REVIEW

Bioinformatics approaches for the functional interpretation of protein lists: From ontology term enrichment to network analysis

Kris Laukens[1,2], *Stefan Naulaerts*[1,2] *and Wim Vanden Berghe*[3]

[1] Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan, Antwerp, Belgium
[2] Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp / Antwerp University Hospital, Antwerp, Belgium
[3] PPES, Department of Biomedical Sciences, University of Antwerp, Wilrijk, Belgium

The main result of a great deal of the published proteomics studies is a list of identified proteins, which then needs to be interpreted in relation to the research question and existing knowledge. In the early days of proteomics this interpretation was only based on expert insights, acquired by digesting a large amount of relevant literature. With the growing size and complexity of the experimental datasets, many computational techniques, databases, and tools have claimed a central role in this task. In this review we discuss commonly and less commonly used methods to functionally interpret experimental proteome lists and compare them with available knowledge. We first address several functional analysis and enrichment techniques based on ontologies and literature. Then we outline how various types of network and pathway information can be used. While the problem of functional interpretation of proteome data is to an extent equivalent to the interpretation of transcriptome or other ''omics'' data, this paper addresses some of the specific challenges and solutions of the proteomics field.

Received: June 30, 2014
Revised: October 16, 2014
Accepted: November 24, 2014

Keywords:
Bioinformatics / Data interpretation / Functional bioinformatics / Overrepresentation analysis / Pathways / Protein networks

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

Most proteomic analyses invariably lead to lists of protein identities. The interpretation of such lists presents a significant challenge in most studies. In many cases these lists are only partially interpreted in the context of the research question or hypothesis. Very often expert interpretation based on digesting large amounts of relevant literature plays the most important role. However many bioinformatics approaches, tools, and databases that can facilitate, automate, and standardize this task are now available. Selecting (and reporting) a suitable approach inherently contributes to the quality, reproducibility, and consistency of the results and (in particular) their interpretation.

Practical information for the selection of approaches to interpret protein lists is rather scattered. Existing efforts to review this task mostly focus on the transcriptomics field [1–3]. The field of proteome data analysis has been broadly reviewed before. In particular, a range of papers addressed upstream data analysis steps such as spectrum processing, identification, and quantification [4–9]. In 2006 Lisacek and colleagues presented a comprehensive overview of the opportunities and challenges of experimental proteome data in the context of systems biology, network analysis, and functional interpretation [10]. More recent papers addressed these challenges as a part of a broader MS proteomics bioinformatics review [6, 7, 9]. We feel that a comprehensive and updated survey of the functional analysis of proteomics results is timely.

Correspondence: Professor Kris Laukens, Advanced Database Research and Modelling (ADReM), University of Antwerp, Middelheimlaan 1 B-2020, Antwerp, Belgium
E-mail: kris.laukens@uantwerpen.be

Colour Online: See the article online to view Figs. 1 and 2 in colour.

This paper aims to consolidate practically relevant approaches and solutions to the problem of how to make functional sense of protein identification data and highlights, where relevant and possible, how they are related to or different from analogous ''omics'' disciplines.

2 From transcriptome to proteome data interpretation

In several regards the functional interpretation of protein lists is similar to the analysis of differential transcriptomics data, for which an extensive range of interpretation tools have been developed ahead of equivalent proteomics developments [11–17], and many of these tools can also be used for proteome data. However proteins constitute the functional machinery of the cell and operate in their own specific spatiotemporal and interaction context. The unique features and limitations of the proteomics field require a careful selection and application of techniques and demand a critical interpretation. Protein lists are particularly different from other ''omics'' lists as a consequence of a combination of the intrinsic biomolecular properties of the proteome level and the limitations of the analytical technology.

A first major biological challenge is the dynamic range of the proteome, which stretches at least seven orders of magnitude, compared to a mRNA dynamic range of three or four orders [18, 19]. Despite the power of most contemporary MS-based setups, an immense fraction of the proteome remains hidden in most studies, resulting in a much higher number of false negatives than in most transcriptome studies. Furthermore, the complexity of a proteome is much larger than the complexity of a transcriptome, due to the extensive range of (often substoichiometric) post-translational processing possibilities that enormously diversify the potential structures of each translation product. Third, the proteome level layer is inherently characterized by a large number of context-specific interactions. In some types of experiments these interactions are entirely ignored, but in others they are specifically the scope of the experimental setup. Intentionally or not, under nondenaturing conditions any targeted purification of proteins can lead to the indirect copurification of many of these interactors that are relevant for the interpretation of most proteome data. In addition several analytical limitations compromise the composition of proteome lists in comparison to other ''omics'' lists. Most proteomics workflows depend on separation and identification strategies, which carry an inherent selectivity toward certain classes of proteins based on features such as their abundance, size, and other (physicochemical) characteristics. They cause an increased variability and an incomplete, biased coverage of the proteome under investigation, especially in comparison to hybridization- or next-generation sequencing-based technologies that dominate the transcriptomics field. Compared to transcriptomics, most proteomics studies also sample a smaller fraction of the full proteome. The consequently smaller number of "interesting" (e.g. differential) proteins can have negative repercussions on the statistical power of functional analysis techniques. More specifically, larger gene lists have a higher sensitivity toward slightly enriched or more specific terms in enrichment analysis workflows (further discussed below) [16]. Biologically meaningful and interesting enrichments may therefore remain below detection thresholds in small protein lists.

To avoid that the abovementioned biological and analytical complications lead to a bias in the downstream interpretation, they should as much as possible be accounted for when performing functional analyses [20]. In many cases we do not know to what extent this affects the results obtained with tools that are often primarily designed for transcriptome data analysis. Nevertheless there are two levels at which the proteome-specific challenges can be studied. Well-established reference proteome repositories [21, 22] can be evaluated as background datasets in the functional analyses to anticipate the intricacies of the protein layer (and is further discussed in the enrichment analysis section of this paper). In the absence of comprehensive reference proteome data, a plethora of machine learning-based tools exist to predict which subset of the theoretical proteome is likely to exhibit certain characteristics (e.g. modifications, interactions, detectability) relevant for the study [23], and to what extent the functional properties discussed further in this paper correlate to these characteristics. As an example, models to predict the likelihood that a given protein will be detected in an experimental workflow can be employed to predict the expected proteome and analyze in silico the effect of physicochemical biases to the functional representativeness of this expected proteome [23].

3 Protein list data

Irrespective of the employed technology, proteome studies typically generate a list of protein entries, with or without quantitation data. In the following paragraphs we will briefly highlight three major classes of protein lists, discuss some of their data properties and the primary challenges of their subsequent functional analysis. A high-level overview of these datasets and the functional analysis is presented in Fig. 1.

The most simple type of proteomics results are the *reference proteomes*, in which for a given biological entity, the entire protein population is analyzed, resulting in a number of identified proteins that are supposedly representative. They have been the scope of many early proteome studies, typically using various 2D-PAGE profiling methods [24, 25] that often resulted in web-accessible databases. Reference proteomes covering a wide range of species and tissues using the full range of profiling methods, are nowadays often published in this journal in the form of a "dataset brief" paper, whereas several large-scale efforts are currently culminating in comprehensive organism-wide reference proteome resources [21, 22]. The identified proteins provide evidence for the

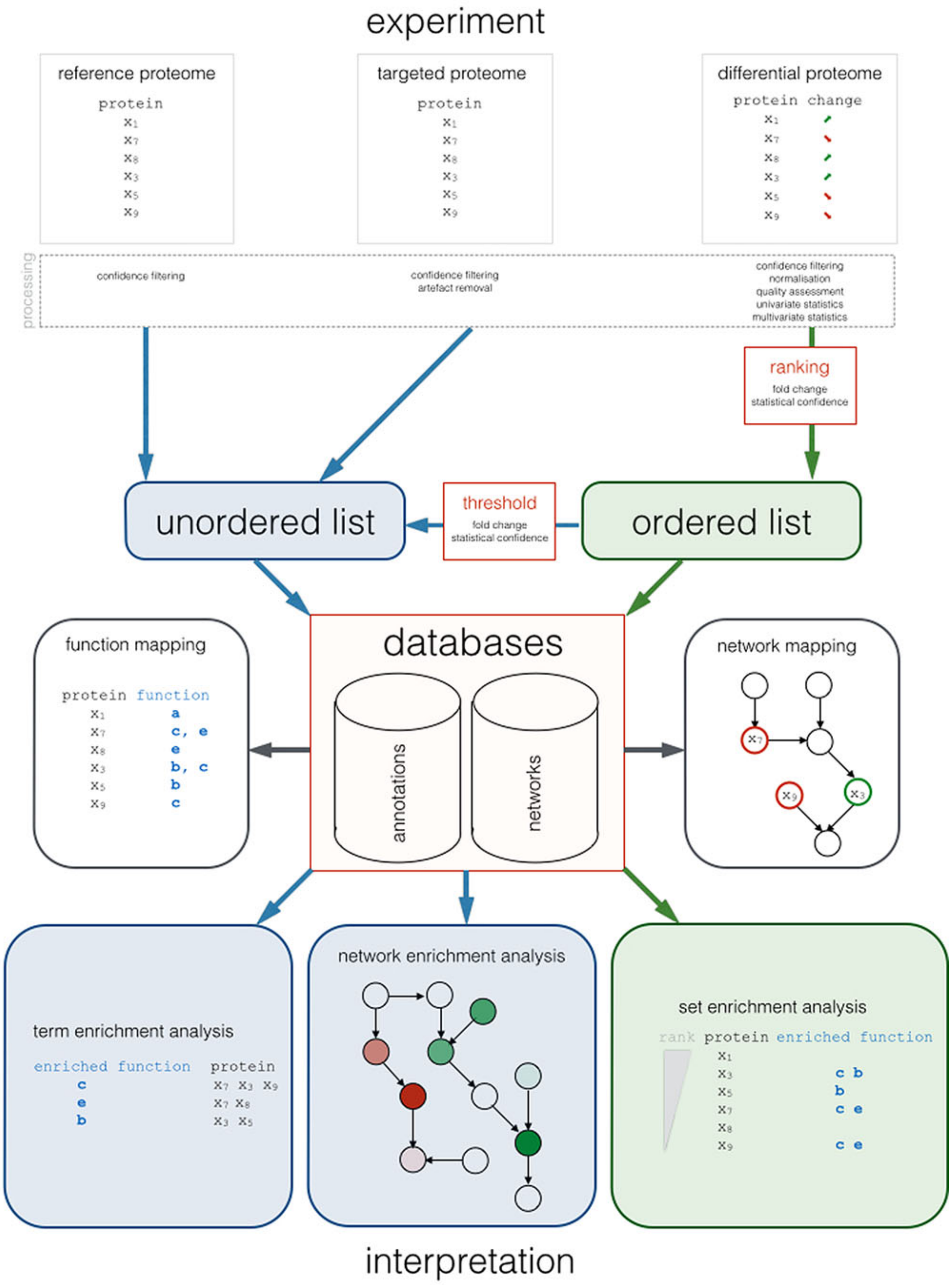


Figure 1. Overview scheme of typical workflows for the analysis of protein lists, from experiment (top) to functional interpretation (bottom). We can distinguish three different types of protein lists (top, with x_n corresponding to a single identified protein entry, and green or red arrows representing differential up- or dow-nregulation, respectively), that are either ordered (green box and arrows) or unordered (blue box and arrows). Typical data processing steps are displayed in the "processing" box, as well as in the "threshold" box to convert ordered to unordered lists. The interpretation is based on various databases (see Table 1 for a comprehensive overview) that contain annotation (e.g. GO terms) or network (e.g. KEGG) data (middle). The subsequent interpretation steps (arrows and boxes) are also colored according to whether they are based on unordered (blue), or ordered (green) lists, or whether both types can be used (black). Node colors in the network representation correspond to the up- (green) or downregulation of protein entries, and assigned or enriched functions are labeled a, b, . . . e.

transcription and translation of genes and can be relevant to improve gene annotations (i.e. proteogenomics). More abundant proteins are likely better represented in such an analysis, and a degree of selection bias towards certain physicochemical properties in relation to the analytical methodology can usually not be excluded. For reference proteomes, data preprocessing is often limited to determining a suitable identification score threshold to retain only confident identifications while limiting the number of false positives, as reviewed before [6, 8, 26]. The downstream analysis typically tries to establish which functions are present in the dataset, how representative the dataset is compared to the full genome, which subcellular compartments are represented, etc.

A second type of protein lists is generated with selective techniques to extract specific subsets of the proteome, i.e. targeted proteomes with particular properties of interest. In interactomics this property may be an observed interaction with a bait protein [27–29]. In chemoproteomics this depends on protein interactions with an affinity tagged chemical [30]. In other cases proteins with specific structural properties (e.g. a PTM) [31] or proteins that are substrates of specific enzymes (e.g. proteases) [32] are targeted. In some cases a data processing step consists of excluding nonspecifically and or indirectly copurified housekeeping, cytoskeleton or artifact proteins. Similar to the reference proteomes, an identification score threshold usually determines which proteins are retained for further interpretation. The functional analysis of the resulting lists can be diverse and is largely imposed by the study goals. For example, in the targeted analysis of substrates for a given class of enzymes (proteases, kinases, ...) the goal might be to reveal which biological functions are affected. If the analysis of interactors is the scope, an important goal could be mapping the observed interactors over the known network data.

The third type of experiment is most relevant for this review as it uses label-based or label-free quantitative approaches to generate differential proteome data [33,34]. Three tasks can be typically distinguished in quantitative proteomics [35]. Class comparison aims to distinguish how abundance levels of individual proteins differ between predefined groups. The observation of differential proteins can be translated into functional biological insights. Class prediction aims to identify sets of proteins that allow distinguishing multiple biological sample types (e.g. healthy versus diseased, life cycle stage) and thus pertains the discovery of effective biomarkers. Computational approaches are then applied to learn which features are robust enough to correctly predict the class of an unknown sample based on its proteome. A third objective could be class discovery, i.e. using proteome profiles to identify subtypes, for example, of diseases based on their proteome profiles. Even though functional interpretation is often not the primary goal of class prediction and class discovery, it can still be relevant to explain not only which proteins differentiate the classes, but also why this is the case. The elucidation of the functional context can also be directly relevant for the classification tasks [36].

Differential proteomics builds on reliable protein quantitation. This can be achieved using a range of techniques [34,35,37], but differential protein identification always needs to be accompanied by appropriate experimental design, data processing, and statistical analysis [8,35,37–40]. An optimal experimental design is crucial to maximize the gain of trustworthy information with the available resources [35]. Data processing includes normalization, the investigation of data quality, and statistical analysis to test whether proteins are differential with sufficient confidence. For the subsequent functional interpretation it is always crucial to choose an optimal threshold that balances sensitivity and specificity.

In the functional analysis of differential proteome data with computational tools we can make a distinction between ordered and unordered protein lists. In an ordered list the proteins are ranked according to a predefined metric (e.g. the degree of up- or down-regulation or its confidence), while in unordered list such a rank is absent, i.e. the presence of a protein on the list is determined by whether it fulfills a predefined criterion (such as exceeding one of the above threshold criteria). An ordered list can always be converted into an unordered list by applying a cutoff threshold. In contrast to differential protein lists, most reference and targeted proteome studies result in unordered lists.

4 Singular interpretation of protein list entries

To interpret proteomics data sets, a typical and still common approach consists of browsing the list and interpreting each observation separately. This task is often accompanied by the digestion of extensive amounts of literature relating to the protein list entries. More accessible than literature, very valuable databases exist to learn more about each individual protein function. The most comprehensive, easily accessible expert-curated resource is Uniprot [41], but most other annotation databases presented in Table 1 can be queried with individual protein accession codes. This allows extending a list of protein identifications with the associated functions for each protein entry (illustrated in Fig. 1).

Whereas such an interpretation per protein has its merits, it has serious limitations. Comparing experimental observations to manually collected knowledge could introduce bias. Due to the complexity of the data the interpretation often culminates into a lengthy discussion section that lists all the potential functional implications of each observed protein. This makes many publications unnecessarily complex and hard to compare with other studies. This approach also carries the danger of skewing the interpretation toward what fits the underlying assumptions and hypotheses, and there is no guarantee that two researchers who interpret the same list end up with the same conclusions. Especially potentially important hidden relationships between the members of a protein list may be missed.

In the next sections a range of computational, database-driven approaches, and tools will be presented that can be used for a more comprehensive, uniform, and objective analysis of experimental results, to allow for consistent hypothesis generation.

5 Enrichment analysis based on controlled vocabularies

Various functional databases exist that contain an experimentally proven or otherwise inferred link between the gene or protein and a specific "function." Usually these functions are part of a controlled vocabulary: they have a clearly described meaning that is defined and supervised by domain specialists. The most well-known functional database is the gene ontology (GO) [42]. It describes the function of genes or proteins in three broad categories: biological process, molecular function, and cellular component. Each of these categories is represented as a tree structure, in which the nodes are terms, and the branches define their relationships. In annotation databases, individual genes are linked to terms of this tree.

Table 1. Overview of common data resources for the functional interpretation of protein lists

Content	Name	Reference	Availability[a)]	Link
Interaction databases				
Protein interactions	BioCarta	[173]	Mixed	http://www.biocarta.com/
	BioGrid	[134]	Public	http://thebiogrid.org
	DIP	[132]	Public	http://bond.unleashedinformatics.com/Action?
	HPRD	[133]	Public	http://www.hprd.org
	IntAct	[135]	Public	http://www.ebi.ac.uk/intact/
	MINT	[136]	Public	http://mint.bio.uniroma2.it/mint/Welcome.do
	String	[142]	Public	http://string-db.org
Pathways	BIND Translation	[174]	Public	http://baderlab.org/BINDTranslation
	BioCyc	[120]	Public	http://biocyc.org
	Ingenuity IPA		Commercial	http://www.ingenuity.com/products/ipa
	INOH	[175]	Public	http://inoh.hgc.jp/download.html
	KEGG	[72]	Mixed	http://www.genome.jp/kegg/kegg2.html
	NetPro		Commercial	http://www.molecularconnections.com/home/en/home/products/netPro
	Pathway Commons	[121]	Public	http://www.pathwaycommons.org/about/
	Panther Pathways	[122]	Public	http://www.pantherdb.org/pathway/
	Reactome	[118]	Public	http://www.reactome.org
	WikiPathways	[119]	Public	http://wikipathways.org/index.php/WikiPathways
Ontology databases				
Functions	Gene Ontology	[42]	Public	http://www.geneontology.org
Protein evolution	Protein Ontology	[143]	Public	http://pir.georgetown.edu/pro/pro.shtml
Aggregated databases				
	UniProt	[41]	Public	http://www.uniprot.org
	NCBI Protein	[176]	Public	http://www.ncbi.nlm.nih.gov/protein

a) Availability under the assumption of academic, noncommercial use.

Gene ontology is a tremendously valuable resource for protein list interpretation but has also some general limitations. Since functional annotation is labor-intensive these databases are never entirely up to date and even for the most important model organisms they do not have full coverage [43]. Therefore they are often supplemented with less trustworthy evidence inferred from electronic sources. Functional annotations tend to be biased to certain areas [44]. Annotations are still black-white simplifications of an often less clear-cut and more complex reality. For an in-depth coverage on issues such as time-delayed curation, imprecision of annotations, and misclassification we refer to [12, 44]. In addition to GO, the techniques outlined below are applicable to other annotation data (listed in Table 1).

5.1 Annotation term enrichment analysis

A very straightforward approach to systematically interpret sets of proteins consists of testing whether the set is enriched with certain functional annotations. This "classic" overrepresentation analysis has been reviewed by multiple authors [7, 10, 16], but for completeness we will briefly outline its principles and present a contemporary list of tools. It involves comparing the frequency of individual functional annotations within a set of protein identifiers against the annotation frequency in a reference list. The enrichment of a term can be statistically tested, e.g. using a hypergeometric distribution or a Fisher's exact test. Since the number of functional terms for which the enrichment is tested is usually large, a suitable multiple testing correction needs to be applied. For a discussion of the statistical aspects of enrichment analysis we refer to [16, 45]. An enormous advantage and cause of the popularity of annotation term enrichment compared to the singular analysis of individual protein list entries is that it allows to easily summarize the functional properties of a results table into a limited set of overrepresented functions. This offers transparent foundations for the construction of an interpretation discussion based on global findings instead of individual protein entries.

Many tools developed over the years for annotation term enrichment are summarized in Table 2 [46–57], and for a comprehensive review we refer to Huang et al. [16]. An illustrated GO enrichment based on an experimental dataset is presented in Fig. 2. Although their simplicity makes them very attractive, methods based on overrepresentation analysis have a tendency to suffer from limited discriminative power [58]. They are also sensitive to the chosen cutoff threshold that determines which proteins are retained in the list [59, 60]. Most of the current approaches assume that proteins are independent and do not capture synergistic effects [61]. Most enrichment methods disregard missing annotations. They also ignore the structure of the functional graph and the partial redundancies between GO terms can complicate the enrichment list, although specific methods have been created to reduce these dependencies or filter the list to the most informative terms [62, 63].

Another inherent problem lies in the selection of a reference set. If a representative list of proteins that are observable by the experimental platform cannot be generated, the full genome is often chosen as the reference. More suitable options however are available. The previously discussed

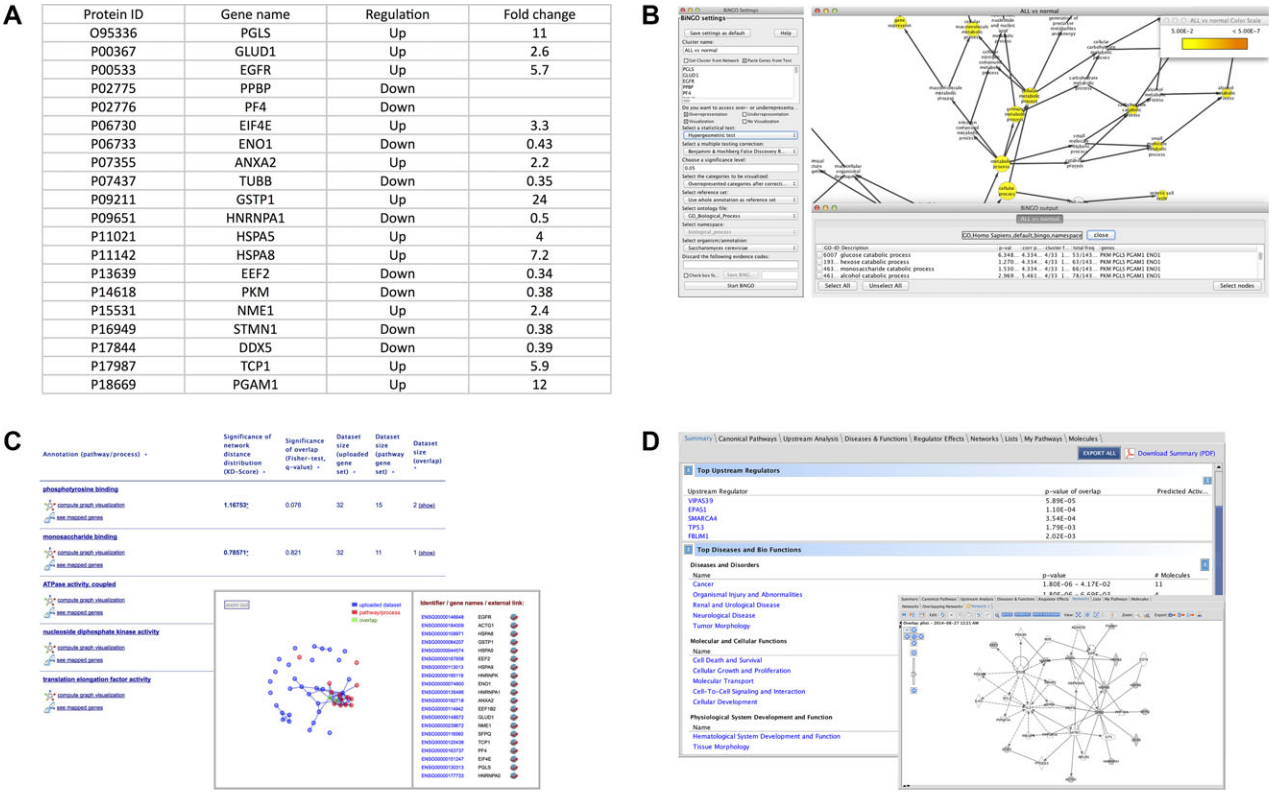
Table 2. Overview of commonly used computational tools for the functional interpretation of protein lists

Section	Name	Reference	Availability[a)]	Link
Keyword enrichment	agriGO	[55]	Public	http://bioinfo.cau.edu.cn/agriGO/
	BinGO	[49]	Public	http://apps.cytoscape.org/apps/bingo
	ClueGO	[54]	Public	http://apps.cytoscape.org/apps/cluego
	DAVID	[52]	Public	http://david.abcc.ncifcrf.gov
	FuncAssociate	[46]	Public	http://llama.mshri.on.ca/cgi/func/funcassociate
	GenMAPP	[51]	Public	http://www.genmapp.org
	GOEAST	[53]	Public	http://omicslab.genetics.ac.cn/GOEAST/
	GOminer	[48]	Public	http://discover.nci.nih.gov/gominer/index.jsp
	GOStat	[177]	Public	http://gostat.wehi.edu.au
	Ingenuity IPA		Commercial	http://www.ingenuity.com/products/ipa
	WebGestalt	[57]	Public	http://bioinfo.vanderbilt.edu/webgestalt/
	WEGO	[50]	Public	http://wego.genomics.org.cn/cgi-bin/wego/index.pl
Set enrichment	ASSESS	[74]	Public	http://people.genome.duke.edu/assess/
	EnrichR	[139]	Public	http://amp.pharm.mssm.edu/Enrichr
	FatiScan	[81]	Public	http://www.gepas.org
	GAGE	[82]	Public	http://www.bioconductor.org/packages/release/bioc/html/gage.html
	Gazer	[75]	Public	http://expressome.kobic.re.kr/GAzer/document.jsp
	GLOBALTEST	[178]	Public	http://www.bioconductor.org/packages/2.0/bioc/html/globaltest.html
	GSEA	[83]	Public	http://www.broadinstitute.org/gsea/index.jsp
	PAGE	[80]	On request	
	GSVA	[179]	Public	http://www.bioconductor.org/packages/release/bioc/html/GSVA.html
Text mining	Anni	[89]	Public	http://biosemantics.org/anni/
	CoCiter	[100]	Public	http://www.picb.ac.cn/hanlab/cociter/
	CoPub	[15]	Public	http://copub.gatcplatform.nl
	FAUN	[97]	Public	https://grits.eecs.utk.edu/faun/
	Martini	[93]	Public	http://martini.embl.de
	LAITOR	[98]	Public	http://laitor.sourceforge.net
	PESCADOR	[99]	Public	http://cbdm.mdc-berlin.de/tools/pescador/
	TXTGate	[96]	Public	http://tomcat.esat.kuleuven.be/txtgate/
Network enrichment	DEGraph	[128]	Public	http://www.bioconductor.org/packages/release/bioc/html/DEGraph.html
	EnrichNet	[58]	Public	http://www.enrichnet.org
	MetaCore	[180]	Commercial	https://portal.genego.com
	NetGSA	[127]	Public	http://www.biostat.washington.edu/˜ashojaie/research.html
	PathNet	[124]	Public	http://www.bioconductor.org/packages/release/bioc/html/PathNet.html
	Pathway-Express	[113]	Registration	http://vortex.cs.wayne.edu/projects.htm
	PWEA	[126]	Public	http://zlab.bu.edu/PWEA/
	SCORE-PAGE	[123]	On request	
	SPIA	[125]	Registration	http://vortex.cs.wayne.edu/ontoexpress/
	THINK-Back-DS	[181]	Public	http://www.eecs.umich.edu/db/think/software.html
	TopoGSA	[181]	Public	http://www.topogsa.org

a) Availability under the assumption of academic, noncommercial use.

biases of proteome analysis techniques toward protein abundance and physicochemical properties demand for adjusted reference sets or adapted statistical tests to alleviate the propagation of selection bias into the enrichment results [20, 64]. Even though the definition of relevant reference sets for proteome list enrichment analysis is nontrivial, we can anticipate that the increasing availability of exhaustive reference proteome databases for various tissues [21, 22] can further foster these developments. However the choice of smaller, adjusted background reference lists comes at a cost of decreased statistical power [16]. For example, in a new enrichment analysis of a published differential protein list [65] we observed a reduction from 127 enriched GO categories (using the full genome as a background reference) to 47 enriched categories when instead a representative reference list of 8553 proteins was used based on the draft proteome [66], according to identical statistical criteria (Supporting Information File 1). The draft proteome itself contains several overrepresented GO categories compared to the full genome, which may explain some observed overrepresented categories that are only observed when the differential data are compared to the full genome. The majority of the lost enrichments however are most likely a consequence of the decreased statistical power with the use of a smaller reference dataset. This analysis illustrates that the impact of proteome-specific reference lists certainly deserves attention. In this context it is important to select an enrichment tool that supports the use of custom background sets, which is not the case for all tools.

Annotation term enrichment is in principle applicable to the three types of protein lists. For reference proteomes enrichment analysis compared to the full (theoretical) proteome can reveal the impact of experimental biases on the observed proteins and the extent of coverage for biologically relevant compartments and functions. For targeted and differential proteome lists the role of enrichment results in the interpretation can be diverse, dependent on the biological context and experimental setup.



Protein ID	Gene name	Regulation	Fold change
O95336	PGLS	Up	11
P00367	GLUD1	Up	2.6
P00533	EGFR	Up	5.7
P02775	PPBP	Down	
P02776	PF4	Down	
P06730	EIF4E	Up	3.3
P06733	ENO1	Down	0.43
P07355	ANXA2	Up	2.2
P07437	TUBB	Down	0.35
P09211	GSTP1	Up	24
P09651	HNRNPA1	Down	0.5
P11021	HSPA5	Up	4
P11142	HSPA8	Up	7.2
P13639	EEF2	Down	0.34
P14618	PKM	Down	0.38
P15531	NME1	Up	2.4
P16949	STMN1	Down	0.38
P17844	DDX5	Down	0.39
P17987	TCP1	Up	5.9
P18669	PGAM1	Up	12

Annotation (pathway/process)	Significance of network distance distribution (XD-Score)	Significance of overlap (Fisher-test, q-value)	Dataset size (uploaded gene set)	Dataset size (pathway gene set)	Dataset size (overlap)
phosphotyrosine binding compute graph visualization see mapped genes	1.16753	0.076	32	15	2 (show)
monosaccharide binding compute graph visualization see mapped genes	0.78571	0.821	32	11	1 (show)
ATPase activity, coupled compute graph visualization see mapped genes					
nucleoside diphosphate kinase activity compute graph visualization see mapped genes					
translation elongation factor activity compute graph visualization see mapped genes					

Identifier / gene names / external link:

ENSG00000146648	EGFR
ENSG00000184009	ACTG1
ENSG00000109971	HSPA8
ENSG00000084207	GSTP1
ENSG00000044574	HSPA5
ENSG00000167658	EEF2
ENSG00000113013	HSPA9
ENSG00000165119	HNRNPK
ENSG00000074800	ENO1
ENSG00000135486	HNRNPA1
ENSG00000182718	ANXA2
ENSG00000114942	EEF1B2
ENSG00000148672	GLUD1
ENSG00000239672	NME1
ENSG00000116560	SFPQ
ENSG00000120438	TCP1
ENSG00000163737	PF4
ENSG00000151247	EIF4E
ENSG00000130313	PGLS
ENSG00000177733	HNRNPA0

Top Upstream Regulators

Upstream Regulator	p-value of overlap	Predicted Activ...
VIPAS39	5.89E-05	
EPAS1	1.10E-04	
SMARCA4	3.54E-04	
TP53	1.79E-03	
FBLIM1	2.02E-03	

Top Diseases and Bio Functions

Diseases and Disorders

Name	p-value	# Molecules
Cancer	1.80E-06 - 4.17E-02	11
Organismal Injury and Abnormalities		
Renal and Urological Disease		
Neurological Disease		
Tumor Morphology		

Molecular and Cellular Functions

Name
Cell Death and Survival
Cellular Growth and Proliferation
Molecular Transport
Cell-To-Cell Signaling and Interaction
Cellular Development

Physiological System Development and Function

Name
Hematological System Development and Function
Tissue Morphology

Figure 2. Functional interpretation of a differential protein list with different tools. Figure 2 shows several tools that are traditionally used for gene expression analysis, but have often been used in functional protein knowledge discovery. Here, a public dataset of the NCI-60 proteome (http://wzw.tum.de/proteomics/nci60) is used that shows differences between leukemia and breast cancer. The protein list is shown in the top left pane (A) and is analyzed for GO term enrichment with BinGO (B), network enrichment with EnrichNet (C) and pathway enrichment with IPA (D). In Fig. 2B, color indicates significance, with the more intense orange indicating a higher *p*-value. EnrichNet (Fig. 2C) shows all calculated values and allows the network visualization of the pathways and processes it identifies. Figure 2D shows the summary display in IPA. Pathway enrichment is calculated here with a Fisher exact test. Longer bars indicate higher significance.

5.2 Set enrichment analysis

Complementary to the classical overrepresentation analysis of unordered lists, it is possible to evaluate the significance of predefined sets of items in ordered lists, which are often generated in differential proteomics studies. This is known as set enrichment analysis. Set enrichment approaches detect whether or not the items of predefined subsets (e.g. based on a common functional annotation) are randomly distributed in the ordered list or have a tendency to cooccur at the top or the bottom of the list. Set enrichment analysis is capable of detecting less pronounced changes and does not require a hard cutoff threshold to be defined as in the term enrichment approach that is based on an unordered list. Originally developed for the analysis of microarray data, this methodology [67, 68] is used to interpret differentially expressed genes in gene set enrichment analysis [58, 69]. To employ this method on sorted lists of proteins by using the protein identifiers a modified approach has been introduced, referred to as protein set enrichment [70].

Conducting a gene set enrichment analysis first requires the definition of the gene sets. These can be created based on the GO [42] or other annotation systems such as InterPro [71] or KEGG [72], or specialized set databases [73–75]. There are multiple methods (see Table 1), but they all follow the same steps. They start with a ranked list of proteins, usually based on their quantitative differences (e.g. based on *t*-statistics). Next, they calculate enrichment scores. Permutation methods then allow the creation of a null distribution that is used to test the significance. Depending on the tool either the proteins or the class labels (samples) are randomized to generate this null distribution [76], an approach that has a substantial impact on the results, dependent on the experimental setup [45]. Comparing the enrichment scores of the observed distribution with the null distribution, allows to calculate associated *p*-values (corrected for multiple testing [77]).

In general, gene set enrichment analysis methods can be categorized based on how they define the null hypothesis [14]. A first group are the self-contained methods, which only consider the elements in the list to test, as null hypothesis, that no genes in the set can be associated with the observed phenotype [78]. Competitive methods are currently most common [45] and require that also genes outside the set are sampled to be compared with the set of interest. A third category are the parametric approaches [79–82], in which a z-score replaces the empirical null distribution.

Gene set analysis has become popular for the analysis of differential expression data as it offers far more statistical power than term enrichment analysis [61, 83]. It addresses three core issues of the classical overrepresentation analysis. There is no requirement of an arbitrary threshold to distinguish significantly differential transcripts or proteins. It accommodates for the dependencies between proteins within a pathway. Finally, set enrichment detects coordinated changes in gene product abundances to discover significantly affected functional classes, which is something overrepresentation approaches are incapable of. Nevertheless these methods also have weaknesses. It is not trivial to compare the performance of the many available techniques. Using real biological datasets for an evaluation is difficult because there is not a real golden standard that defines what is biologically true or false. Synthetic datasets can be used but hardly simulate the true biological complexity. Nevertheless a comparison of several components of gene set analysis can be found in [84]. An additional problem is that competitive approaches [67] incorrectly assume that all the proteins in the list are independent, and consequently inflate the false positive rate [14, 85], although promising solutions to remove this correlation bias have been proposed [86]. Like term enrichment analysis, gene set enrichment analysis also considers each functional set to be independent of others [87], an assumption that is not entirely biologically correct, since protein subsets can be part of several pathways. Finally, without precaution gene set enrichment analysis tends to suffer from inherent biases due to the experimental analytical technology, similar to those discussed above for the unordered list enrichment analysis [88]. Since set enrichment analysis is applied to ranked lists, its application in the scope of proteome analysis is restricted to differential protein list data, as illustrated in Fig. 1.

6 Enrichment analysis based on literature keywords

A wealth of knowledge to assist in the interpretation of protein lists remains hidden in literature. With the growing number of articles relevant to proteomics, it becomes increasingly hard for researchers to keep up with literature. Therefore there is a need for automated text mining tools that make literature content accessible [3, 89]. For example, text-mining methods that consider co-occurrence of terms can be used to rapidly generate hypotheses. MEDLINE abstracts can be easily mined for protein occurrences and can then be linked to MESH concepts or GO terms [89], which can then be used to investigate the relevance of proteins for diseases. Furthermore, literature contexts have been used to infer novel protein–protein interactions [90].

Text mining approaches can be roughly distinguished based on whether they use the words from the abstract [63, 91] or instead tag thesaurus entries [43, 89, 92]. The way in which the text abstracts are queried also differs across methods. For example, some tools use thesauruses to link text keywords to genes [43, 93], while others require Pubmed queries [94]. The existing tools can also be distinguished according to how they handle the associations between terms and genes. The simplest approach is to investigate direct [95] or indirect co-occurrence [96] and then statistically test overrepresentation [15] but other and more advanced methods have been described [97–100]. For an overview of tools, we refer to Table 2.

7 Network approaches

Molecular network data are rapidly becoming available and present a rich information source for the functional interpretation of protein lists. Most network data resources are far from complete, but they provide information that is complementary to other functional analysis techniques [7, 10, 87, 101–104]. Interactions can provide insights in the hidden relationships between individual proteins that cooccur in an experimental list. Even a simple topological characteristic, such as the number of interactions connecting a protein can suggest that it plays a central functional role. If proteins found in a differential proteomics list are members of the same network neighborhood, the chances that their observation is *bona fide* increases [104].

The simplest approach to use network data to interpret protein lists consist of mapping and subsequently highlighting the individual proteins on a pathway or network (also illustrated in Fig. 1). Most network and pathway visualization tools can perform this task [105–107]. It works particularly for small (sub) networks, as complex network data become difficult to visualize. Network data can also be used for overrepresentation analysis and set enrichment as outlined above for ontologies and literature keywords (as shown in Fig. 1). In this case the topology of the network is used to determine how similar individual proteins are. A metric for the similarity or distance between proteins is the shortest path needed to connect two proteins. Overrepresentation of proteins of the experimental list within a small network neighborhood can then be statistically tested, e.g. with a Fisher-exact test. In subgraph prediction differential proteins are mapped onto a reference network followed by clustering [40, 101, 108]. To improve the biological relevance and quality of the detected subgraphs additional quality metrics, such as GO term coherence of subgraph members [9, 63], may be incorporated.

The detection of overrepresented subnetworks can be of great value for the interpretation of proteomics data, and has led to successful applications, e.g. in biomarker discovery and prioritization [109, 110], in the elucidation of disease causes [111] and in the generation of fundamental mechanistic insights in various processes [28,112]. Below we will discuss the major types of relevant network data and outline how they can be of practical use.

7.1 Metabolic pathways

Metabolic pathway-based interpretation and hypothesis generation has grown in importance over the last few years [87, 109, 113, 114]. The use of metabolic pathways allows the generation of hypotheses with regard to the effect of a process on the metabolome. For example, if several proteins affected in a differential analysis are members of the same metabolic pathway, it is likely that this metabolic pathway and thus its resulting metabolites are affected in the study. This may be confirmed in further metabolomics follow-up studies. As metabolic pathways are driven by proteins (i.e. metabolic enzymes), the overlay of experimental proteome data has potentially more explanatory value about the dynamics of the metabolome than the overlay of data from the transcriptional level. A variety of tools, some of which commercial, have been developed for the analysis of enrichments of pathways in protein lists (See Table 2) [58, 87, 115, 116], sometimes in combination with GO-information [48] or protein interaction networks [117]. The knowledge bases at the core of these analyses are often extensively curated, e.g. Reactome [118], KEGG [72] and IPA (www.ingenuity.com) or rely on community feedback and submission, e.g. Wikipathways [119]. Several common pathway databases are presented in Table 1 [72, 120–122], and an illustrated example of pathway analysis executed with IPA is presented in Fig. 2.

Pathway information can be used in enrichment and subnetwork detection analysis in several ways (Table 2). ScorePAGE [123] weighs protein level statistics with a score based on the number of reactions to connect two proteins, and other tools use additional information. PathNet [124] uses the relation of a protein to a disease to reveal significantly overrepresented pathways or use impact factor analysis [125]. A first class of methods, including most of the earlier methods [113, 123, 125], creates an aggregated pathway score after node level analysis. A second type uses a topology-driven approach in the node level analysis itself, e.g. by using the correlation coefficient to score node pairs [126]. Others employ multivariate approaches [127, 128]. Many techniques have been reviewed by Mitrea and colleagues [129].

Pathway databases are not always entirely consistent, as various components are simultaneously involved in multiple processes, making a clear distinction of individual pathways difficult. The complexity of the data and the required intensive curation efforts result in data with lower coverage than resources such as GO [109]. Soh and colleagues [114] reported that even popular curated pathway databases show only limited overlap and that each source is partially incomplete, but this should improve over time.

7.2 Binary protein–protein interactions

Despite being extremely meaningful, pathway databases only cover a very specific functional subspace of the proteome. An entirely different level consists of the networks of pairwise interactions between proteins. Interactome data have grown exponentially, especially since the emergence of high-throughput interaction mapping methods [102, 130, 131]. The network of protein–protein interactions can be a powerful information source to reveal and explain hidden relationships within protein lists. For example proteins that are found to be coregulated in differential protein lists and that are neighbors in the interaction network suggest that together they play a role in the affected biological processes.

Important protein–protein interaction databases that capture this information are presented in Table 1 [132–136]. Their flexibility and information value have resulted in a range of tools [137–139] (see Table 2) that allow the user to incorporate this data in the interpretation task. In Fig. 2 an example of network enrichment is shown based on the EnrichNet tool.

Despite their value and availability protein interaction data come with pitfalls. Protein–protein interaction can be transient or permanent [130], leading to variable consistency and reliability of across experimental platforms [140] and databases that mix these classes. Interaction databases tend to be biased toward certain functions and components. Many databases contain also interactions based on indirect evidence, for example, based on genetic or metabolic evidence or on colocalization data [130] and thus require careful interpretation. Even protein–protein interaction information is largely incomplete and mostly limited to a small number of model species. To address the scattering of information over multiple databases [140], several recent (and nontrivial) efforts have been initiated to consolidate resources, e.g. Mintact [141] and String [142]. Complementary to the various interactome databases, the protein ontology consortium has developed a structured representation of relationships among proteins and their different variants and modified forms [143].

7.3 Regulatory networks

In addition to metabolic pathways and binary protein interactions, we can use regulatory network information to interpret experimental protein lists. In general, this information can be relevant to reveal regulatory associations between proteins in a list. A first type of such association is the effector-target relationship, in which a protein and its (direct or indirect) target(s) are observed in the same list. Another relevant association that can be found is the coregulated target association, which indicates that multiple proteins of a list are under influence of a common regulator. The attribution of multiple proteins of a differential protein list to the same regulatory pathway can potentially explain their coobservation as a consequence of the same regulatory mechanisms. Below we will discuss the most relevant types of regulatory network information.

7.3.1 PTM networks

As PTMs are both unique to the protein level and constitute a key level of regulation [144], the network of proteins that regulate each other through PTMs is becoming a valuable component in the pipeline toward the interpretation of experimental protein lists. For example, this level could explain the correlated observation of a modifying enzyme and its substrate, or of multiple substrates that are under influence of the same modification enzyme. These types of connections are relevant for the interpretation of both differential protein lists and many targeted proteomes, information that would probably never be revealed through any of the other tools described in this paper. Crosstalk between PTM regulators further complicates the biological reality and adds substantial challenges [145,146] and attempts have been made to describe the global nature of the network that connects functionally associated PTMs. Nevertheless our current knowledge on PTMs and their role is still fragmented and incomplete [147]. New techniques to systematically identify modifications [148] have resulted in several modification databases [149–153]. Promising efforts are being made to develop centralized resources for known and predicted PTM associations in projects such as PTMcode [154,155]. The systematic inclusion of PTM network data in data analysis pipelines is not yet obvious, but over the years these resources may become valuable complementary tools for the in-depth interpretation of protein lists and will likely follow equivalent paradigms as the incorporation of other network data in functional analysis workflows.

7.3.2 Transcriptional regulation networks

Regulatory networks are also controlled at the transcriptomics level, either by miRNAs binding to complementary sequences or sequence-specific transcription factors (TFs).

Transcription factors are key effectors that regulate gene expression in eukaryotes and have thus been extensively studied. They recognize an extensive range of genomic binding sites [156] and regulate large amounts of downstream genes [157], which complicates the analysis of their eventual effects. Additionally, several transcription factors regulate others, resulting in a complex web of interactions. Usually, TFs are studied in a network context with the goal of discovering motifs, such as related TFs binding functionally coherent genes or the discovery of coexpressed TFs in an attempt to gain a better understanding of the underlying processes. Often these networks are constructed based on sequence motif templates obtained from databases such as TRANSFAC [158] or databases containing experimental evidence obtained from yeast-one hybrid and ChIP [159–161]. Regardless of origin, these networks can be connected to existing protein–protein interactions, thus linking the proteomics and genomics contexts with each other. However, one has to be cautious when dealing with the edges in the integrated network. In the default transcription factor network, nodes are regulatory DNA motifs or transcription factors, and the interactions between these elements are directed [162], while binary protein–protein interactions are undirected. As such, the result of mapping the TF network on protein–protein interaction data is a network consisting of both directed and undirected interactions, which poses problems for several network analysis tools.

In contrast to transcription factors, much is still unknown about the mode of action of the miRNAs and the scale of their involvement in regulation, as highthroughput methods for their detection and mapping only recently emerged [163]. However, it is suggested that they provide a layer for fine-tuning regulation by impacting the levels of complementary mRNA, [164]. A detailed review of their effects is available in Inui et al. [165]. Although very little experimental evidence is available, various computational tools already exist to predict miRNA targets [166–168]. In addition to miRNAs, various other regulatory evidence types can be used to strengthen or weaken relations between proteins, such as gene coexpression and phenotypic profiling.

7.3.3 Other types of signal transduction

Although PTMs and transcriptional regulation networks are an essential part of cellular signaling cascades, other signal transduction mechanisms cannot be ignored. Most signaling cascades are multistep processes that require several inputs (receptor binding their ligands) that are combined by means of various scaffold proteins that direct the information flow [169]. The presence of scaffold proteins and their ligands in an experimental list can thus point to specific intracellular signal transduction cascades [170]. Similarly, secondary messengers and G-coupled proteins can be used to reconstruct signaling pathways by looking at their interaction partners and the other proteins present in the list.

8 Conclusions and perspectives

Despite the availability of numerous tools and resources, experimental protein lists are often only partially interpreted. Computational approaches used to analyze then are diverse, but not enough is known about their comparative performances on proteome data. It is therefore imperative to come up with more uniform approaches for the optimal functional interpretation of proteome data and the evaluation thereof.

While the techniques and resources addressed in this review are valuable tools that can shed important insights in often complex data coming from proteomics experiments, the field is currently rather scattered. Choices for a given method are often driven by practical considerations, such as familiarity with certain tools, or the availability of commercial licenses. The availability of multiple tools, multiple parameters, and the extensive lists of results that many tools generate inherently carry the risk of cherry picking, i.e. reporting the results that optimally fit the assumptions. Despite the availability of numerous tools, important opportunities for further research are the problem of selection bias in proteomics and its impact on functional analysis. New emerging network data and proteome-specific features such as the presence of PTMs, the dynamics of protein complexes and the emergence of protein–drug interaction resources [171] are likely interesting avenues for further developments, especially with increasing availability of relevant high quality data. Better approaches to evaluate tools and functional interpretation results are needed to enable this field to become a mature subfield of proteome bioinformatics. In addition, there is an immense need for a robust framework to handle nonmodel organism proteomes, through a combination of limited existing knowledge about the species and orthology mapping to integrate knowledge from better characterized organisms [172].

Systematic functional interpretation deserves more recognition as an inherent part of any proteome analysis workflow. With this paper we aim to contribute to the awareness of the importance of defining an optimal functional interpretation strategy and properly reporting it alongside the publication of proteomics results.

This work was supported by the Research Foundation-Flanders (FWO) [grant number: G.0903.13N]; the agency for Innovation by Science and Technology (IWT) [grant number 120025]; and the University of Antwerp [BOF docpro to S.N.].

The authors have declared no conflict of interest.

9 References

[1] Curtis, R. K., Orešič, M., Vidal-Puig, A., Pathways to the analysis of microarray data. *Trends Biotechnol.* 2005, *23*, 429–435.

[2] Werner, T., Bioinformatics applications for pathway analysis of microarray data. *Curr. Opin. Biotechnol.* 2008, *19*, 50–54.

[3] Faro, A., Giordano, D., Spampinato, C., Combining literature text mining with microarray data: advances for system biology modeling. *Brief. Bioinform.* 2011, bbr018.

[4] Boguski, M. S., McIntosh, M. W., Biomedical informatics for proteomics. *Nature* 2003, *422*, 233–237.

[5] Palagi, P. M., Hernandez, P., Walther, D., Appel, R. D., Proteome informatics I: bioinformatics tools for processing experimental data. *Proteomics* 2006, *6*, 5435–44.

[6] Kumar, C., Mann, M., Bioinformatics analysis of mass spectrometry-based proteomics data sets. *Febs Lett.* 2009, *583*, 1703–1712.

[7] Malik, R., Dulla, K., Nigg, E. A., Körner, R., From proteome lists to biological impact- tools and strategies for the analysis of large MS data sets. *Proteomics* 2010, *10*, 1270–1283.

[8] Matthiesen, R., Azevedo, L., Amorim, A., Carvalho, A.S., Discussion on common data analysis strategies used in MS-based proteomics. *Proteomics* 2011, *11*, 604–619.

[9] Goh, W. W. B., Wong, L., Computational proteomics: designing a comprehensive analytical strategy. *Drug Discov. Today* 2014, *19*, 266–274.

[10] Lisacek, F., Cohen-Boulakia, S., Appel, R. D., Proteome informatics II: Bioinformatics for comparative proteomics. *Proteomics* 2006, *6*, 5445–5466.

[11] Drăghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C., Krawetz, S. A., Global functional profiling of gene expression. *Genomics* 2003, *81*, 98–104.

[12] Khatri, P., Drăghici, S., Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 2005, *21*, 3587–3595.

[13] Gamberoni, G., Lamma, E., Lodo, G., Marchesini, J. et al., Fun&Co: identification of key functional differences in transcriptomes. *Bioinformatics* 2007, *23*, 2725–2732.

[14] Goeman, J. J., Bühlmann, P., Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007, *23*, 980–987.

[15] Frijters, R., Heupers, B., Beek, P. van, Bouwhuis, M. et al., CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res.* 2008, *36*, W406–W410.

[16] Huang, D. W., Sherman, B. T., Lempicki, R. A., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009, *37*, 1–13.

[17] Geistlinger, L., Csaba, G., Küffner, R., Mulder, N., Zimmer, R., From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics* 2011, *27*, i366–i373.

[18] Schwanhäusser, B., Busse, D., Li, N., Dittmar, G. et al., Global quantification of mammalian gene expression control. *Nature* 2011, *473*, 337–342.

[19] Zubarev, R. A., The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* 2013, *13*, 723–726.

[20] Louie, B., Higdon, R., Kolker, E., The necessity of adjusting tests of protein category enrichment in discovery proteomics. *Bioinformatics* 2010, *26*, 3007–3011.

[21] Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S. et al., A draft map of the human proteome. *Nature* 2014, *509*, 575–581.

[22] Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M. et al., Mass-spectrometry-based draft of the human proteome. *Nature* 2014, *509*, 582–587.

[23] Kelchtermans, P., Bittremieux, W., De Grave, K., Degroeve, S. et al., Machine learning applications in proteomics research: How the past can boost the future. *Proteomics* 2014, *14*, 353–366.

[24] Mathesius, U., Imin, N., Chen, H., Djordjevic, M. A. et al., Evaluation of proteome reference maps for cross-species identification of proteins by peptide mass fingerprinting. *Proteomics* 2002, *2*, 1288–1303.

[25] Laukens, K., Deckers, P., Esmans, E., Van Onckelen, H., Witters, E., Construction of a two-dimensional gel electrophoresis protein database for the Nicotiana tabacum cv. Bright Yellow-2 cell suspension culture. *Proteomics* 2004, *4*, 720–727.

[26] Cottrell, J. S., Protein identification using MS/MS data. *J. Proteomics* 2011, *74*, 1842–1851.

[27] Lievens, S., Eyckerman, S., Lemmens, I., Tavernier, J., Large-scale protein interactome mapping: strategies and opportunities. *Expert Rev. Proteomics* 2010, *7*, 679–690.

[28] Van Leene, J., Hollunder, J., Eeckhout, D., Persiau, G. et al., Targeted interactomics reveals a complex core cell cycle machinery in Arabidopsis thaliana. *Mol. Syst. Biol.* 2010, *6*, 397.

[29] Corominas, R., Yang, X., Lin, G. N., Kang, S. et al., Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nat. Commun.* 2014, *5*, 3650.

[30] Fadden, P., Huang, K. H., Veal, J. M., Steed, P. M. et al., Application of chemoproteomics to drug discovery: identification of a clinical candidate targeting Hsp90. *Chem. Biol.* 2010, *17*, 686–694.

[31] Roux, P. P., Thibault, P., The coming of age of phosphoproteomics—from large data sets to inference of protein functions. *Mol. Cell. Proteomics* 2013, *12*, 3453–3464.

[32] Van den Berg, B. H. J., Tholey, A., Mass spectrometry-based proteomics strategies for protease cleavage site identification. *Proteomics* 2012, *12*, 516–529.

[33] Bantscheff, M., Lemeer, S., Savitski, M. M., Kuster, B., Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* 2012, *404*, 939–965.

[34] Neilson, K. A., Ali, N. A., Muralidharan, S., Mirzaei, M. et al., Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics* 2011, *11*, 535–553.

[35] Oberg, A. L., Mahoney, D. W., Statistical methods for quantitative mass spectrometry proteomic experiments with labeling. *BMC Bioinformatics* 2012, *13*, S7.

[36] Liu, W., Li, C., Xu, Y., Yang, H. et al., Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics* 2013, *29*, 2169–2177.

[37] Matzke, M. M., Brown, J. N., Gritsenko, M. A., Metz, T. O. et al., A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments. *Proteomics* 2013, *13*, 493–503.

[38] Carpentier, S. C., Panis, B., Swennen, R., Lammertyn, J., Finding the significant markers: statistical analysis of proteomic data. *Methods Mol. Biol. Clifton Nj* 2008, *428*, 327–347.

[39] Jung, K., Statistics in experimental design, preprocessing, and analysis of proteomics data. *Methods Mol. Biol.* 2011, *696*, 259–272.

[40] Naulaerts, S., Meysman, P., Bittremieux, W., Vu, T. N. et al., A primer to frequent itemset mining for bioinformatics. *Brief. Bioinform.* 2013. doi: 10.1093/bib/bbt074.

[41] The UniProt consortium, reorganizing the protein space at the universal protein resource (UniProt). *Nucleic Acids Res.* 2011, *40*, D71–D75.

[42] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D. et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 2000, *25*, 25–29.

[43] Jelier, R., Goeman, J. J., Hettne, K. M., Schuemie, M. J. et al., Literature-aided interpretation of gene expression data with the weighted global test. *Brief. Bioinform.* 2011, *12*, 518–529.

[44] Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C., Friedberg, I., Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *Plos Comput. Biol.* 2013, *9*, e1003063.

[45] Maciejewski, H., Gene set analysis methods: statistical models and methodological differences. *Brief. Bioinform.* 2014, *15*, 504–518.

[46] Berriz, G. F., King, O. D., Bryant, B., Sander, C., Roth, F. P., Characterizing gene sets with FuncAssociate. *Bioinformatics* 2003, *19*, 2502–2504.

[47] Khatri, P., Draghici, S., Ostermeier, G. C., Krawetz, S. A., Profiling gene expression using onto-express. *Genomics* 2002, *79*, 266–270.

[48] Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D. et al., GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* 2003, *4*, R28.

[49] Maere, S., Heymans, K., Kuiper, M., BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* 2005, *21*, 3448–3449.

[50] Ye, J., Fang, L., Zheng, H., Zhang, Y. et al., WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 2006, *34*, W293–297.

[51] Salomonis, N., Hanspers, K., Zambon, A. C., Vranizan, K. et al., GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics* 2007, *8*, 217.

[52] Huang, D. W., Sherman, B. T., Lempicki, R. A., Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 2008, *4*, 44–57.

[53] Zheng, Q., Wang, X. J., GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.* 2008, *36*, W358–W363.

[54] Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P. et al., ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 2009, *25*, 1091–1093.

[55] Du, Z., Zhou, X., Ling, Y., Zhang, Z., Su, Z., agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* 2010, *38*, W64–W70.

[56] Zhang, S., Cao, J., Kong, Y. M., Scheuermann, R. H., GOBayes: Gene Ontology-based overrepresentation analysis using a Bayesian approach. *Bioinformatics* 2010, *26*, 905–911.

[57] Wang, J., Duncan, D., Shi, Z., Zhang, B., WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* 2013, *41*, W77–W83.

[58] Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., Valencia, A., EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* 2012, *28*, i451–i457.

[59] Pan, W., A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 2002, *18*, 546–554.

[60] Dalman, M. R., Deeter, A., Nimishakavi, G., Duan, Z. H., Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics* 2012, *13*, S11.

[61] Nam, D., Kim, S. Y., Gene-set approach for expression pattern analysis. *Brief. Bioinform.* 2008, *9*, 189–197.

[62] Grossmann, S., Bauer, S., Robinson, P. N., Vingron, M., Improved detection of overrepresentation of Gene-Ontology annotations with parent–child analysis. *Bioinformatics* 2007, *23*, 3024–3031.

[63] Huang, W. L., Tung, C. W., Ho, S. W., Hwang, S. F., Ho, S. Y., ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics* 2008, *9*, 80.

[64] Young, M. D., Wakefield, M. J., Smyth, G. K., Oshlack, A., Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 2010, *11*, R14.

[65] Chen, R., Yi, E. C., Donohoe, S., Pan, S. et al., Pancreatic cancer proteome: the proteins that underlie invasion, metastasis, and immunologic escape. *Gastroenterology* 2005, *129*, 1187–1197.

[66] Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S. et al., A draft map of the human proteome. *Nature* 2014, *509*, 575–581.

[67] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S. et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 2005, *102*, 15545–15550.

[68] Keller, A., Backes, C., Lenhof, H. P., Computation of significance scores of unweighted Gene Set Enrichment Analyses. *BMC Bioinformatics* 2007, *8*, 290.

[69] Hill, W. D., Davies, G., van de Lagemaat, L. N., Christoforou, A. et al., Human cognitive ability is influenced by genetic variation in components of postsynaptic signalling complexes assembled by NMDA receptors and MAGUK proteins. *Transl. Psychiatry* 2014, *4*, e341.

[70] Cha, S., Imielinski, M. B., Rejtar, T., Richardson, E. A. et al., In Situ proteomic analysis of human breast cancer epithelial cells using laser capture microdissection (LCM)-LC/MS: Annotation by protein set enrichment analysis (PSEA) and gene ontology (GO). *Mol. Cell. Proteomics* 2010, mcp.M110.000398.

[71] Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A. et al., InterPro: the integrative protein signature database. *Nucleic Acids Res.* 2009, *37*, D211–D215.

[72] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M. et al., Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 2014, *42*, D199–D205.

[73] Tomfohr, J., Lu, J., Kepler, T. B., Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 2005, *6*, 225.

[74] Edelman, E., Porrello, A., Guinney, J., Balakumaran, B. et al., Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics* 2006, *22*, e108–e116.

[75] Kim, S. B., Yang, S., Kim, S. K., Kim, S. C. et al., GAzer: gene set analyzer. *Bioinformatics* 2007, *23*, 1697–1699.

[76] Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J. et al., Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. U. S. A.* 2005, *102*, 13544–13549.

[77] Benjamini, Y., Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 1995, *57*, 289–300.

[78] Tarca, A. L., Bhatti, G., Romero, R., A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *Plos One* 2013, *8*, e79217.

[79] Tu, K., Yu, H., Zhu, M., MEGO: gene functional module expression based on gene ontology. *Biotechniques* 2005, *38*, 277–283.

[80] Kim, S. Y., Volsky, D. J., PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 2005, *6*, 144.

[81] Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J. et al., From genes to functional classes in the study of biological systems. *BMC Bioinformatics* 2007, *8*, 114.

[82] Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., Woolf, P. J., GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* 2009, *10*, 161.

[83] Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A. et al., PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 2003, *34*, 267–273.

[84] Hung, J. H., Yang, T. H., Hu, Z., Weng, Z., DeLisi, C., Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief. Bioinform.* 2012, *13*, 281–291.

[85] Dinu, I., Potter, J. D., Mueller, T., Liu, Q. et al., Gene-set analysis and reduction. *Brief. Bioinform.* 2009, *10*, 24–34.

[86] Nam, D., De-correlating expression in gene-set analysis. *Bioinformatics* 2010, *26*, i511–i516.

[87] Khatri, P., Sirota, M., Butte, A. J., Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 2012, *8*, e1002375.

[88] Geeleher, P., Hartnett, L., Egan, L. J., Golden, A. et al., Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics* 2013, *29*, 1851–1857.

[89] Jelier, R., Schuemie, M. J., Veldhoven, A., Dorssers, L. C. et al., Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol.* 2008, *9*, R96.

[90] Van Haagen, H. H. H. B. M., 't Hoen, P. A. C., Botelho Bovo, A., de Morrée, A. et al., Novel protein-protein interactions inferred from literature context. *PLoS One* 2009, *4*, e7894.

[91] Chaussabel, D., Sher, A., Mining microarray expression data by literature profiling. *Genome Biol.* 2002, *3*, research0055.1–research0055.16.

[92] Alako, B. T., Veldhoven, A., Baal, S. van, Jelier, R. et al., CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics* 2005, *6*, 51.

[93] Soldatos, T. G., O'Donoghue, S. I., Satagopam, V. P., Jensen, L. J. et al., Martini: using literature keywords to compare gene sets. *Nucleic Acids Res.* 2010, *38*, 26–38.

[94] Wang, J. H., Zhao, L. F., Lin, P., Su, X. R. et al., GenCLiP 2.0: a web server for functional clustering of genes and construction of molecular networks based on free terms. *Bioinformatics* 2014, *30*, 2534–2536.

[95] Jenssen, T. K., Laegreid, A., Komorowski, J., Hovig, E., A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 2001, *28*, 21–28.

[96] Glenisson, P., Coessens, B., Vooren, S. V., Mathys, J. et al., TXTGate: profiling gene groups with text-based information. *Genome Biol.* 2004, *5*, R43.

[97] Tjioe, E., Berry, M. W., Homayouni, R., Discovering gene functional relationships using FAUN (Feature Annotation Using Nonnegative matrix factorization). *BMC Bioinformatics* 2010, *11*, S14.

[98] Barbosa-Silva, A., Soldatos, T. G., Magalhães, I. L., Pavlopoulos, G. A. et al., LAITOR—literature assistant for identification of terms co-occurrences and relationships. *BMC Bioinformatics* 2010, *11*, 70.

[99] Barbosa-Silva, A., Fontaine, J. F., Donnard, E. R., Stussi, F. et al., PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. *BMC Bioinformatics* 2011, *12*, 435.

[100] Qiao, N., Huang, Y., Naveed, H., Green, C. D., Han, J. D. J., CoCiter: An efficient tool to infer gene function by assessing the significance of literature co-citation. *PLoS One* 2013, *8*, e74074.

[101] Goh, W. W. B., Lee, Y. H., Zubaidah, R. M., Jin, J. et al., Network-based pipeline for analyzing ms data: an application toward liver cancer. *J. Proteome Res.* 2011, *10*, 2261–2272.

[102] Goh, W. W. B., Lee, Y. H., Chung, M., Wong, L., How advancement in biological network analysis methods empowers proteomics. *Proteomics* 2012, *12*, 550–563.

[103] Goh, W. W. B., Fan, M., Low, H. S., Sergot, M., Wong, L., Enhancing the utility of proteomics signature profiling (PSP) with pathway derived Subnets (PDSs), performance analysis and specialised ontologies. *BMC Genomics* 2013, *14*, 35.

[104] Goh, W. W. B., Wong, L., Networks in proteomics analysis of cancer. *Curr. Opin. Biotechnol.* 2013, *24*, 1122–1128.

[105] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S. et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003, *13*, 2498–2504.

[106] Pavlopoulos, G. A., Wegener, A. L., Schneider, R., A survey of visualization tools for biological network analysis. *Biodata Min.* 2008, *1*, 12.

[107] Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A. et al., Visualization of omics data for systems biology. *Nat. Methods* 2010, *7*, S56–S68.

[108] Li, J., Zimmerman, L. J., Park, B. H., Tabb, D. L. et al., Network-assisted protein identification and data interpretation in shotgun proteomics. *Mol. Syst. Biol.* 2009, *5*.

[109] Wu, C. C., Hsu, C. W., Chen, C. D., Yu, C. J. et al., Candidate Serological biomarkers for cancer identified from the secretomes of 23 cancer cell lines and the human protein atlas. *Mol. Cell. Proteomics* 2010, *9*, 1100–1117.

[110] Imielinski, M., Cha, S., Rejtar, T., Richardson, E. A. et al., Integrated proteomic, transcriptomic, and biological network analysis of breast carcinoma reveals molecular features of tumorigenesis and clinical relapse. *Mol. Cell. Proteomics* 2012, *11*, M111.014910.

[111] Higareda-Almaraz, J. C., Enríquez-Gasca, M. R., Hernández-Ortiz, M., Resendis-Antonio, O., Encarnación-Guevara, S., Proteomic patterns of cervical cancer cell lines, a network perspective. *BMC Syst. Biol.* 2011, *5*, 96.

[112] Van Damme, E., Laukens, K., Dang, T. H., Van Ostade, X., A manually curated network of the PML nuclear body interactome reveals an important role for PML-NBs in SUMOylation dynamics. *Int. J. Biol. Sci.* 2010, *6*, 51–67.

[113] Draghici, S., Khatri, P., Tarca, A. L., Amin, K. et al., A systems biology approach for pathway level analysis. *Genome Res.* 2007, *17*, 1537–1545.

[114] Soh, D., Dong, D., Guo, Y., Wong, L., Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinformatics* 2010, *11*, 449.

[115] Cavalieri, D., Castagnini, C., Toti, S., Maciag, K. et al., Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases. *Bioinforma. Oxf. Engl.* 2007, *23*, 2631–2632.

[116] Foroushani, A. B. K., Brinkman, F. S. L., Lynn, D. J., Pathway-GPS and SIGORA: identifying relevant pathways based on the over-representation of their gene-pair signatures. *PeerJ* 2013, *1*, e229.

[117] Doderer, M. S., Anguiano, Z., Suresh, U., Dashnamoorthy, R. et al., Pathway Distiller—multisource biological pathway consolidation. *BMC Genomics* 2012, *13*, 1–16.

[118] Matthews, L., Gopinath, G., Gillespie, M., Caudy, M. et al., Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* 2009, *37*, D619–622.

[119] Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K. et al., WikiPathways: pathway editing for the People. *PLoS Biol* 2008, *6*, e184.

[120] Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L. et al., Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* 2005, *33*, 6083–6089.

[121] Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I. et al., Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* 2011, *39*, D685–D690.

[122] Mi, H., Muruganujan, A., Thomas, P. D., PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* 2013, *41*, D377–D386.

[123] Rahnenführer, J., Domingues, F. S., Maydt, J., Lengauer, T., Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat. Appl. Genet. Mol. Biol.* 2004, *3*, 1–29.

[124] Dutta, B., Wallqvist, A., Reifman, J., PathNet: a tool for pathway analysis using topological information. *Source Code Biol. Med.* 2012, *7*, 10.

[125] Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S. et al., A novel signaling pathway impact analysis. *Bioinformatics* 2009, *25*, 75–82.

[126] Hung, J. H., Whitfield, T. W., Yang, T. H., Hu, Z. et al., Identification of functional modules that correlate with phenotypic difference: the influence of network topology. *Genome Biol.* 2010, *11*, R23.

[127] Shojaie, A., Michailidis, G., Analysis of Gene Sets Based on the Underlying Regulatory Network. *J. Comput. Biol.* 2009, *16*, 407–426.

[128] Jacob, L., Neuvial, P., Dudoit, S., Gains in Power from Structured Two-Sample Tests of Means on Graphs. *Ann. Appl. Stat.* 2012, *6*, 561–600.

[129] Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S. et al., Methods and approaches in the topology-based analysis of biological pathways. *Front. Physiol.* 2013, *4*, 278.

[130] De Las Rivas, J., Fontanillo, C., Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *Plos Comput Biol* 2010, *6*, e1000807.

[131] Tang, H., Zhong, F., Xie, H., A quick guide to biomolecular network studies: Construction, analysis, applications, and resources. *Biochem. Biophys. Res. Commun.* 2012, *424*, 7–11.

[132] Xenarios, I., Salwínski, L., Duan, X. J., Higney, P. et al., DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 2002, *30*, 303–305.

[133] Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S. et al., Human protein reference database—2009 update. *Nucleic Acids Res.* 2009, *37*, D767–D772.

[134] Stark, C., Breitkreutz, B. J., Chatr-aryamontri, A., Boucher, L. et al., The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.* 2011, *39*, D698–D704.

[135] Kerrien, S., Aranda, B., Breuza, L., Bridge, A. et al., The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 2012, *40*, D841–D846.

[136] Licata, L., Briganti, L., Peluso, D., Perfetto, L. et al., MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 2012, *40*, D857–D861.

[137] Lachmann, A., Ma'ayan, A., Lists2Networks: Integrated analysis of gene/protein lists. *BMC Bioinformatics* 2010, *11*, 87.

[138] Schaefer, M. H., Fontaine, J. F., Vinayagam, A., Porras, P. et al., HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores. *PLoS One* 2012, *7*, e31826.

[139] Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q. et al., Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013, *14*, 128.

[140] Lehne, B., Schlitt, T., Protein–protein interaction databases: keeping up with growing interactomes. *Hum. Genomics* 2009, *3*, 291–297.

[141] Orchard, S., Ammari, M., Aranda, B., Breuza, L. et al., The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 2014, *42*, D358–D363.

[142] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M. et al., STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 2013, *41*, D808–D815.

[143] Bult, C. J., Drabkin, H. J., Evsikov, A., Natale, D. et al., The representation of protein complexes in the Protein Ontology (PRO). *BMC Bioinformatics* 2011, *12*, 371.

[144] Khoury, G. A., Baliban, R. C., Floudas, C. A., Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Reports* 2011, *1*.

[145] Beltrao, P., Bork, P., Krogan, N. J., van Noort, V., Evolution and functional cross-talk of protein post-translational modifications. *Mol. Syst. Biol.* 2013, *9*, 714.

[146] Minguez, P., Parca, L., Diella, F., Mende, D. R. et al., Deciphering a global network of functionally associated post-translational modifications. *Mol. Syst. Biol.* 2012, *8*, 599.

[147] Deribe, Y. L., Pawson, T., Dikic, I., Post-translational modifications in signal integration. *Nat. Struct. Mol. Biol.* 2010, *17*, 666–672.

[148] Choudhary, C., Mann, M., Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* 2010, *11*, 427–439.

[149] Lee, T. Y., dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.* 2006, *34*, D622–D627.

[150] Gnad, F., Gunawardena, J., Mann, M., PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res.* 2010, *39*, D253–D260.

[151] Dinkel, H., Chica, C., Via, A., Gould, C. M. et al., Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res.* 2010, *39*, D261–D267.

[152] Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B. et al., PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* 2012, *40*, D261–D270.

[153] Horn, H., Schoof, E. M., Kim, J., Robin, X. et al., KinomeXplorer: an integrated platform for kinome biology studies. *Nat. Methods* 2014, *11*, 603–604.

[154] Benayoun, B. A., Veitia, R. A., A post-translational modification code for transcription factors: sorting through a sea of signals. *Trends Cell Biol.* 2009, *19*, 189–197.

[155] Minguez, P., Letunic, I., Parca, L., Bork, P., PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins. *Nucleic Acids Res.* 2013, *41*, D306–D311.

[156] MacQuarrie, K. L., Fong, A. P., Morse, R. H., Tapscott, S. J., Genome-wide transcription factor binding: beyond direct target regulation. *Trends Genet.* 2011, *27*, 141–148.

[157] Johnson, D. S., Mortazavi, A., Myers, R. M., Wold, B., Genome-wide mapping of in vivo protein-dna interactions. *Science* 2007, *316*, 1497–1502.

[158] Wingender, E., The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* 2008, *9*, 326–332.

[159] Deplancke, B., Dupuy, D., Vidal, M., Walhout, A. J. M., A Gateway-Compatible Yeast One-Hybrid System. *Genome Res.* 2004, *14*, 2093–2101.

[160] Walhout, A. J. M., Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Res.* 2006, *16*, 1445–1454.

[161] Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., Luscombe, N. M., A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 2009, *10*, 252–263.

[162] Zhu, C., Byers, K. J. R. P., McCord, R. P., Shi, Z. et al., High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.* 2009, *19*, 556–566.

[163] Hafner, M., Landthaler, M., Burger, L., Khorshid, M. et al., Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 2010, *141*, 129–141.

[164] Guo, A. Y., Sun, J., Jia, P., Zhao, Z., A Novel microRNA and transcription factor mediated regulatory network in schizophrenia. *BMC Syst. Biol.* 2010, *4*, 10.

[165] Inui, M., Martello, G., Piccolo, S., MicroRNA control of signal transduction. *Nat. Rev. Mol. Cell Biol.* 2010, *11*, 252–263.

[166] John, B., Enright, A. J., Aravin, A., Tuschl, T. et al., Human MicroRNA Targets. *PLoS Biol* 2004, *2*, e363.

[167] Lewis, B. P., Burge, C. B., Bartel, D. P., Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell* 2005, *120*, 15–20.

[168] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., Segal, E., The role of site accessibility in microRNA target recognition. *Nat. Genet.* 2007, *39*, 1278–1284.

[169] Garrenton, L. S., Young, S. L., Thorner, J., Function of the MAPK scaffold protein, Ste5, requires a cryptic PH domain. *Genes Dev.* 2006, *20*, 1946–1958.

[170] Good, M. C., Zalatan, J. G., Lim, W. A., Scaffold Proteins: Hubs for Controlling the Flow of Cellular Information. *Science* 2011, *332*, 680–686.

[171] Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S., Blicher, T. H. et al., STITCH 4: integration of protein–chemical interactions with user data. *Nucleic Acids Res.* 2014, *42*, D401–D407.

[172] Remmerie, N., De Vijlder, T., Laukens, K., Dang, T. H. et al., Next generation functional proteomics in non-model plants: A survey on techniques and applications for the analysis of protein complexes and post-translational modifications. *Phytochemistry* 2011, *72*, 1192–1218.

[173] Nishimura, D., BioCarta. Biotech Softw. *Internet Rep.* 2001, *2*, 117–120.

[174] Bader, G. D., Betel, D., Hogue, C. W. V., BIND: the Biomolecular Interaction Network Database. *Nucl. Acids Res.* 2003, *31*, 248–250.

[175] Yamamoto, S., Sakai, N., Nakamura, H., Fukagawa, H. et al., INOH: ontology-based highly structured database of signal transduction pathways. *Database* 2011, *2011*, bar052–bar052.

[176] Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F. et al., RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* 2014, *42*, D756–D763.

[177] Beißbarth, T., Speed, T. P., GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 2004, *20*, 1464–1465.

[178] Goeman, J. J., Geer, S. A. van de, Kort, F. de, Houwelingen, H. C. van, A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004, *20*, 93–99.

[179] Hänzelmann, S., Castelo, R., Guinney, J., GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 2013, *14*, 7.

[180] Bessarabova, M., Ishkin, A., JeBailey, L., Nikolskaya, T., Nikolsky, Y., Knowledge-based analysis of proteomics data. *BMC Bioinformatics* 2012, *13*, S13.

[181] Farfán, F., Ma, J., Sartor, M. A., Michailidis, G., Jagadish, H. V., Think back: knowledge-based interpretation of high throughput data. *BMC Bioinformatics* 2012, *13*, 1–7.