eds.), Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, Austria, 2003, pp. 20–22.
[35] R.L. Seal, S.M. Gordon, M.J. Lush, M.W. Wright, E.A. Bruford, genenames.org: the HGNC resources in 2011, Nucleic Acids Res. 39 (2011) D514–D519.
[36] T.J. Griffin, S.P. Gygi, T. Ideker, B. Rist, J. Eng, L. Hood, R. Aebersold, Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*, Mol. Cell. Proteomics 1 (2002) 323–333.
[37] R. Brockmann, A. Beyer, J.J. Heinisch, T. Wilhelm, Posttranscriptional expression regulation: what determines translation rates? Plos Comput. Biol. 3 (2007) 531–539.
[38] Q. Tian, S.B. Stepaniants, M. Mao, L. Weng, M.C. Feetham, M.J. Doyle, E.C. Yi, H. Dai, V. Thorsson, J. Eng, D. Goodlett, J.P. Berger, B. Gunter, P.S. Linseley, R.B. Stoughton, R. Aebersold, S.J. Collins, W.A. Hanlon, L.E. Hood, Integrated genomic and proteomic analyses of gene expression in mammalian cells, Mol. Cell. Proteomics 3 (2004) 960–969.
[39] G.A. Chen, T.G. Gharib, C.C. Huang, J.M.G. Taylor, D.E. Misek, S.L.R. Kardia, T.J. Giordano, M.D. Iannettoni, M.B. Orringer, S.M. Hanash, D.G. Beer, Discordant protein and mRNA expression in lung adenocarcinomas, Mol. Cell. Proteomics 1 (2002) 304–313.
[40] D. Greenbaum, C. Colangelo, K. Williams, M. Gerstein, Comparing protein abundance and mRNA expression levels on a genomic scale, Genome Biol. 4 (2003) 117.
[41] D.A. Day, M.F. Tuite, Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview, J. Endocrinol. 157 (1998) 361–371.
[42] B. Pradet-Balade, F. Boulme, H. Beug, E.W. Mullner, J.A. Garcia-Sanz, Translation control: bridging the gap between genomics and proteomics? Trends Biochem. Sci. 26 (2001) 225–229.
[43] L. Nie, G. Wu, W.W. Zhang, Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in *Desulfovibrio vulgaris*: a quantitative analysis, Genetics 174 (2006) 2229–2243.
[44] R.D. Abreu, L.O. Penalva, E.M. Marcotte, C. Vogel, Global signatures of protein and mRNA expression levels, Mol. Biosyst. 5 (2009) 1512–1526.
[45] K. Jung, A. Gannoun, K. Stühler, B. Sitek, H.E. Meyer, W. Urfer, Analysis of dynamic protein expression data, RevStat Stat. J. 3 (2005) 99–111.
[46] K. Jung, A. Gannoun, B. Sitek, O. Apostolov, A. Schramm, H.E. Meyer, K. Stühler, W. Urfer, Statistical evaluation of methods for the analysis of dynamic protein expression data from a tumor study, RevStat Stat. J. 4 (2006) 67–80.
[47] L. Nie, G. Wu, F.J. Brockman, W.W. Zhang, Integrated analysis of transcriptomic and proteomic data of *Desulfovibrio vulgaris*: zero-inflated Poisson regression models to predict abundance of undetected proteins, Bioinformatics 22 (2006) 1641–1647.
[48] W. Torres-Garcia, W.W. Zhang, G.C. Runger, R.H. Johnson, D.R. Meldrum, Integrative analysis of transcriptomic and proteomic data of *Desulfovibrio vulgaris*: a non-linear model to predict abundance of undetected proteins, Bioinformatics 25 (2009) 1905–1914.
[49] M.P. Washburn, A. Koller, G. Oshiro, R.R. Ulaszek, D. Plouffe, C. Deciu, E. Winzeler, J.R. Yates III, Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 3107–3112.
[50] A. Beyer, J. Hollunder, H.P. Nasheuer, T. Wilhelm, Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale, Mol. Cell. Proteomics 3 (2004) 1083–1092.
[51] L. Nie, G. Wu, W.W. Zhang, Correlation between mRNA and protein abundance in *Desulfovibrio vulgaris*: a multiple regression to identify sources of variations, Biochem. Biophys. Res. Commun. 339 (2006) 603–610.
[52] E.Z. Yu, A.E.C. Burba, M. Gerstein, PARE: a tool for comparing protein abundance and mRNA expression data, BMC Bioinforma. 8 (2007) 309.
[53] S. Rogers, M. Girolami, W. Kolch, K.M. Waters, T. Liu, B. Thrall, H.S. Wiley, Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models, Bioinformatics 24 (2008) 2894–2900.
[54] S. Nishizuka, L. Charboneau, L. Young, S. Major, W.C. Reinhold, M. Waltham, H. Kouros-Mehr, K.J. Bussey, J.K. Lee, V. Espina, P.J. Munson, E. Petricoin, L.A. Liotta, J.N. Weinstein, Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 14229–14234.
[55] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[56] M. Bantscheff, S. Lemeer, M.M. Savitski, B. Kuster, Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present, Anal. Bioanal. Chem. 404 (2012) 939–965.
[57] S.E. Ong, M. Mann, Mass spectrometry-based proteomics turns quantitative, Nat. Chem. Biol. 1 (2005) 252–262.
[58] M.H. Elliott, D.S. Smith, C.E. Parker, C. Borchers, Current trends in quantitative proteomics, J. Mass Spectrom. 44 (2009) 1637–1660.
[59] S.A. Gerber, J. Rush, O. Stemman, M.W. Kirschner, S.P. Gygi, Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 6940–6945.
[60] F. Ozsolak, P.M. Milos, RNA sequencing: advances, challenges and opportunities, Nat. Rev. Genet. 12 (2011) 87–98.
[61] Z. Wang, M. Gerstein, M. Snyder, RNA-seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.
[62] http://www.psidev.info/mzquantml, (Access Date: 13.02.2013).
[63] http://www.fged.org/, (Access Date: 13.02.2013).
[64] P.T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W.L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B.J. Aronow, A. Robinson, D. Bassett, C.J. Stoeckert, A. Brazma, Design and implementation of microarray gene expression markup language (MAGE-ML), Genome Biol. 3 (2002).
[65] T.F. Rayner, P. Rocca-Serra, P.T. Spellman, H.C. Causton, A. Farne, E. Holloway, R.A. Irizarry, J.M. Liu, D.S. Maier, M. Miller, K. Petersen, J. Quackenbush, G. Sherlock, C.J. Stoeckert, J. White, P.L. Whetzel, F. Wymore, H. Parkinson, U. Sarkans, C.A. Ball, A. Brazma, A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB, BMC Bioinforma. 7 (2006).