The continuously evolving field of proteomics produces increasing amounts of data while improving the quality of protein identifications. Albeit quantitative measurements are becoming more popular, many proteomic studies are still based on non-quantitative methods for protein identification. These studies result in potentially large sets of identified proteins, where the biological interpretation of proteins can be challenging. Systems biology develops innovative network-based methods, which allow an integrated analysis of these data. Here we present a novel approach, which combines prior knowledge of protein-protein interactions (PPI) with proteomics data using functional similarity measurements of interacting proteins. This integrated network analysis exactly identifies functional similarity measurements of interacting proteins. This integrated network analysis exactly identifies network modules with a maximal consistent functional similarity reflecting biological processes of the investigated cells. We validated our approach on small (H9N2 virus-infected gastric cells) and large (blood constituents) proteomic data sets. Using this novel algorithm, we identified characteristic functional modules in virus-infected cells, comprising key signaling proteins (e.g., the stress-related kinase RAF1) and demonstrate that this method allows a module-based functional characterization of cell types. Analysis of a large proteome data set of blood constituents resulted in clear separation of blood cells according to their developmental origin. A detailed investigation of the T-cell proteome further illustrates how the algorithm partitions large networks into functional sub-networks each representing specific cellular functions. These results demonstrate that the integrated network approach not only allows a detailed analysis of proteome networks but also yields a functional decomposition of complex proteomic data sets and thereby provides deeper insights into the underlying cellular processes of the investigated system. 

Molecular & Cellular Proteomics 13: 10.1074/mcp.M113.032839, 1877–1889, 2014.

Proteome studies using mass spectrometry are a widely used technique for the analysis of protein samples and investigation of signaling pathways (1, 2). Albeit quantitative measurements are possible (2, 3), they are technically demanding and more costly than simple qualitative approaches, which usually produce a list of proteins identified in the analyzed sample. These lists are often hard to interpret and may include incomplete data because of the insufficient coverage of MS-based proteomics, for example for low-abundance proteins or proteins with specific physical characteristics (4). For subsequent analysis, single proteins or small subsets of proteins of interest are usually selected from the data sets on the basis of biological background knowledge, mostly because of their known association with particular biological pathways.

Protein-protein interaction (PPI) networks, which contain experimentally validated interactions between proteins, can supply a basic structure for the network context in and between pathways. Using PPI information to create biological networks of the analyzed data and searching for functional modules can therefore be a powerful tool for the identification of key pathways in cellular systems. In this context, a functional module is defined as a group or cluster of proteins having similar biological function in the same network neighborhood (5), which act in a concerted manner to accomplish specialized cellular functions. Much effort has been spent in recent years to optimize methods for identification of proteins ensuring the validity of measurements (6–8). The development of functional analysis methods and further biological interpretation of results, however, has just started. Previously, some approaches have already focused on the network investigation of proteomic data (9).

Data-driven approaches to analyze large-scale data sets can produce an unbiased view on the entire data set and help to focus on small subsets of functionally relevant proteins (10–15). Further methods for investigating functional and disease-related subnetworks of genes or proteins have also been introduced in recent years (16–22). However, these ap-
Functional Module Search in Protein Networks

Approaches are based on protein/gene information on the nodes and the mere topology of the PPI network and although some of them use edge weights (12, 18, 19), functional information for the interaction connecting these nodes is not used.

Large-scale bioinformatics analyses profit heavily from the development of a standardized functional annotation of the gene ontology (GO)\(^1\) (23). GO is a hierarchical structure containing functional terms, which are grouped into so-called ontologies. There are three main ontologies: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) containing a set of terms grouped hierarchically. Proteins are assigned biological terms according to their function in an either manual or automated way. Interestingly, two genes can be compared using the information from their assigned functional GO annotations, which is called semantic similarity. Several similarity measures have been defined between GO terms, which have been extended to similarity measures between two or more proteins (24). Functional similarity between two proteins is an adequate measurement for how well they interact in a biological network based on the hypothesis that proteins with related function tend to interact with each other (5).

Using the GO semantic similarity of two genes as a basis and integrating this knowledge to a predefined human PPI network of experimentally detected interactions from literature, we have developed an exact functional module detection algorithm for the analysis of qualitative proteomics data sets. Our investigation involves five important steps: a) the development of a functional interaction score for each edge based on the GO semantic similarity of the interacting proteins; b) the testing and validation of the information signal contained in the functionally enhanced human PPI network; c) the application of the new algorithm to a small proteomic data set of H9N2 virus-infected gastric cells (25); d) the extension of the method for the decomposition of networks into functional modules; and finally e) the characterization of different blood constituents based on a large proteome data set (26). Thus we demonstrate the benefits of combining proteome data with functionally scored PPI networks and illustrate multiple options for network analysis of qualitative proteomic data sets. To make this method available to all proteome researchers, we have created a website (http://gosim.bioapps.biozentrum.uni-wuerzburg.de/gosim.php) which gives a detailed tutorial of the algorithm including all scripts and programs required. This will help users to understand the method by examples and allows to apply the procedure step by step using own sample proteins.

\(^1\) The abbreviations used are: GO, Gene Ontology; PPI, Protein-Protein Interactions; BP, Biological Process; MF, Molecular Function; CC, Cellular Component; DAG, Directed Acyclic Graph; MICA, Most Informative Common Ancestor; BUM, Beta Uniform Mixture; ILP, Integer Linear Programming; PBMC, Peripheral Blood Mononuclear Cell.

**EXPERIMENTAL PROCEDURES**

**Human Interactome**—The protein-protein interaction (PPI) network was assembled based on interaction and site-specific phosphorylation information from Human Protein Reference Database (HPRD) (version 9) (27) and then combined with phosphorylation data from Phosphosite (28) to complement the phosphosite data set obtained from HPRD. For all analyses, the largest connected component consisting of 10,688 proteins and 55,195 interactions was used.

**Gene Ontology (GO) Functional Enrichment**—Gene Ontology (GO) information was extracted from the GO database (23) (accessed May 2011). This database contains biological terms, describing known gene function, which are hierarchically clustered and modeled by a directed acyclic graph (DAG) containing three main branches: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). Terms are structured according to the semantic hierarchy of biological terms with the most specific biological functions found at the lowest levels of the GO DAG.

All enrichment analyses of GO terms were performed with the BINGO plug-in (29) of the network analysis framework Cytoscape (30). Statistically significant categories were selected according to their \( p \) value using a hypergeometric test (29). All \( p \) values were adjusted for multiple testing using the Benjamini and Hochberg correction (31).

**Calculation of GO Semantic Similarity Scores for the Human PPI Network**—The Gene ontology semantic similarity scores were calculated based on the method of Fröhlich et al. (32) which calculates functional similarity between two genes by determining the average of best matching GO term similarity for both genes (GOSim package). This method is based on the approach initially developed by Schlicker et al. (33), which uses the probability of annotation of the most informative common ancestor (MICA) as a weighting factor to provide graph placement. The GOSim package is incorporated in the statistical software R (R version 2.13.0 (34)).

The functional similarities range between 0 and 1 with a value of 0 representing no similarity and 1 assigned to proteins with very high semantic similarity. The similarity of interacting proteins was individually computed on all three ontologies (Biological process (BP), Molecular function (MF), and Cellular component (CC)) for each interaction pair separately.

**Adjustment of Functional Interaction Scores**—To obtain an overview of the functional information in the PPI network and test the proportion of signal in the calculated GO scores (for BP, MF, and CC) we compared the GO-similarity measures in the actual PPI network with that measured on the background sets created by randomly rewiring the edges in the network. The human PPI interactome was randomized, keeping the number of edge occurrences for each gene constant. We then obtained an empirical \( p \) value for each GO score on each interaction based on the performed randomization and the real interactome. All resulting empirical \( p \) values were then used for fitting a BUM model to the interactome, which divides the scores into a signal and a noise component.

To investigate the network from all aspects simultaneously, a combination of the three main ontology scores was needed. Therefore, we used a method of \( p \) value aggregation, previously applied in microarray analysis to combine different sets of data (11). The calculated empirical \( p \) values for all three ontologies were thus combined into a single \( p \) value using a uniform order statistic (order 1) (11). We calculated a threshold \( p \) value (FDR = 0.03), which controls the false discovery rate for interaction scores. \( p \) values below this threshold were considered significant and scored positively whereas those above the threshold were assumed to have arisen from the null model and obtained negative scores. The aggregated \( p \) values from the BUM model were converted to composite network edge scores (= functional interaction scores) using the BioNet package (10). The
edges missing in all three ontology annotations (2456 edges) were assigned the average value of all edge scores. A detailed description of this method is available in the Supplementary Materials, Fig. S1.

**Node Scores**—Nodes in the network represent the proteins from the proteomic sample. The network node scores were calculated based on the presence or absence of a protein in the proteome sample and the interactome edge scores, where proteins from the proteomic sample were scored as the negative average value of their connecting edges. The rest of the interactome proteins were given the average of all edge scores. The role of the node score is to counteract the contribution of the functional interaction scores. This is similar to the role of the $\beta$ parameter used by Huang and Fraenkel (12).

**Signal and Noise Decomposition of GO-similarity $p$ Values**—The Beta Uniform Mixture (BUM) model introduced by Pounds and Morris (35) is a statistical method for decomposing raw $p$ values into a signal and a noise component. Here the distribution of $p$ values is regarded as a mixture of two components: One that describes the signal component, and one that is assumed to have arisen from background noise. It is well known from theory that $p$ values are distributed uniformly under the null hypothesis (noise). For the signal component, an overrepresentation of small $p$ values will be observed and the $p$ value distribution will have a peak near zero. In the BUM model, the signal component is modeled by a beta distribution, whereas the noise component is modeled by the uniform distribution. This leads to a mixture model of a $B(a; 1)$ beta distribution (signal) and a uniform distribution (noise) with mixture parameter $\lambda$ and shape parameter $a$ of the beta distribution. The log likelihood is thus defined as

$$\log \mathcal{L}(\lambda, a; x) = \sum_{i=1}^{n} \log(\lambda + (1 - \lambda)ax^{\lambda - 1})$$

A detailed derivation of this equation can be found in Dittrich et al. (11). Consequently, the maximum-likelihood estimations of the unknown parameters are given by

$$[\hat{\lambda}, \hat{a}] = \arg \max_{\lambda, a} \mathcal{L}(\lambda, a; x)$$

The parameter estimates have been obtained using numerical optimization as implemented in the BioNet package (10). Using the fitted parameters of this model the distribution of $p$ values can then be decomposed into a signal and a noise component. Based on this decomposition an upper bound of the noise proportion can be assessed by the so-called $\Pi$-upper value as detailed by Pounds and Morris (35). As this $\Pi$-upper value denotes an upper bound for the assessment of the noise proportion and consequently the value 1-$\Pi$-upper can be regarded as lower bound estimate of the signal proportion in the $p$ value distribution. This parameter can thus be used to obtain a (conservative) assessment of the signal content of a $p$ value distribution. In our analysis, a BUM model was fitted to the $p$ value distribution for each ontology separately.

**Module Detection Algorithm**—For network analysis we applied a recently devised algorithm (heinz, heaviest induced subgraph), which computes provably optimal solutions to the maximal-scoring subgraph (MSS) problem in reasonable running time using integer linear programming (ILP) (11). Given the node and edge scores as defined in the previous section, the algorithm identifies the highest-scoring connected subnetwork, where the score of a subnetwork is defined as the sum of the node and edge weights of an underlying spanning tree. The problem is closely related to the well-known prize-collecting Steiner tree problem (PCST); see (36) for a more detailed description of this relationship. Indeed we transform the MSS problem into a PCST problem and use the PCST algorithm by Ljubić et al. (37) to compute optimal solutions. In contrast to the classical Steiner tree problem, which tries to connect a given set of nodes using linking nodes in the network, the PCST problem identifies a subtree that maximizes the profit of all nodes contained the subtree minus the costs of all edges in the subtree. This optimization does not require the entire set of all input nodes to appear in the resulting solution but rather the set of maximally scoring nodes (here constrained by the functional interaction context as reflected by the edge weights). This is of particular importance for large data sets as it allows to focus on functionally relevant submodules by successive partitioning of the entire network as exemplified by the T-Cell proteome data set. Steiner tree models have also been used by other groups for functional module detection (12, 38). Here the heinz algorithm was used with node scores based on the pre calculated node and edge scores. When using the algorithm without functional interaction scores, proteins in the sample were assigned a constant positive node score, whereas all the rest of the proteins with known interactions (therefore available in the human interactome) obtained a constant negative score.

### RESULTS

**Analyzing Integrated Networks from Proteomic Data**—Large-scale proteome analysis has become an important tool for investigating cellular functions. Here, we introduce a new approach for proteome network analysis based on functional similarity between proteins and PPI data. Identification of functional modules within the network, containing proteins from the probe along with linking proteins from the human proteome, presents a more detailed picture of the network context and signaling mechanisms of the identified proteins. Therefore, we investigate large proteomic data sets with a novel approach, weighting nodes and edges of the network according to the functional similarity of the interacting proteins in the PPI network based on their Gene Ontology annotation (23) (Fig. 1), which is represented in a directed acyclic graph (DAG) with specificity of terms increasing toward the bottom of each branch (Fig. 1A).

First, we mapped all proteins identified in a sample to the human interactome network (10,688 nodes and 55,196 edges) where all interactions had been weighted with a score based on semantic similarity of the GO annotation of the interacting proteins. To achieve this we derived a novel score based on a previously developed measurement of semantic similarity (32, 33), which is adjusted and adapted to the network context. We will further refer to these scores as functional interaction scores as they represent the functional similarity between each interacting pair. The principle of our method is as follows. Proteins identified in the sample are assigned positive values derived from the functional interactions scores, whereas linking proteins from the interactome network (i.e., connecting proteins not identified in the sample) obtain a negative background score (Fig. 1B) (for details see Experimental Procedures). Subsequently, the algorithm searches for the maximally scoring subgraph, thus identifying the most likely functional module based on interaction and protein scores (Fig. 1A). In particular, the use of functional interaction scores ensures that the algorithm prioritizes paths with a high similarity over those with a low similarity (Fig. 1C). Linking proteins from the human interactome are included only if this...
FIG. 1. Flowchart of the subnetwork extraction algorithm: A, Proteins identified by mass spectrometry are mapped to the human interactome network. Functional interaction scores (edge scores) based on GO semantic similarity of the interacting proteins and protein scores (node scores) are integrated into the interactome. Functionally relevant modules can then be extracted using the module detection algorithm (right hand side). B, Calculation of protein scores. Proteins identified in the proteome sample are given a score based on the number of adjacent interactions (edges). The score for each node is the negative average of the scores of all adjacent edges. The remaining proteins are assigned the average of all interactome functional scores, which is always negative and ensures proper concentration of the solutions. Proteins from the proteome study contained in the resulting maximum scoring subnetwork are highlighted in red. C, Effect of functional interaction scores. From alternative equally scoring subnetworks, the one with a higher overall score (i.e. functionally most similar path) is preferred automatically. Example: Protein A and B are proteins identified in the sample, which are connected over the linking protein C, as the functional similarity between A-C and C-B is higher than any other connecting paths. In this sense prior knowledge of functional pathways is superimposed onto the often sparse data sets of identified proteins.
increases the maximum score of the resulting subnetwork. These additional proteins can represent important players for signal transduction, which might potentially be present in vivo without having been identified in the sample (i.e., potential false negatives). As proteome studies are often fractionated and detection of some proteins proves to be extremely difficult, this approach is useful for unraveling the network context of identified proteins along with proteins missing in the original sample because of technical difficulties.

Scoring Proteins and Interactions Based on Functional Similarity—A prerequisite for functional interaction scores is the availability of information, which measures the similarity of two proteins in terms of their GO semantic similarity (24). The rationale behind a scoring system based on GO similarity is that proteins with a high functional similarity have a higher probability of being involved in the same cellular processes, thus the interaction between them will be given a higher priority. In this sense, the edge score reflects the functional relevance of the particular interaction in the PPI network. To estimate the information content of the interactions in the network in terms of functional scores, we compared the original functional scores with the scores of the network after 1000 randomizations where the size of the network and the distribution of node degree (number of interactions) were kept constant, but the topology of the edges was changed. We then extracted empirical p values based on this comparison to the rewiring background null model (Experimental Procedures). The distribution of these p values clearly evidences a strong signal of the interactome network in terms of semantic similarity. To quantify this we used a beta uniform mixture model (BUM) to decompose the p value distribution into a signal and a noise component (represented by the \( \Pi_{\text{upper}} \) value introduced by Pounds and Morris (35), for a more detailed explanation see Material and Methods). The model clearly captured a signal in the interactome, reflected by the accumulation of small p values in the histogram (1 - (\( \Pi_{\text{upper}} \) value) = 54%) (Fig. 2A). Here, the density describes relative frequency of p values depicted in the histogram (which is a normalized frequency measure so the area of the entire histogram sums up to one) and the overlaid curve visualizes the predictions of the fitted model. Comparing two randomized networks, the obtained p values were uniformly distributed (Fig. 2B), indicating the expected lack of signal content.

Thus, we demonstrated that the human interactome contains interacting proteins with high functional similarity significantly more often than the randomized network model. We fitted a BUM model to the p values of all three GO main categories (BP, MF, and CC) and subsequently aggregated these p values into one combined p value using an order statistic as previously described (see “Experimental Procedures, Adjustment of Functional Interaction Scores”) (11). A detailed flowchart of all involved steps is presented in the Supplemental materials (supplemental Fig. S1). When comparing the information content of each individual score,
the BP scores showed the highest signal content (54%), followed closely by the combined functional interaction score based on all three adjusted ontology scores (44%) (supplementary Fig. S2). The higher signal content of BP terms alone can be explained by the high annotation quality and abundance of BP terms when compared with MF and CC terms.

Identification of Functional Modules in the Network—Integrating functional interaction scores to the module detection algorithm helps to connect proteins identified in a sample based on the functional context of the proteins within the network. To investigate the effect of functional interaction score for the network analysis, we applied the extended functional modules approach on a small proteomic data set from a human gastric cell line (AGC, originating from a gastric epithelial cell line) analyzed after infection with avian influenza virus (H9N2). The study identified 22 proteins using mass spectrometry approaches based on 2-DE coupled with MALDI-TOF MS and nano-ESI-MS/MS (25). We performed two separate analyses of this protein sample: one with and one without interaction scores in the algorithm (see Experimental Procedures). As the functional interaction scores are based on GO annotations, which are extracted from literature information, we expected the algorithm including functional interaction scores to derive functional modules with higher functional similarity and higher number of publications supporting each interaction. Information of the edges is based on well-studied interactions, which are tractable experimentally and can serve as a solid basis for hypothesis generation after network analysis.

The method incorporating functional interaction scores (Fig. 3B) yielded a module of 39 proteins, whereas the method without interaction scores extracted 37 proteins (Fig. 3A). Both networks contained partially overlapping linking proteins and interactions (14 interactions and six linking proteins are the same). There were 28 proteins in common for the two solutions along with unique proteins, which appeared only in
the solution with functional interaction scores (AKT1, EPB41, GNG4, HRAS, HSPA4, JAK2, KRT5, PDK1, RAF1, YWHAB, and ACTG1). Comparing the literature support of the resulting modules we found a higher number of literature citations for each interaction in the solution with functional scores compared with the solution without scores (Fig. 3). In the network with interaction scores there were six interactions with more than one citation (Fig. 3B), indicating the algorithm chooses well-supported interactions.

A detailed examination of the interaction paths of both modules confirmed that the introduction of functional scores improves the biological interpretation of the observed cell response after virus infection. Although cytoskeleton proteins, mainly keratins, were overrepresented in both modules (data not shown), the keratin cluster was connected to different nodes in the two solutions indicating that functional information influences the resulting network. In particular, the keratin cluster in the functional score solution additionally contained keratin 1 (KRT1), complementing the keratin cluster. Notably, this subgraph provides more defined pathway connecting keratins to the rest of the proteins. The linking kinase RAF1 plays a major role under stress conditions and the phosphorylation-dependent disruption of RAF1 interaction with the keratin protein KRT18 has been previously described during stress response (39). The Ras GTP-ase (HRAS) found in the solution with functional interaction scores activates RAF1 (40) in the signaling cascade and indicates how activation of RAF1 might have been triggered under stress conditions in gastric cells.

To examine the differences between the two networks modules in respect to their interaction scores and functional similarity, we further considered the functional interaction scores of the network where they were initially disregarded (Fig. 3A) and tested whether there is a significant difference between the two score sets (supplemental Fig. S3; supplemental Tables S1 and S2). As expected the solution with functional interaction scores showed a significantly higher similarity score (Mann-Whitney U test, \( p = 0.002 \), indicating that the interacting pairs in this solution were connecting functionally similar proteins.

This example demonstrates that the algorithm including interaction scores improves the functional relevance of the obtained modules and focuses on modules and paths containing substantial biological information to explain cellular responses. Furthermore, proteins initially missing in the measured sample, but crucial for signal transduction, can be identified.

Further applications include the analysis of a mass spectrometry data set from vincristine-resistant human gastric cancer cells, where up- and down-regulated proteins in the sample were visualized in the network solution (supplemental Fig. S4).

Module-based Functional Characterization of Blood Constituents—Using our novel approach we characterized functionally distinct protein data sets from a recently published mass spectrometry study of human blood cell constituents (26). We applied the module detection algorithm to the protein samples of all seven blood constituents separately: plasma, T cells, neutrophils, monocytes, platelets, erythrocytes, and peripheral blood mononuclear cells (PBMCs). The resulting module of each constituent was subsequently tested for functional enrichment of specific GO terms. At first, we performed a group-specific analysis and set the significance threshold for terms to a corrected \( p \) value of at least \( 1 \times 10^{-5} \) in at least one of the blood constituents. Unspecific GO terms (containing less than three or more than 600 proteins) were excluded from the analysis. Thus, enriched functions in one or more of the data sets could be investigated and visualized as a heatmap (Fig. 4). Clustering of GO terms was performed with hierarchical clustering based on the negative logarithm of the GO enrichment \( p \) value.

Blood constituents clustered in a way consistent with their differentiation stages. As expected, plasma proteins built an out-group distant from all other blood constituents because they represent proteins secreted into the blood plasma. Two big groups were formed, which can be assigned to the myeloid and lymphoid progenitor cells during hematopoiesis. The myeloid progenitor cells erythrocytes, platelets, and neutrophils were separated from the lymphoid derived cells: lymphocytes and PBMCs (representing a mixture of cells with a round nucleus, including T cells and monocytes). Monocytes clustered in this group, although they are derived from myeloid progenitor cells, probably because of their functional similarity with differentiated T cells and PBMCs.

GO enrichment analysis revealed terms specific to platelets such as the term “platelet activation” and hemostatic processes, whereas other terms such as “cell killing” and “leukocyte-mediated cytotoxicity” were enriched in monocytes and biologically differentiate these from other blood cell types. Ubiquitin and proteasome processes were highly enriched in three cell types: T cells, monocytes, and PBMC. The term “translation” was overrepresented in T cells and PBMC when compared with all other constituents. Monocytes were enriched for oxygen species metabolism functions (Fig. 4).

In a next step we focused on biological processes characterizing a single blood constituent by taking in account significantly enriched terms in exclusively one cell type with a \( p \) value of at least 0.05, which were not overrepresented in any other data set (supplemental Table S3). T cells were characterized by functions such as “positive regulation of vesicle fusion,” “T-cell receptor signaling” and “positive regulation of apoptosis,” whereas monocytes were distinguished by leukocyte and neutrophil chemotaxis. On the other hand, neutrophils had an overrepresentation of the term “response to bacterium,” consistent with their role in bacterial defense and inflammation (41, 42). Plasma proteins showed the most significantly enriched differential terms, associated mainly with acute inflammatory response and complement activation.
Functional Module Search in Protein Networks

- Proteasome
  - Overrepresentation
  - -log(p-value)
  - Molecular & Cellular Proteomics 13.7

- Translation
  - Actin cytoskeleton assembly

- Membrane organization

- Oxygen species metabolism
  - Cell killing
  - Platelet activation

- Hemostasis / blood coagulation

- Acute immune response

- Relative expression of module members in plasma cells, monocytes, T-cells, and PBMCs.
PBMCs were mainly represented with functions related to antigen processing and presentation, in accordance with their role in immune regulation, whereas enrichment analysis in platelets revealed processes such as positive regulation of endocytosis and cell adhesion. Similar results were also obtained during functional analysis of the platelet proteome (43), where endocytosis was the maximally enriched pathway in platelets.

In conclusion, specific functions for each cell type could be extracted, indicating the resulting modules are cell-type specific and focus on the characteristic functional modules of the analyzed cells. This includes specific terms and modules for T-cell receptor signaling, neutrophil response to bacteria, inflammation in plasma proteins, and endocytosis in platelets.

**Partitioning the T-cell Network into Distinct Functional Modules**—The large size of networks from proteomic analyses makes it difficult to search for a particular smaller functional module inside the resulting network. Therefore, we introduce a method for decomposing the network into smaller functional modules of a predefined size. The resulting network of T cells from the previous analysis of blood constituents was used as an example.

From altogether 970 T-cell proteins, 861 could be mapped to the human interactome. We first assembled the smallest subnetwork containing all T-cell proteins in the data set comprising 3004 interactions between 1026 proteins (Fig. 5A). We partitioned this large network into non-overlapping modules of a predefined size (in this case 50 proteins) applying the module detection algorithm with and without interaction scores. Finally, all resulting modules were tested for functional enrichment.

The analysis without interaction scores yielded only very few significant GO-terms, which were mainly very general and unspecific to T-cell signaling (supplemental Table S4).

In contrast, the analysis with functional interaction scores yielded much more significant terms and not only general processes such as ubiquitin/proteasome regulation (regulation of ubiquitin-protein ligase activity: 2.39 $\times 10^{-6}$) and RNA splicing (1.32 $\times 10^{-5}$) (Fig. 5B), but also many T-cell specific pathways. In the second solution the terms T-cell differentiation (2.73 $\times 10^{-5}$) and T-cell receptor signaling pathway (2.72 $\times 10^{-5}$) were enriched when compared with the set of all T-cell proteins (Fig. 5C). Also, processes associated with signaling, kinase cascades and phosphorylation were overrepresented. The third solution contained enrichment of the following terms associated with cell movement: actin cytoskeleton organization (4.33 $\times 10^{-5}$), leukocyte cell-cell adhesion (1.29 $\times 10^{-5}$) and regulation of actin cytoskeleton organization (4.33 $\times 10^{-5}$) (Fig. 5D). A detailed summary of all enriched GO terms and their $p$ values can be found in the supplemental Tables S4 and S5. These results evidence that the extraction of characteristic modules heavily benefits from the integration of functional interaction scores and allows a semantic partitioning of the network in characteristic submodules.

**DISCUSSION**

Fast evolving mass spectrometry techniques generate large proteome data sets and only systematic analysis methods can reveal new insights into functional properties of the cellular system (44, 45). In contrast to transcriptomics, where often large numbers of biological replicates from genome-wide measurements are available along with a set of well-established analytical methods, proteomics still faces challenges with genome-wide quantitative measurements and low numbers of replicates. This is particularly true for qualitative proteomics, where only the presence or absence of proteins is determined without quantifying the abundance or change of protein expression. The biological interpretation of lists containing identified proteins in the sample is often challenging. Moreover, because of technical difficulties in protein detection (e.g. transmembrane proteins in gel-based systems) the probability of false negative results (i.e. missed proteins) is relatively high, which makes the meaningful interpretation of these results even more difficult.

Here, integrated network analysis can be particularly useful to overcome some of the experimental and statistical limitations. Interaction networks provide a wealth of information, which can be exploited by integrated network analyses. In particular, these molecular networks typically provide a functional scaffold on which the systemic analysis and interpretation of proteome data can be found. Integrated network analysis can pick up proteins, which have not been directly detected in a sample (potentially false negatives) but are part of functional complexes or pathways; the presence of these proteins is inferred by their functional network context. This is one of the central points, where integrated network analysis also provides the potential to improve the protein detection error rate (4). For the analysis of a proteome data set it is

![Fig. 4. Functional proteome characterization of human blood constituents. For all available proteome data sets from seven blood constituents (plasma, T cells, neutrophils, monocytes, platelets, PBMC (peripheral blood mononuclear cell), and erythrocytes) functional modules have been identified and characterized by functional GO term (biological processes) enrichment analysis. Clustering of module GO terms (based on phred-scaled enrichment $p$ values) with an enrichment $p$ value of at least $1 \times 10^{-5}$ in at least one of the cell types (126 GO terms in total) reveals a blood constituents clustering consistent with their developmental stages. Significantly overrepresented terms are depicted in red, whereas the remaining terms are shown in blue. Maximally enriched terms of high biological importance are marked additionally with an asterisk. Specific biological processes characterizing single types of blood constituents are overrepresented, e.g. platelet activation (platelets), cell killing and oxygen species metabolism (monocytes), translation and protein degradation (T cells, PBMC), and acute immune response (plasma). Other terms characterize specific subgroups, e.g. protein degradation process (monocytes, plasma, and PBMC), hemostasis, blood coagulation, and wound healing (plasma and platelets).](https://example.com/Fig4.png)
therefore highly desirable to integrate and examine the interaction partners and network properties (9, 46).

However, merely mapping the identified proteins in a sample to a protein-protein interaction network is not enough for profound biological investigation and interpretation of the data. Plainly, a minimal analysis could simply focus on the connected components of the proteins in the network. However, if the number of identified proteins is large, the connected network components derived may become huge and thus will not be amendable for biological interpretation. Moreover, the modules might become “functionally diluted” in the sense that they comprise different parts of many diverse pathways. Furthermore, because of the high connectivity of interactome components often more than one possible path can be found to connect the sample proteins. The resulting modules might not be unique, as there might be more than one possibility for connecting the positive nodes with a maximum-scoring subgraph.

One solution to this problem is the integration of quantitative edge weights. A weighting score of the edges allows concentrating the resulting submodule onto paths with higher functional resemblance of interacting proteins. Here we introduced a scoring based on GO similarity, which ensures that proteins are connected by functionally similar and thus biologically more relevant interactions and it can also serve as an indicator of interaction confidence in the network. This is additionally supported by the observation that high scoring interactions have a higher literature support (supplemental Fig. S3).

Protein interactions are weighted based on GO annotations, which comprise information on three main ontologies. Based on this annotation functional similarity between pro-
teins can be quantified on each of the three ontologies separately (47). However, to obtain a combined score it is necessary to aggregate the similarity of the three ontologies into one similarity measure. Previously published methods essentially average the score over all three categories for an aggregated score (33, 48). The main problem with this aggregation is that if GO information is missing even for one ontology, the final score is not defined. To circumvent this problem we introduce a new aggregation statistic based on an order statistic which automatically handles missing ontology information intrinsically (11). Furthermore, this scoring is adapted to the context of network analysis based on a statistical model that measures the joint signal content of the network topology and GO similarity.

As GO annotation is an important measure to validate PPIs detected in high throughput interaction studies, and vice versa PPI information contributes to GO annotation the usage of GO as basis for functional interaction scores may possibly lead to a bias through redundancy.

This might partially contribute to the signal content in terms of functional similarity for the interacting proteins in the network (see Fig. 2). Interestingly, a recent study analogously found that the number of PPI interactions between genes within a GO-term is significantly higher than would be expected for random networks (49).

Integrating these scores into the proteomic network, it can then be decomposed into functional modules using an exact algorithm (11). Thus our algorithm provably identifies the maximal scoring subgraph in networks, where nodes and edges are scored (10, 11, 50). This optimal solution contains the maximum positive scoring nodes (proteins) and interactions connecting them. Modules identified by the algorithm can represent parts of protein complexes or pathways with the same function making this approach highly suitable to identify protein complexes or pathways.

Although several methods of module detection have been published in recent years, most of them focus on the analysis of gene expression data in the context of PPI networks pioneered by the work of Ideker et al. (13). Most of these works use heuristic algorithms which generally only give approximate instead of exact optimal solutions, often resulting in different modules (for an evaluation see (50)). As these approaches usually require quantitative measurements at the network nodes they do not easily transfer to proteome data where often only proteins detection calls are available. Several methods implemented in the MATISSE framework (17, 51, 52) search for modules in PPI networks using gene expression data. Based on a heuristic search algorithm sub-modules in a given PPI network showing highly correlated expression profiles are identified (51). Although probabilities for individual interactions might be integrated (52) the explicit use of GO similarity as an additional metric for scoring interactions is not available. Methods directly investigating Proteome data (without the requirement for quantitative measurements) include the PPI-spider algorithm approach by Antonov et al. (9) which incrementally adds interconnecting network nodes to the list of identified input proteins in order to link the entire set of input nodes together. This approach does not use edge weights to prefer edges with a potential higher functional relevance and consequently does not try to identify an “optimal” module in terms a clearly defined objective function. Thus, a particular advantage of our algorithm is the combination of optimized and improved functional scores with an exact algorithm for the identification of provably optimal solutions. In fact, the method not only uses GO scores but it optimizes these scores by adapting them to the network context after a signal and noise decomposition, thereby focusing primarily on relevant interactions.

Using this integrated network approach we first studied a small data set derived from virus infected human gastric cells and compared it to an approach without functional score information. Clear differences in the topology and biological interpretation of the two resulting networks became apparent: Although the main functional complex of keratin proteins was maintained in both, the connecting path to the other proteins in the sample was different.

A detailed comparison between some of the given interactions is presented in the supplement (supplemental Table S6). The most prominent example for a different path chosen by the functionally enhanced algorithm is the interaction between the keratin complex (KRT18) with YWHAB (a 14–3-3 protein) and the kinase RAF1. The solution without interaction scores contains the Protein kinase C (PRKCE) as link to the keratin complex. Although the phosphorylation of KRT18 at Ser53 by PRKCE is already known, the functional role of this phosphorylation has not yet been confirmed. It has been speculated that it may play an in vivo role in filament reorganization (53, 54). Focusing on the solution with functional scores, the association of keratin 18 with 14–3-3 proteins and thereby with Raf1 kinase is known to regulate cell signaling (39). During cell stress the binding between Raf1 and the keratin cluster is disrupted (39), which might be a useful hint, as a virus infection causes stress to the gastric cells. KRT18 phosphorylation is reported to regulate various keratin functions including the binding to 14–3-3 proteins, involvement in the modulation of cell cycle progression and organizing keratin filaments (55, 56) along with a role in keratin protein turnover by ubiquitination (57) or during apoptosis (58). There is also evidence of a possible phosphorylation of KRT18 by Raf1, which causes the disruption in the complex (39). Hence, including functional scores enhances the biological insights of the resulting functional modules as shown here for gastric cells.

Based on functional modules we characterized a large-scale proteomic data set of blood constituents. An interesting aspect was the enrichment of proteasome and ubiquitin processes in particularly three cell types: T cells, monocytes and PBMC. The existence and role of the proteasome in immune
as well as non-immune cells has recently been reviewed by Ebstein et al. (59). Interestingly, these three cell types constitutively express three proteasome subunits (PSMB8, PSMB10, and PSMB9). These are normally not included in the proteasome but induced after stimulation building the so-called immunoproteasome, which plays an important role in antigen presentation by MHC class I molecules, cell proliferation, cell signaling and cytokine production (59). This type of proteasome has been identified mainly in professional antigen-presenting cells such as macrophages (60), activated or resting B-cells (61) and monocyte-derived dendritic cells (62). In our results monocytes are enriched for the GO term “proteasome.” As monocytes replenish the pool of macrophages and dendritic cells in the body, we can assume a presence of immunoproteasome in these cells as well. T cells show a constitutive expression of all three molecules and the immunoproteasome facilitates protein homeostasis and cell proliferation in these cells (63). Because PBMCs are mainly a mixture of macrophages, monocytes, and lymphocytes we can expect that we would identify the immunoproteasome here as well. The exclusive expression of the immunoproteasome subunits in the monocyte, T cell and PBMC samples could explain the relative overrepresentation of “proteasome” related GO terms when compared with the profile of the other blood constituents.

Although qualitative proteomic data sets only measure the presence of proteins in a sample, quantitative proteomics can reveal exact changes in abundance of signaling proteins. As technologies are constantly evolving more and more quantitative data is becoming available. Hence, the focus of our future work will be on the extension of the algorithm to the analysis of quantitative proteome data and particularly the more complex analysis of phosphoproteome data.

In summary, we have demonstrated in this study that a detailed investigation of integrated protein-protein interaction networks can be accomplished by using a functional module-based network algorithm. The presented approach allows a functional decomposition of complex proteomic data sets and offers a systems biological perspective on proteome data in the context of cellular interaction networks. It provides insights into pathway structures beyond the scope of traditional analysis approaches, in particular hints in large-scale proteomics and description of networks, modules, and submodules with specific function.

* This work was supported by Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research)/Systems Biology of ADP Receptor Activation (D.B. and S.N.) and the DFG (Transregio 124, TP B2).
[S] This article contains supplemental Figs. S1 to S4 and Tables S1 to S6.
¶ Both authors contributed equally to this work.
† To whom correspondence should be addressed: Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, D-97074 Würzburg, Germany. Tel.: +49 931-845577; E-mail: marcus.dittrich@biozentrum.uni-wuerzburg.de.

REFERENCES
1. Choudhary, C., and Mann, M. (2010) Decoding signalling networks by mass spectrometry-based proteomics. Nat. Rev. Mol. Cell Biol. 11, 427–439
2. Freytag, C., von Kriegsheim, A., Matallanas, D., and Koich, W. (2008) Proteomics and phosphoproteomics for the mapping of cellular signaling networks. Proteomics 8, 4402–4415
3. Bachi, A., and Bonaldi, T. (2008) Quantitative proteomics as a new piece of the systems biology puzzle. J. Proteomics 71, 357–367
4. Goh, W. W., Lee, Y. H., Chung, M., and Wong, L. (2012) How advancement in biological network analysis methods empowers proteomics. Proteomics 12, 550–563
5. Barabasi, A. L., Gulbahce, N., and Loscalzo, J. (2011) Network medicine: a network-based approach to human disease. Nat. Rev. Genet. 12, 56–68
6. Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O., and Aebersold, R. (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol. Cell. Proteomics 8, 2405–2417
7. Xiao, C. L., Chen, X. Z., Du, Y. L., Sun, X., Zhang, G., and He, Q. Y. (2013) Binomial probability distribution model-based protein identification algorithm for tandem mass spectrometry utilizing peak intensity information. J. Proteome Res. 12, 328–335
8. Haas, W., Faherty, B., Gerber, S., Elias, J., Beausoeil, S., Bakalarski, C., Li, X., Villén, J., and Gygi, S. (2008) Optimization and use of peptide mass measurement accuracy in shotgun proteomics. Mol. Cell. Proteomics 7, 1326–1337
9. Antonov, A. V., Dietmann, S., Rodchenkov, I., and Mewes, H. W. (2009) PPI spider: a tool for the interpretation of proteomics data in the context of protein-protein interaction networks. Proteomics 9, 2740–2749
10. Beisser, D., Klau, G. W., Dandekar, T., Muller, T., and Dittrich, M. T. (2010) BioNet: an R-Package for the functional analysis of biological networks. Bioinformatics 26, 1129–1130
11. Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Muller, T. (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 24, i223–i231
12. Huang, S. S., and Fraenkel, E. (2009) Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. Sci. Signal. 2, ra40
13. Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002) Discovering regulatory and signaling circuits in molecular interaction networks. Bioinformatics 18, S233–240
14. Scott, M. S., Perkins, T., Bunnell, S., Pepin, F., Thomas, D. Y., and Hallett, M. (2005) Identifying regulatory subnetworks for a set of genes. Mol. Cell. Proteomics 4, 683–692
15. Zhang, S., and Zhao, Z. (2011) GenRev: Exploring functional relevance of genes in molecular networks. Genomics 99, 183–188
16. Zhao, X. M., Wang, R. S., Chen, L., and Aihaara, K. (2008) Uncovering signal transduction networks from high-throughput data by integer linear programming. Nucleic Acids Res. 36, e48
17. Uilitsky, I., Krishnamurthy, A., Karp, R. M., and Shamir, R. (2010) Optimally discriminative subnetwork markers predict response to chemotherapy. Bioinformatics 26, 1129–1130
18. Vandin, F., Upfal, E., and Raphael, B. J. (2011) Algorithms for detecting significantly mutated pathways in cancer. J. Comput. Biol. 18, 507–522
19. Qiu, Y. Q., Zhang, S., Zhang, X. S., and Chen, L. (2010) Detecting disease associated modules and prioritizing active genes based on high-throughput data. BMC Bioinformatics 11, 26
20. Su, J., Yoon, B. J., and Dougherty, E. R. (2010) Identification of diagnostic subnetwork markers for cancer in human protein–protein interaction network. BMC Bioinformatics 6, S9
21. Chowdhury, S. A., and Koyuturk, M. (2010) Identification of coordinately dysregulated subnetworks in complex phenotypes. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 133–144
22. Dao, P., Wang, K., Collins, C., Esté, M., Lapuk, A., and Sahinalp, S. C. (2011) Optimally discriminative subnetwork markers predict response to chemotherapy. Bioinformatics 27, 205–213
23. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat.
