

REVIEW

Open Access

Bioinformatic analysis of proteomics data

Andreas Schmidt, Ignasi Forne, Axel Imhof*

From High-Throughput Omics and Data Integration Workshop
Barcelona, Spain. 13-15 February 2013

Abstract

Most biochemical reactions in a cell are regulated by highly specialized proteins, which are the prime mediators of the cellular phenotype. Therefore the identification, quantitation and characterization of all proteins in a cell are of utmost importance to understand the molecular processes that mediate cellular physiology. With the advent of robust and reliable mass spectrometers that are able to analyze complex protein mixtures within a reasonable timeframe, the systematic analysis of all proteins in a cell becomes feasible. Besides the ongoing improvements of analytical hardware, standardized methods to analyze and study all proteins have to be developed that allow the generation of testable new hypothesis based on the enormous pre-existing amount of biological information. Here we discuss current strategies on how to gather, filter and analyze proteomic data sets using available software packages.

Background

Proteins are involved in almost all physiological aspects of cellular life from the catalysis of biochemical reactions within the intermediary metabolism to the processing and integration of internal and external signals. The misregulation of protein expression results in pathological states such as cancer, neurodegenerative diseases and metabolic imbalances. Proteins are synthesized by translating the information encoded in a RNA molecule to a polypeptide chain, which adopts a specific three dimensional structure. Proteins are subjected to a constant turnover making protein homeostasis a very important feature of their regulation. Many proteins function within large multimeric complexes that are highly dosage dependent. The recent developments in gathering large scale genomic, transcriptomic and proteomic data pose substantial challenges to the bioinformatic processing of these data, which have yet not been completely solved. In case of the proteomic investigation, the challenges occur at all levels ranging from sample preparation and data gathering over the raw data integration and database searching to the functional interpretation of large datasets. Based on an iterative strategy of proteomic analysis, data interpretation and

systematic challenges, hypothesis can be developed and modified, which will eventually lead to the generation of new knowledge (Figure 1).

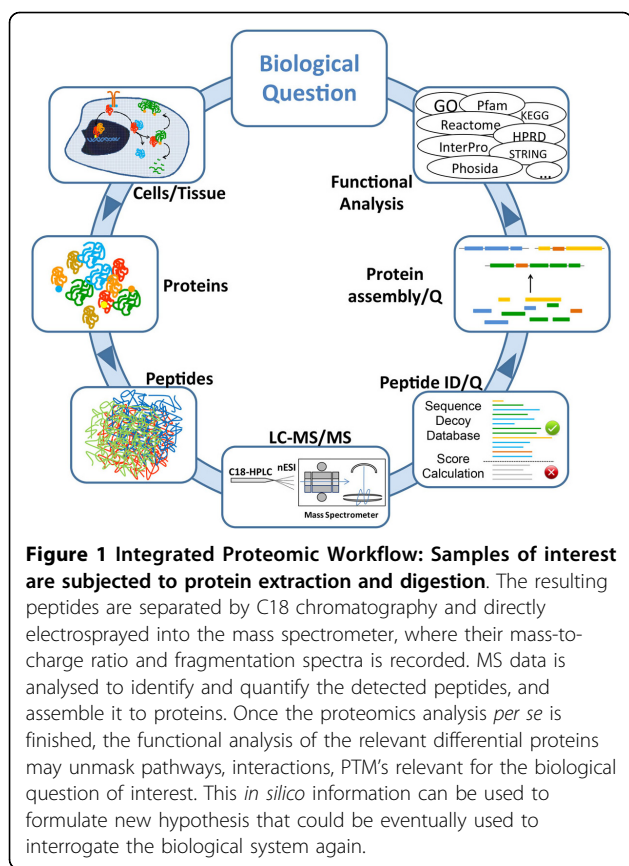
Mass spectrometry data analysis

All proteins from a sample of interest are usually extracted and digested with one or several proteases (typically trypsin alone or in combination with Lys-C [1]) to generate a defined set of peptides. Several enrichment and fractionation steps can be introduced at protein or peptide level in this general workflow when sample complexity has to be reduced or when a specific subset of proteins/peptides should be analysed (i.e. organelle specific proteome [2,3] or stoichiometric post-translational modified peptides [4]).

The peptides obtained are subsequently analysed by liquid chromatography coupled to mass spectrometry (LC-MS). The two most common approaches here are either designed to achieve a deep coverage of the proteome (shotgun MS [5]) or to collect as much quantitative information as possible for a defined set of proteins/peptides (targeted MS [6]). During the analysis peptides eluting from the chromatography are selected according to defined rules (see below) and further fragmented within the mass spectrometer. The resulting tandem mass spectra (MS^2) provide information about the sequence of the peptide, which is key to their identification. For a shotgun

* Correspondence: imhof@lmu.de

Munich Center of Integrated Protein Science and Adolf-Butenandt Institute, Ludwig Maximilians University of Munich, 80336 Munich, Germany



approach, no prior knowledge of the peptides present in the sample is required to define peptide selection criteria during the LC-MS analysis. Therefore, the peptides eluting from the chromatographic column are identified in a data-dependent mode [7], where continuously the N most abundant peptides at a given retention time are selected for fragmentation and their masses excluded for further selection during a defined time. By using this dynamic exclusion [8], less abundant peptides are also selected for fragmentation.

The data can be displayed as a 3-D map with the mass-to-charge ratios (m/z), retention times (RT) and intensities for the observed peptides as axis, together with fragmentation spectra (MS^2) for those peptides that were selected during any of the data dependent cycles. The intensity of a certain peptide m/z can be plotted along the RT to obtain the corresponding chromatographic peak. The area under this curve (AUC) can be employed to quantify the corresponding peptide. On the other hand, the peptide identification is achieved through its fragmentation spectrum.

The large number of MS^2 spectra generated by the last generations of mass spectrometers requires automated search engines capable of identifying and quantifying the analysed peptides. It is not the aim of this

review to detail the existing algorithms (see [9] for this purpose), but to give a general idea how they work and which kind of data should be expected from them. Briefly, search algorithms aim to explain a recorded MS^2 spectrum by a peptide sequence from a pre-defined database, returning a list of peptide sequences that fit to the experimental data with a certain probability score or false discovery rate (FDR). The databases are normally protein databases translated from genomic data [10], although other strategies like spectral libraries [11] or mRNA databases [12] have been successfully applied. A final step is then required to assemble the identified peptides into proteins, which can be challenging, in particular when dealing with redundant peptides or alternatively spliced proteins [13]. In any of these cases, several strategies have been described to reduce the false discovery rate of such matching approaches both at peptide identification and protein assembling level [14].

This general shotgun/discovery approach leads to the identification of thousands of proteins with a dynamic range of 10^4 - 10^5 [15] within a complete cellular lysate. However, this method presents still two main drawbacks: sensitivity and reproducibility. Normally, complete coverage of proteins and complexes involved in the same signaling pathway or belonging to the same functional family is not achieved. Additionally, reproducibility in protein identification among replicates can vary between 30 and 60% [16,17]. These limitations have been successfully addressed by the so-called targeted proteomics [6]. This approach is based on a general method called selected reaction monitoring (SRM), where predefined peptides at scheduled RT are selected and fragmented, and two or three fragments monitored. Due to the increased scan speed and mass window selectivity of the current mass analyzers, SRM can be simultaneously performed on multiple analytes. This capability lead to the multiplexing of SRMs in a method called multiple reaction monitoring (MRM). The multiplexing capability have been used to quantify several hundreds of proteins in a broad dynamic range, down to proteins present at very low copy number in the cell (~ 50 copies/cell) in the background of the whole range of protein concentration in eukaryotic cells [18,19].

The AUC of the monitored fragments can then be used for quantification. By spiking the peptide mixture with isotopically labelled standard peptides, such targeted approaches can also be used to determine absolute rather than relative quantitation levels of proteins [20] or post-translational modifications [21]. However, as previous knowledge about the proteins is required, such targeted approaches are usually performed in combination or subsequent to a shotgun approach. Similarly to the genomic data, shot gun proteomic studies can also be uploaded to dedicated proteome repositories [22], which can also be

used for database searching. The cooperation of the largest repositories PRIDE, Proteome Commons and Peptide Atlas within the Proteome Exchange project <http://www.proteomeexchange.org> allow direct access to most of the stored proteomic datasets and provides a highly valuable source for bioinformatics data mining [23-25].

GO Term identification and enrichment analysis

The output of a proteome analysis either in a shotgun approach or a more targeted method is usually a long list of identified factors, that have a probability score and ideally also a quantitative value associated with them. In order to understand and interpret these data and to generate testable hypothesis on the systemic response of the proteome to a challenge, the list has to be further classified and filtered. The first step for a functional analysis of a large protein list is to connect the protein name to a unique identifier. While gene names have been standardized, protein names can differ between different databases and even releases of the same database. Although many of the large databases have been curated throughout the recent years, this can pose quite a bioinformatic challenge and can lead to a substantial loss of information. Several web-based algorithms exist to connect protein names to their corresponding gene names, such as PICR or CRONOS [26];[27]. However some functional databases like the Uniprot knowledge base, Ensembl or the outdated IPI number (International Protein Index)[28-30] can use protein identifiers as input.

A first step for functional interpretation of the resultant protein list is to connect the protein identifier with its associated Gene Ontology terms (<http://www.geneontology.org>, [31]). Introduction of the Gene Ontology helped to overcome the redundancy in terminology for biological processes [32]. Thereby, genes are associated to hierarchically clustered, functional terms that describe the “biological process”, “molecular function” or “cellular component” which have a unique identification number. A specific GO term can be related to more than one parent terms, as long as the whole structure resembles an acyclic graph. This list of terms is not yet complete and changes with new discoveries, making GO terms redundant or obsolete. Another drawback of the use of GO terms for functional annotations is the fact that most (95%) of the GO terms annotations are done computational, while the minority is manually curated and based on experimental details [32]. For single proteins the simplest way to perform a GO term annotation is to look up the corresponding terms with the Amigo tool provided on the GO website [33]. For larger data sets and systematic approaches some database search algorithms for proteomic data such as MaxQuant, Proteome Discoverer and X!tandem [34,35] have implemented a GO-term annotation step. As not all protein entries are fully annotated with the corresponding GO

terms, it is possible to retrieve GO-terms from the closest related protein via BLAST similarity search in the BLAST2GO tool [36].

The first step after GO-term annotation is a GO-term enrichment analysis to compare the abundance of specific GO-terms in the dataset with the natural abundance in the organism or a reference dataset, e.g. different cell lines, inhibitor treatment or growth states [37]. To extract functions that are significantly enriched in one sample over a second dataset, a p-value is calculated based which shows overrepresentation of a specific GO term, thereby it is necessary to cluster related GO-terms. This calculation can be done by most of the previously mentioned programs, but there is a plethora of other, mostly web-based software tools available ([38] http://neurolex.org/wiki/Category:Resource:Gene_Ontology_Tools). For instance, the DAVID and Babelomics software resources are often mentioned when it is necessary to analyze large gene list but currently there are more than 60 tools calculating GO term enrichment [38-40]. Most of these tools can be classified into three different types of enrichment algorithms, with singular enrichment analysis (SEA) being the most simple algorithms that test one annotation term at a time for a list of interesting genes [41]. GStat, BinGO, or EasyGO are based on SEA algorithms. More sophisticated algorithms are gene set enrichment algorithms (GSEA) that take all genes of analysis into account, not only gene with significant change of abundance. Nevertheless, GSEA requires a quantitative measurement to rank the genes and is used in GSEA/P-GSEA and Gene Trail. Finally, modular enrichment analysis (MEA) include relationships between annotation terms which prevents loss of important biological correlations due to lacking relationships and reduces redundancy [41]. Those programs are not only limited to GO term enrichment, but they have also modules to search for protein networks (see below), convert protein identifiers, as well as link to further information and publications that substantiate the observed gene function. Especially the DAVID software resources offer a plethora of other tools for instance for gene and annotation term clustering, mapping of genes to pathways and diseases as well as advanced statistics. A second important choice for the result of GO term enrichment is the reference dataset, which is either predefined by the tool, for instance all genes of the organism, or can be selected manually (all identified proteins) [42]. Weinert et al. have applied the DAVID GO term enrichment algorithm to study conservation of acetylation sites between human and drosophila from the extracted GO-terms of acetylated proteins [43]. In their study, they showed the conservation of protein acetylation in the respiratory chain, translational processes, but also in ubiquitinating

enzymes. Bates et al. could show that the Abl-kinase dependent reprogramming of B-cells is to a major part post-transcriptionally regulated, by comparing the abundance of mRNA levels with protein abundance upon imatinib inhibitor treatment [44]. Despite the usefulness of GO terms for a functional annotation and filtering of large proteomic data sets the assignment is highly dependent on the algorithm used for annotation. Recently, fourteen GO enrichment algorithms have been tested on the same dataset. The obtained results showed a rather high discrepancy for p-values of certain GO terms [42].

Pathway analysis

A pathway describe the series of chemical reactions in the cell that lead to an observable biological effect. Proteins involved in the chemical reaction and those that have regulatory influence are combined in so-called pathway databases. Similarly to the previously described GO term enrichment analysis, protein or gene lists can also be scrutinized for pathway abundances which might be more meaningful because it moves the data interpretation away from the gene-centric view towards the identification of functional biological processes. Furthermore, functionally independent proteins can share some GO term associations, for instance for very general terms such as “binding” or “cytoplasmic”. A high number of resources and databases is available to extract pathway constraints from biological data (Figure 1). Comprehensive pathway databases such as KEGG, Reactome, Ingenuity Pathway Knowledge Base or BioCarta include a high number of diverse interaction data, which could arise from intracellular reactions such as metabolism or signaling pathways, genetic interactions or drug development studies [45-47]. Apart from the comprehensive resources, highly specific databases have been developed for signal transduction processes such as PANTHER, GenMAPP or PID [48-50]. Recently, several databases were created which comprise pathways active in cancer. Such databases like Netpath [51], should help to identify cancer relevant proteins and genes from a complex dataset. In fact, public databases share a high degree of connectivity, allowing rapid distribution of novel findings. A comprehensive list of more than 300 pathway and interaction data resources can be found on the *pathguide* website <http://pathguide.org>[52]. Nowadays enrichment analyses are available with almost all pathway database resources, so that a list of significantly altered proteins, with regard to abundance and/or post-translational modifications, is sufficient to extract data on pathway abundance. However, similar to the GO term annotation the identification of pathways affected under certain conditions is highly dependent on the algorithm. Müller and colleagues published a comparison of the Ingenuity Pathway Analysis (IPA) software and String for the analysis of several artificial datasets [53].

The tested datasets consisted of core proteins and associated proteins of 5 different pathways, Wnt, App, and Ins signaling, mitochondrial apoptosis as well as tau phosphorylation, respectively, which were retrieved from literature mining and a set of background proteins from proteomic analysis of HEK293 cells that that were falsely annotated as significantly regulated proteins in several repeats. They reported similar results for both pathway analysis algorithms, but also that neither algorithm could reach a sufficient p-value for reliable pathway enrichment. Additional features in IPA, such as annotation of protein family and localization, might help the experienced researcher to identify falsely annotated protein hits.

Analysis of protein-protein-interactions

The majority of proteins do not act as independent entities. They form either transient or stable complexes with other proteins that act as scaffolds or regulate the protein activity. A protein can be involved in multiple complexes of varying composition and to completely understand a biological system it is necessary to analyze the abundant protein complexes as well as the conditions that lead to their formation or dissociation. Information on protein interactions in complexes is deposited in interaction databases such as MINT, BioGRID, IntAct or HRPD [54-57], associated with the biological process in which they are functionally important. Not all annotated interactions in public databases are based on experimental observations. Dependent on the database used one can find a rather high percentage of predicted interactions and interactions based on literature mining such as STRING or iRefWeb [37,58,59]. For this purpose, a variety of literature mining tools to screen PubMed abstracts has been developed of which chilibot and sciminer are most favored [60,61]. These interactions are the result of sophisticated algorithms that are trained on the existing set of protein-protein interactions. Furthermore, most large interaction databases have implemented simple algorithms that allow mapping of interaction proteins on the resource website.

Protein interactions are often displayed as large interaction networks illustrating the high degree of connectivity and the presence of promiscuous hub proteins. A widely used resource for interaction data is STRING, which is not only a database itself, but connects to several other data resources to and is therefore also capable of literature mining [59,62]. Further, STRING is also capable of drawing simple protein networks based on the provided gene list and the available interactions in its databases. Cytoscape has evolved as a powerful graphical tool to draw interaction networks of high complexity and for incorporation and comparison of datasets from different experimental procedures. Cytoscape has only limited information stored, but interconnects excessively to other

databases to obtain information. Recently, EnrichNet was launched, a web-based platform integrating pathway and interaction analysis in 6 different databases (KeGG, BioCarta, Gene Ontology, Reactome, Wiki and NCI pathways) with functional associations and connecting these data with molecular function (Interpro) and protein complex information (Corum) [63]. This tool creates pathway lists and highly interactive function maps, which can also be downloaded and visualized in cytoscape. A study of the targets of cullin-ring dependent ubiquitination revealed that a large fraction of the observed proteins become modified upon activity of the SCF complex [64]. Analysis of the obtained list of SCF regulated proteins by cytoscape revealed a high degree of interconnectivity.

Protein domain and motif analysis

When working in not yet or just recently-sequenced organisms, data bases might not contain the complete set of protein descriptions. Similarly, proteins of unknown function might also be identified from highly curated databases of well studied organisms. Those proteins often lack the previously described information on interactions and pathway affiliations so that they would not be found in such studies. To learn more about the function of those proteins and how they interact with members of certain pathways, it is helpful to analyze their amino acid sequence for specific folds of protein domains or for motifs for post-translational modifications. The simplest analysis represents a BLAST search against the database of known protein sequences to find if proteins with similar amino acid sequences have been described in other organisms [65]. Further, the amino acid sequence can be analyzed by programs such as Pfam, Interpro, SMART or also DAVID, to learn if the identified protein shares a specific protein fold with other proteins [39,66-68]. These algorithms apply hidden Markov models (HMMs) to classify proteins on basis of their amino acid sequence and predict the occurrence of a specific protein domain. Knowing about the abundance of a specific fold, could help to implement unknown proteins into biological networks. Secondly, algorithms such as MotifX or PhosphoMotif Finder analyze the sequence environment of post-translational modification sites [69,70], thereby reporting enrichment of certain amino acid motifs which can help to identify the modifying enzyme.

Conclusions

The development of methods to systematically study all proteins in a cell and their subsequent functional annotation opens up new pathways of research. In the future it is very likely that such studies will uncover new principles of how biological systems operate hopefully leading to an improved treatment of human pathologies.

Over the last ten years the analytical hardware has reached a level of sophistication of a more mature scientific field. However, the bioinformatic interpretation and the processing of the data are still in its infancy. Besides reliable and robust algorithms, international standards for data processing and deposition as well as their interpretation have to be developed and agreed upon in order to unleash the full potential of proteomic research.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AS, IF and AI wrote subsections of the paper and were involved in assembling the manuscript

Declarations

The publication costs for this article were partly funded by a grant from the European Union (STATEGRA, 257082) and partly supported by COST-BMBS, Action BM1006 "next Generation Sequencing Data Analysis Network", SeqAhead.

This article has been published as part of BMC Systems Biology Volume 8 Supplement 2, 2014: Selected articles from the High-Throughput Omics and Data Integration Workshop. The full contents of the supplement are available online at <http://www.biomedcentral.com/bmcsystbiol/supplements/8/S2>.

Published: 13 March 2014

References

1. Wiśniewski JR, Mann M: Consecutive Proteolytic Digestion in an Enzyme Reactor Increases Depth of Proteomic and Phosphoproteomic Analysis. *Anal Chem* 2012, **84**(6):2631-2637.
2. Lee YH, Tan HT, Chung MCM: Subcellular fractionation methods and strategies for proteomics. *Proteomics* 2010, **10**(22):3935-3956.
3. Altaalar AFM, Heck AJR: Trends in ultrasensitive proteomics. *Current Opinion in Chemical Biology* 2012, **16**(1-2):206-213.
4. Johnson H, Eyers C: Analysis of Post-translational Modifications by LC-MS/MS. In *LC-MS/MS in Proteomics. Volume 658*. Humana Press/Cutillas PR, Timms JF 2010:93-108.
5. Hein MY, Sharma K, Cox J, Mann M: Chapter 1 - Proteomic Analysis of Cellular Systems. *Handbook of Systems Biology* San Diego: Academic Press; 2013, 3-25.
6. Picotti P, Aebersold R: Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Meth* 2012, **9**(6):555-566.
7. Kalli A, Smith GT, Sweredoski MJ, Hess S: Evaluation and Optimization of Mass Spectrometric Settings during Data-Dependent Acquisition Mode: Focus on LTQ-Orbitrap Mass Analyzers. *J Proteome Res* 2013.
8. Graumann J, Scheltema RA, Zhang Y, Cox J, Mann M: A Framework for Intelligent Data Acquisition and Real-Time Database Searching for Shotgun Proteomics. *Molecular & Cellular Proteomics* 2012, **11**(3).
9. Hoopmann MR, Moritz RL: Current algorithmic solutions for peptide-based proteomics data generation and identification. *Current Opinion in Biotechnology* 2013, **24**(1):31-38.
10. Kumar C, Mann M: Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Letters* 2009, **583**(11):1703-1712.
11. Lam H: Building and Searching Tandem Mass Spectral Libraries for Peptide Identification. *Molecular & Cellular Proteomics* 2011, **10**(12).
12. Chalkley RJ, Bandeira N, Chambers MCC, JS, Deutsch EW, Kapp EAM, Lam HNH, Neubert TA, Sun RX, Vitek O, Weintraub ST: iPRG 2013: Using RNA-Seq data for Peptide and Protein Identification. *ABRF Poster* 2013 2013.
13. Nesvizhskii AI, Aebersold R: Interpretation of Shotgun Proteomic Data: The Protein Inference Problem. *Molecular & Cellular Proteomics* 2005, **4**(10):1419-1440.

14. Nesvizhskii AI: **A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics.** *J Proteomics* 2010, **73**(11):2092-2123.
15. Nagaraj N, Alexander Kulak N, Cox J, Neuhäuser N, Mayr K, Hoerning O, Vorm O, Mann M: **System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap.** *Molecular & Cellular Proteomics* 2012, **11**(3).
16. Liu H, Sadygov RG, Yates JR: **A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics.** *Anal Chem* 2004, **76**(14):4193-4201.
17. Tabb DL, Vega-Montoto L, Rudnick PA, Varyath AM, Ham AJL, Bunk DM, Kilpatrick LE, Billheimer DD, Blackman RK, Cardasis HL, et al: **Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography-Tandem Mass Spectrometry.** *J Proteome Res* 2009, **9**(2):761-776.
18. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R: **Full Dynamic Range Proteome Analysis of *S. cerevisiae* by Targeted Proteomics.** *Cell* 2009, **138**(4):795-806.
19. Picotti P, Clement-Ziza M, Lam H, Campbell DS, Schmidt A, Deutsch EW, Rost H, Sun Z, Rinner O, Reiter L, et al: **A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis.** *Nature* 2013, **494**(7436):266-270.
20. Klein A, Israel L, Lackey SW, Nargang FE, Imhof A, Baumeister W, Neupert W, Thomas DR: **Characterization of the insertase for beta-barrel proteins of the outer mitochondrial membrane.** *J Cell Biol* 2012, **199**(4):599-611.
21. Kettenbach AN, Rush J, Gerber SA: **Absolute quantification of protein and post-translational modification abundance with stable isotope-labeled synthetic peptides.** *Nat Protoc* 2011, **6**(2):175-186.
22. Riffle M, Eng J: **Proteomics data repositories.** *Proteomics* 2009, **9**(20):4653-4663.
23. Desiere F, Deutsch EW, Nesvizhskii AI, Mallik P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, et al: **Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry.** *Genome Biology* 2004, **6**.
24. Falkner JA, Ulintz PJ, Andrews PC: **A Code and Data Archival and Dissemination Tool for the Proteomics Community.** *American Biotechnology Laboratory* 2006.
25. Vizcaino AJ, Côté RG, Csordas A, Dianas JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J, et al: **The Proteomics Identifications (PRIDE) database and associated tools: status in 2013.** *Nucleic Acids Res* 2012, **41**(D1):D1063-D1069.
26. Wein SP, Côté RG, Dumousseau M, Reisinger F, Hermjakob H, Vizcaino AJ: **Improvements in the protein identifier cross-reference service.** *Nucleic Acids Res* 2012, **40**:W276-W280.
27. Waegle B, Dunger-Kaltenbach I, Fobo G, Montrone C, Mewes HW, Ruepp A: **CRONOS: the cross-reference navigation server.** *Bioinformatics* 2009, **25**(1):141-143.
28. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R: **The International Protein Index: An integrated database for proteomics experiments.** *Proteomics* 2004, **4**(7):1985-1988.
29. Consortium TU: **Reorganizing the protein space at the Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2011, **40**(D1):D71-D75.
30. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, et al: **Ensemble 2012.** *Nucleic Acids Res* 2012, **40**(D1):D84-D90.
31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology.** The Gene Ontology Consortium. *Nature genetics* 2000, **25**(1):25-29.
32. Yon Rhee S, Wood V, Dolinski K, Draghici S: **Use and misuse of the gene ontology annotations.** *Nat Rev Genet* 2008, **9**(7):509-515.
33. Huang D, Sherman B, Lempicki R: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protocols* 2008, **4**(1):44-57.
34. Cox J, Mann M: **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.** *Nat Biotech* 2008, **26**(12):1367-1372.
35. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra.** *Bioinformatics* 2004, **20**(9):1466-1467.
36. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics (Oxford, England)* 2005, **21**(18):3674-3676.
37. Malik R, Dulla K, Nigg E, Körner R: **From proteome lists to biological impact - tools and strategies for the analysis of large MS data sets.** *Proteomics* 2010, **10**:1270-1283.
38. Huang D, Sherman B, Lempicki R: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37**(1):1-13.
39. Jiao X, Sherman B, Huang D, Stephens R, Baseler M, Lane C, Lempicki R: **DAVID-WS: A Stateful Web Service to Facilitate Gene/Protein List Analysis.** *Bioinformatics* 2012, doi:10.1093/bioinformatics/bts251.
40. Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, Tárrega J, Pascual-Montano A, Nogales-Cadenas R, Santoyo J, et al: **Babelomics: an integrative platform for the analysis of transcriptomics, proteomics, and genomic data with advance functional profiling.** *Nucleic Acids Res* 2010, **38**(suppl 2):W210-W213.
41. Tipney H, Hunter L: **An introduction to effective use of enrichment analysis software.** *Human Genomics* 2010, **4**(3):202-206.
42. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems.** *Bioinformatics* 2005, **21**(18):3587-3595.
43. Weinert B, Wagner S, Horn H, Henriksen P, Liu W, Olsen J, Jensen L, Choudhary C: **Proteome-Wide Mapping of the Drosophila Acetylome Demonstrates a High Degree of Conservation of Lysine Acetylation.** *Sci Signal* 2011, **4**(183):ra48.
44. Bates J, Salzman J, May D, Garcia P, McIntosh M, Schlissel M, Brown P: **Extensive gene-specific translational reprogramming in a model of B cell differentiation and Abl-dependent transformation.** *PLoS one* 2012, **7**(5).
45. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, et al: **Reactome: a database of reactions, pathways and biological processes.** *Nucleic Acids Res* 2011, **39**(Database):D691-D697.
46. Kanehisa M, Goto S: **Kyoto encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 2000, **28**(1):27-30.
47. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular datasets.** *Nucleic Acids Res* 2012, **40**(D1):D109-D114.
48. Mi H, Guo N, Kejariwal A, Thomas PD: **PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways.** *Nucleic Acids Res* 2007, **35**:D247-D252.
49. Salomonis N, Hanspers K, Zamboni AC, Vranizan K, Lawlor SC, Dahlquist KD, Doninger SW, Stuart J, Conklin BR, Pico AR: **GenMAPP 2: new features and resources for pathway analysis.** *BMC bioinformatics* 2007, **8**.
50. Schaefer CF, Anthony K, Krupa S, Buchhoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database.** *Nucleic Acids Res* 2009, **37**(D1):D674-D679.
51. Kandasamy K, Mohan S, Raju R, Keerthikumar S, Kumar G, Venugopal A, Telikicherla D, Navarro D, Mathivanan S, Pecquet C, et al: **NetPath: a public resource of curated signal transduction pathways.** *Genome Biology* 2010, **11**(1):R3.
52. Bader G, Cary M, Sander C: **Pathguide: a pathway resource list.** *Nucleic Acids Res* 2006, **34**(Database):D504-D506.
53. Müller T, Schrötter A, Loosse C, Helling S, Stephan C, Ahrens M, Uzkoreit J, Eisenacher M, Meyer HE, Marcus K: **Sense and Nonsense of Pathway Analysis Software in Proteomics.** *J Proteome Res* 2001, **10**(12):5398-5408.
54. Chatranyamontri A, Ceol A, Montecchi Palazzi L, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular Interaction database.** *Nucleic Acids Res* 2007, **35**(Database):D572-D574.
55. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, et al: **The IntAct molecular interaction database in 2012.** *Nucleic Acids Res* 2012, **40**(D1):D841-D846.
56. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al: **Human Protein Reference Database-2009 update.** *Nucleic Acids Res* 2009, **37**(Database):D767-D772.
57. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Res* 2006, **34**(suppl 1):D535-D539.
58. Turner B, Razick S, Turinsky A, Vlasblom J, Crowdy E, Cho E, Morrison K, Donaldson I, Wodak S: **iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence.** *Database: the journal of biological databases and curation* 2010, **2010**.

59. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mehning C, et al: **STRING v9.1: protein-protein interaction networks, with increased coverage and integration.** *Nucleic Acids Res* 2013, **41**(D1):D808-D815.
60. Chen H, Sharp BM: **Content-rich biological network constructed by mining PubMed abstracts.** *BMC Bioinformatics* 2004, **5**.
61. Hur J, Schuyler AD, States DJ, Feldmann EL: **Sciminer: web based literature mining tool for target identification and functional enrichment analysis.** *Bioinformatics* 2009, **25**(6):838-840.
62. Snel B, Lehmann G, Bork P, Huyen MA: **STRING: a web-server to retrieve and display the repeatedly occurring neighborhood of a gene.** *Nucleic Acids Res* 2000, **28**(18):3442-3444.
63. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A: **EnrichNet: network-based gene set enrichment analysis.** *Bioinformatics* 2012, **28**(18):i451-i457.
64. Emanuele M, Elia A, Xu Q, Thoma C, Izhar L, Leng Y, Guo A, Chen YN, Rush J, Hsu P, et al: **Global Identification of Modular Cullin-RING Ligase Substrates.** *Cell* 2011, **147**(2):459-474.
65. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.
66. Punta M, Coggill P, Eberhardt R, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40**(Database):D290-D301.
67. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, et al: **Interpro in 2011: new developments in the family and domain prediction database.** *Nucleic Acids Res* 2012, **40**(D1):D306-D312.
68. Letunic I, Doerks T, Bork P: **SMART 7: recent updates to the protein domain annotation resource.** *Nucleic Acids Res* 2012, **40**(D1):D302-D305.
69. Amanchy R, Periaswamy B, Mathivanan S, Reddy R, Tattikota S, Pandey A: **A curated compendium of phosphorylation motifs.** *Nat Biotech* 2007, **25**(3):285-286.
70. Chou M, Schwartz D: **Biological Sequence Motif Discovery Using motif-x.** *Current Protocols in Bioinformatics* 2002.

doi:10.1186/1752-0509-8-S2-S3

Cite this article as: Schmidt et al.: Bioinformatic analysis of proteomics data. *BMC Systems Biology* 2014 **8**(Suppl 2):S3.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

