From proteomics toward systems biology: integration of different types of proteomics data into network models

Sangchul Rho*, Sungyong You*, Yongsoo Kim* & Daehee Hwang*
School of Interdisciplinary Bioscience and Bioengineering, Pohang University of Science and Technology, Pohang, Korea

INTRODUCTION

Mammalian organisms consist of various systems, i.e., bodies, organs, tissues, cells, and subcellular compartments, and these systems carry out diverse fundamental activities by signaling different biological networks defined by their nodes (e.g., DNA, mRNAs, proteins, and metabolites in cellular systems) and edges (the various interactions between nodes). When perturbed by an environmental or genetic event (e.g., diseases, nutrient changes, exposures to pathogens and harmful substances), one or more networks or particular portions of networks, called network modules, become activated to execute appropriate functions. Malfunctions of these networks or modules result in failures to respond appropriately, and thus give rise to diseases.

The advances in high-throughput technologies that facilitate the probing of both nodes (e.g., abundances of mRNAs, proteins, and metabolites, and types and degrees of post-translational modifications-PTMs) and edges (e.g., protein-protein, protein-DNA, chemical-protein, and chemical-DNA interactions; abbreviated to PPIs, PDIs, CPIs, and CDIs, respectively) in biological networks offer new opportunities to understand how living organisms execute necessary functions at system levels via complex networking operations (1, 2). These systems biology approaches to the understanding of complex network operations typically involve the following three cardinal processes (3): 1) the generation of global data after a perturbation (e.g., data on mRNA abundances and signaling pathway phosphorylation in diseased systems), 2) the integration of such information into network models that describe key biological events arising from perturbations (e.g., abnormal signaling in lung cancer systems), and 3) the generation of experimentally testable hypotheses concerning the mechanisms underlying key processes (e.g., mechanisms associated with disease initiation and progression). These three processes are achieved by identifying key network modules and exploring their dynamic transitions after perturbations.

Over the past decade proteomics studies have been considered to be of central importance to biologic system studies, and have enhanced our knowledge of the functions of biological networks (e.g., signaling, and gene and metabolic regulation) by generating a tremendous amount of information on; cellular states (at different levels), genomics (e.g., DNA sequencing and ChIP-sequencing), and transcriptomics (e.g., microarrays and Serial Analysis of Gene Expression-SAGE). These proteomic technologies can be largely categorized as being antibody or mass spectrometry (MS) based (4, 5). The latter provides a wider spectrum of information than the former on the functional states of proteins, and on their abundances, modifications, and interactions at different levels. Thus, proteomics has played crucial roles in systems biology research (2, 6) by providing valuable sets of information for integrative network modeling at system levels. However, despite these advantages, proteomics technologies also suffer from undersampling issues related to small detected proteome sizes, i.e., they are
incapable of managing the several thousand proteins in complex samples (7).

Many approaches have improved the performances of MS-based proteomic analysis, e.g., sample preparation methods, Liquid Chromatography (LC), and MS related improvements have been proposed to overcome these undersampling issues (7).

This review demonstrates how current proteomics technologies can help us understand complex network operations at system levels. Subsequent sections are organized as follows: i) systems biology approaches using proteomic data, ii) a brief summary of proteomics technologies, iii) an overview of the integrative proteomics data analysis pipeline (IPDAP), iv) the reconstruction of networks including key network modules, v) an exploration of network models to generate testable hypotheses using Cytoscape and its plugins, and vi) a discussion concerning the further integration of other types of non-proteomic data.

**Systems biology approaches using proteomic data.**

Systems biology approaches can also be applied to proteomic data, as shown in Fig. 1. First, problems of concern are defined for biological and medical systems. For example, the identification of the specific cellular processes of embryonic stem (ES) cells, and their key regulators, key mechanisms of amyloid beta accumulation, and the effects of chronic inflammation on cancer progression and metastasis. Second, the relevant biological systems are perturbed by various means to collect informative datasets intended to answer given questions. For example, when investigating cellular processes specific to ES cells, growth factors could be administered to ES cells and samples could be collected at certain times as differentiation proceeds. Such time intervals should be carefully determined to provide meaningful information regarding early and late differentiation molecular signatures, which allow the identification of cellular

---

**Fig. 1.** A systems biology framework using proteomic data, incorporating: 1) problem formulation, 2) system perturbation, 3) the generation of comprehensive sets of protein abundances, PTM extents, and interactions, 4) the identification of differentially expressed proteins and proteins with different PTM states, 5) network modeling and analysis using various tools (refer to the text for detailed descriptions), 6) generation of hypotheses (the red line in the networks) and their validations (refer to the text). If experimental testing does not succeed after an initial cycle, additional experiments can be designed to collect previously unavailable information, and the same computational scheme reapplied. This process can be repeated until a working hypothesis is generated.
From proteomics toward systems biology
Sangchul Rho, et al.

processes specific to undifferentiated cells. Preliminary RT-PCR or western blots can be carried out for known markers of differentiation stages to better determine time frames. Third, various types of comprehensive proteomic data can be generated from samples collected during the course of differentiation. For example, when investigating the effects of chronic inflammation on tumor progression, MS-based methods can be used to generate multiple types of complementary proteomic datasets, such as: 1) protein abundances, 2) extents of phosphorylations of key signaling proteins, 3) amounts of ubiquitinated proteins, and 4) protein interactions where no interactions are known (refer to 'subproteome capturing methods' and 'labeling methods for protein quantification' in Table 1).

The following three computational steps are performed for all collected data:
1. The identification of proteins whose abundances and extents of phosphorylation and ubiquitination are found to significantly vary (8). For example, in the case of chronic inflammation, this might involve a set of Differentially Expressed Proteins (DEPs) (i.e., a subset of interleukins), whose abundances are increased, and a set of signaling proteins, such as, JAK kinases and stats showing increased phosphorylation.

2. The modeling of networks describing the interactions between selected proteins, using PPIs and PDIs, and thus, of the relationships between biological processes in which these proteins are involved (for an example of a network refer to Fig. 1, and for further details on network modeling section refer to or Fig. 3). Network models can be visualized using Cytoscape (9) and then explored using several Cytoscape plugins to identify: i) enriched Gene Ontology (GO) terms; ii) the pathways of the protein components overrepresented in networks; iii) clusters of network nodes (also called network modules), each of which represents a biological process in which corresponding nodal proteins are involved; iv) active clusters among network clusters; and v) expansions of particular network portions (refer to the network analysis section and Fig. 4).

3. The generation of hypotheses for given problems by analyzing interactions between key network modules and their dynamic transitions. To analyze the effects of chronic inflammation on metastasis, we could identify network modules induced by chronic inflammation and determine how these network modules affect the levels of cytokines/chemokines and their receptors (e.g., CCL12 and CXCR4), which have been reported to be associated with metastasis (10), and then identify relationships between identified network modules and determine how the activities of proteins relate to such modules (i.e., abundances and phosphorylation states) in time. By collectively integrating this information, a mechanistic relationship between the source of chronic inflammation and CXCR4 expression can be predicted based on a knowledge of pathways including the source, intermediate proteins, and CXCR4 expression (refer to the pathway indicated by the line on the network in Fig. 1).

Finally, mechanistic hypotheses should be experimentally validated using various molecular and cellular biology tools, such as, siRNAs, in cell line models, knock-out mice, and chemical inhibition experiments in animal models, and if experimental testing fails, supportive experiments can be designed to collect missing information. Finally, the computational steps can be reapplied to integrate both old and new data.

**Table 1.** Types of proteomics data and the methods typically being used during sample preparation, LC-MS/MS analysis, and computational analysis (for the detailed descriptions of these methods, refer to the text and the references provided)

<table>
<thead>
<tr>
<th>Data type</th>
<th>Sample preparation</th>
<th>LC-MS/MS</th>
<th>Computations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abundances</td>
<td>• Isotopic labeling: SILAC, ICAT, iTRAQ, etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Label free: Spectral counting, alignment-based analysis, etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTMs</td>
<td>• Affinity capture: IMAC, TiO₂, dendrimer (52), and PAC (53)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>• Affinity capture: IMAC, TiO₂, dendrimer (52), and PAC (53)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycosylation</td>
<td>• Lectin affinity chromatography (LAC) (54)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Hydrazide resin (55)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ubiquitination</td>
<td>• Affinity capture: Nickel-affinity chromatography (56), immunoaffinity purification, etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interactions</td>
<td>• Fragmentation of large peptides : ECD (59), ETD (60)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPIs</td>
<td>• MASCOT, ProteinProspector(64), SEQUEST, GutenTag (65), XITANDEM, etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Spotfire (66), Cytoscape, Pathway Studio (67), etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPs</td>
<td>• Chemical affinity matrices (58)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Bead based method</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• STITCH, Sonar MS/MS, etc.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ta to generate a working hypothesis.

**Generation of proteomic data using mass spectrometry-based techniques**

MS-based proteomics technology, a crucial component in systems biology research, provides different types of global data. MS is an analytical tool that measures the mass-to-charge (m/z) ratios of ionized analytes (i.e., proteins or peptides in the case of proteomics) (11). Moreover, the measured peak intensities of certain peptides are proportional to their abundances (see Fig. 2A). To identify peptides detected in first MS scans, the fragment of a particular peptide (e.g., the one indicated by the arrow in Fig. 2A) can be isolated and further fragmented by Collision-Induced Dissociation (CID). The second MS scan (hence MS/MS analysis) then represents the fragmentation pattern of a fragment of the required peptide (Fig. 2B). Subsequently, the protein concerned can be identified by searching for a peptide that produces the observed peak pattern (3, 12). In MS-based proteomic analyses, MS is commonly combined with LC systems that separate peptides by molecular weight (LC-MS/MS analysis) to effectively analyze complex samples. However, due to the detection limit of MS, only abundant peptides can be detected (commonly referred to as an undersampling issue) when a sample includes a large number of proteins at comparable concentrations (7). However, several thousands of proteins can be detected using current technologies when additional fractionation methods, such as, 2-D gel electrophoresis (13) or Multidimensional Protein Identification Technology (MudPIT) using Strong Cation Exchange (SCX) (14, 15), are used to further resolve sample complexity.

MS-based proteomic analysis commonly involves the following steps: 1) sample preparation, 2) LC-MS/MS analysis, and 3) computational analysis of the MS data for protein quantification and identification (16). When different combinations of methods are used for each of these three steps, the results obtained can provide a variety of information on various protein states (Table 1). First, the abundances of proteins in complex samples are normally measured using isotope-labeling methods, such as, iCAT (17), SILAC (18), or iTRAQ (19). Peptide abundances obtained using the iCAT and SILAC approaches can be estimated using elution curve areas and several computational tools like XPRESS (20) and ASAPRatio (21). In addition, peptide abundances determined using the iTRAQ approach can be estimated using Libra (12) as peak intensities of reporter ions with m/z ratios of 114, 115, 116, and 117 in MS/MS scans. In addition to these labeling methods, two label-free quantification methods, i.e., alignment-based (22) and spectral count-based (23) methods, have been widely employed using computational tools, such as, Specarray (24) and...
ProtQuant (25). Second, extents of the PTMs of proteins can be measured using subproteome capture methods, such as, the phosphoproteome and ubiquitinated proteome capture methods (see Table 1 for details). The performances of these capture methods have been summarized in several reviews (26-28). These methods can also be combined with the isotope-labeling approaches described above to quantitatively measure the extents of PTMs. Third, interaction information (i.e., network edges, such as, PPIs and CPs) are normally measured using the following methods: i) immuno-precipitation (IP)-based methods (29), ii) Tandem Affinity Purification (TAP)-tagging approaches (30), and iii) chemically conjugated bead-based methods (31).

Integrative proteomics data analysis

LC-MS/MS analysis produces huge amounts of data (Fig. 2A and 2B), and requires the support of a data analysis system that includes the following: 1) data conversion, visualization, storage, and exchange facilities, 2) basic analytical functions, e.g., protein quantification (not shown) and identification (Fig. 2C and 2D), and 3) systems biology functional analyses systems. Fig. 2 shows the Integrative Proteomics Data Analysis Pipeline (IPDAP) that our group has developed for systems biology research using proteomic data. IPDAP is built on two platforms, i.e., a Computational Proteomics Laboratory Database (CPAS) (32) and Systems Biology Experiment Analysis Management System (SBEAMS) (33). In addition, it also includes a Trans-Proteomic Pipeline (TPP) (34), a gaggled integrative systems biology system (Fig. 2E), and other tools (e.g., DAVID (35) and SpectrumLook).

When a new raw MS dataset is generated by LC-MS/MS, IPDAP first stores the raw data into CPAS, and then converts this raw data into mzXML, a standard data format, as indicated by the arrow between the "Raw LC-MS/MS data" box and CPAS in Fig. 2H. Spectra can be viewed using SpectrumLook. IPDAP then performs the following basic analyses:

1. A database search is first done using either X! Tandem (36) or SEQUEST (37) for protein identification purposes (Fig. 2C). These tools generate theoretical MS/MS (MS2) spectra for the peptides theoretically tryptic-digested using all protein sequences for a given species in the database. By comparing MS2 spectra (Fig. 2B) with all theoretical spectra, these tools then identify peptides that best match for each spectrum by selecting a theoretical tryptic peptide with the highest similarity measure (e.g., highest cross-correlation Xcorr score in SEQUEST; Fig. 2C).

2. PeptideProphet (38) in the TPP then identifies the best match by calculating the probability that the best match is correct. PeptideProphet performs this by: 1) Generating a meta-measure (e.g., an f-value in SEQUEST) by combining all measures (e.g., Xcorr, dCn, and Sp in Fig. 2C) for each best match; 2) Fitting two distributions of true positives and negatives (Correct and Incorrect) to the distribution of meta-measures of all best-fitting peptides (Fig. 2D); and 3) Estimating the probability that the best match is correct using a Bayesian method and the two estimated distributions. A subset of best matches is selected using a probability cutoff (e.g., 0.95). ProteinProphet (39) then estimates the probability of the protein being correct by combining the probabilities assigned to different detected peptides belonging to the same protein. However, these database search tools are not well-suited for the identification of various types of PTM peptides, and thus, other tools like MODi (40), which have been specially designed for the identification of post-translationally modified proteins (PTM peptides), are used.

3. IPDAP finally quantifies proteins using the TPP tools (XPRESS, ASAPRatio estimator, and Libra, which are used in conjunction with isotope-labeling methods), as described in the previous section. The results of protein identification and quantification are stored in CPAS, as indicated by the arrow between the "Protein Identification (TPP)" box and CPAS in Fig. 2.

All of the basic analysis tools described above in IPDAP provide a list of proteins and of their abundances and/or the extents of PTMs. For functional analysis, proteins should be mapped into operational biological networks, i.e., we need to determine how these proteins interact with each other, what pathways and functional groups are enriched in these proteins, and how these pathways change over time in order to understand what subsets of pathways play key roles during the early phases of biological events (e.g., disease initiation and progression). The mapping of such lists of proteins into networks requires a variety of systems biology software tools, as shown in the "Systems Biology Tools" box in Fig. 2. These tools are connected to our IPDAP system via Gaggle, an interface in the CPAS platform that facilitates data exchange between many software tools (Fig. 2E). A selection of software standalone and web-based tools are separately installed or incorporated for direct usage in IPDAP. These include: 1) interaction database querying systems (BIND (41) and HPRD (42)); 2) tools for basic statistical analyses (TIGR MeV) (43), such as, clustering and Principal Component Analysis tools (PCA tools; Fig. 2F); 3) network modeling and analysis tools (STRING (44) and Cytoscape; Fig. 2G); and a pathway analysis tool (KEGG) (45).

In addition to these tools, other commercial tools, such as, GeneGo and Ingenuity Pathway Analysis (IPA) (46), that can also be incorporated into IPDAP.

Network modeling using proteomic data

Network modeling is a key step for processing proteomic data in systems biology, because a network model provides: 1) a means of understanding how detected proteins are associated with underlying network operations, and 2) a platform into which other useful information, such as, abundances and PTM extents can be integrated. Thus, IPDAP provides a general solution for network modeling and for visualizing the pipeline, as shown in Fig. 3. Network models can be reconstructed in IPDAP as follows: 1) a list of detected proteins is transferred to-
gether with their abundances and/or PTM extents from CPAS to Gaggle Boss (47) as a data routing hub, which transfers data between data software packages, as indicated by the arrow numbered ‘1’ in Fig. 3A. Gaggle Boss can then pass the transferred information into one of a number of Gaggle Geese. Information is normally first passed to Data Matrix Viewer (DMV), which allows information to be viewed and processed, as indicated by the arrow numbered by 2 in Fig. 3B. Proteins and their abundances can then be ‘broadcast’ via Gaggle Boss to TIGR MeV, in which statistical analyses can be performed, e.g., for the identification of DEPs, clustering, PCA tools, and the generation of heatmaps, as indicated by arrow ‘3’ in Fig. 3C and 3D. Identified DEPs and their abundances (and/or PTM extents) can finally be sent to Cytoscape for network modeling. During this transmission, a sif file is generated using the first neighbors of DEPs and queried PPIs, and network attribute files are generated using protein abundances and PTM extents. Cytoscape Java web start then imports these files for network visualization, as indicated by the arrows numbered ‘4’ and ‘5’ in Fig. 3C and 3D. The red and green colors represent increases and decreases in the abundances (or PTM extents) of the corresponding network nodes, respectively. Both coloring schemes using node colors (abundances) and node boundary colors (PTM extents) and can also be employed to visualize changes in these two types of information. When time-course data are available, multiple node attribute files can be generated corresponding to the dynamic transitions of network models.

**Network analysis for the generation of testable hypotheses using Cytoscape Plugins**

Having developed a network model, as described above, networks can be analyzed using topologic information (e.g., connectivity), and also using combined information of various types, such as, GO annotations and known pathways. Fig. 4 shows the devised network analysis platform with several Cytoscape Plugins and computational tools (Biological Networks Gene Ontology tool (BiNGO) (48), Database for Annotation, Visualization and Integrated Discovery (DAVID), MCODE (49), ActiveModules (50), and Michigan Molecular Interaction (MiMI) (51)). Analysis outputs can be summarized as: 1) the identification of network modules enriched by GO terms and pathways; 2) the exploration of network structures using clustering and active network portions (i.e., subnetworks) in terms of abundances and PTM extents; and 3) network reduction and expansion. We include below a brief description of each of these output forms.

1. The identification of GO terms enriched by network nodal proteins. GO Biologic Process (GOBP) trees with nodes corresponding to GOBP terms can be generated using the

---

**Fig. 3.** Network modeling pipeline in IPDAP. (A) Gaggle Boss GUI. Abundances and PTM extents computed using ASAPRatio, XPRESS, or Libra and a list of proteins exported to Gaggle from CPAS (arrow 1). (B) DMV GUI showing data import from Gaggle Boss (arrow 2). The selected sub-matrix is then broadcast to TIGR MeV via Gaggle (arrow 3). (C) Heatmap of DEPs generated by TIGR MeV. DEPs identified by TIGR MeV are then forwarded to Cytoscape via Gaggle (arrows 4 and 5). (D) Network model visualized by Cytoscape. Protein abundances can also be represented by node colors. A red and green indicate increased and decreased abundance, respectively, versus controls. Refer to the text for a detailed description of the tools used.
BiNGO plugin (Fig. 4A). A dark orange color indicates that a GOBP term is overrepresented among network nodal proteins (refer to the GOBP node ‘Positive regulation of apoptosis’ in the box in Fig. 4A). Node size is proportional to the number of proteins belonging to the GO term. If GOBP trees for two conditions are compared (e.g., the early and late stages of disease progression), GOBP terms that differ in terms of their degrees of enrichment can be identified, as can sets of network nodal proteins belonging to such GOBP terms. Accordingly, a subnetwork can be reconstructed that describes at the molecular level biological processes which differ during the early and late stages of disease progression, which facilitates the identification of early diagnostic markers as key nodes in subnetworks. Pathway enrichment analysis can also be done for network nodal proteins using DAVID (Fig. 4B), and the results obtained (e.g., enriched pathways; refer to the box in Fig. 4B) can be interpreted in the same manner as those of GO enrichment analysis (e.g., the identification of enriched pathways that differ in terms of degree of enrichment under two conditions).

Exploration of network structures and active subnetworks. The MCODE plugin can be used to generate network clusters (refer to the box in Fig. 4C) within which proteins are densely connected, whereas proteins across different network clusters loosely interact (refer to Fig. 4C figure caption for a detailed description). Both the core network modules and their dynamic relationships can be identified by integrating time-dependent abundance and PTM information. By analyzing such dynamic relationships within core modules, the mechanisms of given biological issues (e.g., disease progression) can be identified. In addition, active networks can be identified among network modules (see the subnetwork in the box in Fig. 4E) using jActiveModules, which select networks with high collec-

![Network Analysis using Cytoscape and its plugins.](image)

**Fig. 4.** Network Analysis using Cytoscape and its plugins. Given an original network, various types of analysis can be performed using Cytoscape plugins. (A) GO enrichment analysis using BiNGO. The box shows an enriched GOBP term, named “Positive regulation of apoptosis”, in the GOBP tree. (B) Pathway enrichment analysis using DAVID. The box shows a part of the enriched pathway, origin recognition complex. (C) Network clustering using MCODE. One of the network clusters is shown in the box. (D) Network expansion using MiMI. The box shows a subnetwork expanded from a particular portion of the network (the circle indicates a disconnected subnetwork) by adding PPIs (added nodes are diamond shaped). (E) Active modules identified using jActiveModules. An exemplary active module is shown in the box. Refer to the text for detailed descriptions of the tools.
tive abundances (i.e., the sum of normalized abundances) or PTM extents (Fig. 4E). When this information is collated, a reliable mechanistic hypothesis may be predicted for a given biological problem.

2. Network reduction or expansion. The initial networks can be reduced by selecting only active or network modules enriched by GO terms and pathways identified using the plugins described above. In addition, particular network portions can be expanded using the MiMI plugin by retrieving PPIs from the MiMI database and incorporating into the subnetwork concerned (Fig. 4D) or disconnected (refer to Fig. 4D for an example of a disconnected network portion). The diamond shaped nodes in Fig. 4D represent added nodes that connect the disconnected portion to the main network.

Discussion

This review presents a systems biology framework for studying biological systems at system levels using proteomic data. This framework has been applied to cellular and tissue systems during proteomic systems studies, and was found to improve our understanding of the underlying mechanisms related to various medical problems. However, despite the promise shown by the devised IPDAP system, several technical issues have yet to be resolved, i.e., 1) small system coverage due to the small sizes of detected proteomes, 2) incomplete interactome sets (e.g., PPIs and CPIs), and 3) high false positive error rates in terms of determining interactions. In particular, more software tools must be developed that more effectively perform data integration and network analyses. Despite these issues, if certain specific combinations of tools and application sequences are chosen to address given problems, we expect to gain novel insight of the mechanisms underlying key biological events in complex disease systems, which we believe will contribute to the identification of early diagnostic markers and better drug targets with fewer side effects.

REFERENCES


From proteomics toward systems biology
Sangchul Rho, et al.

33. http://sboams.org


