Chapter 19

Integrating Omics Data for Signaling Pathways, Interactome Reconstruction, and Functional Analysis

Paolo Tieri, Alberto de la Fuente, Alberto Termanini, and Claudio Franceschi

Abstract

Omics data and computational approaches are today providing a key to disentangle the complex architecture of living systems. The integration and analysis of data of different nature allows to extract meaningful representations of signaling pathways and protein interactions networks, helpful in achieving an increased understanding of such intricate biochemical processes. We here describe a general workflow and relative hurdles in integrating online Omics data and analyzing reconstructed representations by using the available computational platforms.

Key words: Pathway, Interactome, Signaling, Network, Protein interactions, Data integration, Data retrieval, Systems biology, Bioinformatics

1. Introduction

Network abstractions and network analysis are today common in science. This approach has been applied for the representation of complex systems, and has achieved a certain success, from social studies (1) to engineering (2) and biology (3–10).

Despite its intrinsically limited perspective, such conceptualization enables complex biological systems to be considered as a whole and open for mathematical analysis, aiming to the discovery of salient systemic features and providing an accurate and analytic view at the glance of entities, relations, and functions that characterize them. This approach also allows to highlight how the qualities and behavior of single elements influence the network topology and dynamics, how network structure impinges upon

Bernd Mayer (ed.), *Bioinformatics for Omics Data: Methods and Protocols*, Methods in Molecular Biology, vol. 719,
DOI 10.1007/978-1-61779-027-0_19,

processes spreading over the network, or the effect of perturbations on network performance (11, 12). In this regard, the network abstraction of biochemical signaling pathways can represent a useful functional view that can complement analyses and approaches from molecular biology and the various Omics.

Biochemical pathways are usually referred to as intracellular processes whose scale can in some way be placed between small events, such as protein complexes formation or enzyme catalysis, and cell-wide or larger events, such as cell death or inflammation. These processes can be divided into separate steps, which seldom follow a linear and unambiguous succession. It is not yet simple to define a pathway in terms of its components, steps, dynamics and function, given its manifold, hazy, and intricate nature. Actually, pathways and signaling cascades are not isolated entities. A signaling pathway can be triggered by different extra- or intracellular events, may cover different parallel paths and branches, may intersect, be competitive or cooperative or interdependent with other events, each step may have diverging functions, and so on. Pathways, in conclusion, are processes characterized by high complexity (13–15). Abstractions and models of biological networks and pathways discussed here are mainly protein interaction networks (PINs) and protein-signaling networks (PSNs). PINs represent protein–protein binding events on a proteome-wide scale. Nodes and undirected edges represent proteins and binding events among them. In PSNs, nodes and directed edges represent phosphoproteins and phosphorylation reactions. The two models can be combined and enriched with additional layers, such as transcriptional regulatory networks, among others.

Omics data and computational approaches are today providing a key to disentangle the complexity of objects like signaling pathways, assisted by dedicated online databases and specific software tools. Through such methodology, it is possible to integrate data of different nature to extract meaningful representations and useful information, finally leading to an increased understanding of the biochemical process under examination. Nevertheless, the workflow for the integrated reconstruction and analysis of signaling pathways, interactomes, and biological networks is hampered by difficulties of diverse nature, such as lack of data, annotation differences or multiple interpretations, data integration problems and other difficulties (16–18). Materials and workflow described here want to demonstrate a general approach for gathering information of interest from some of the existing pathway and protein interactions databases, for integrating and analyzing data and reconstructed representations by using the available tools, and to understand which kind of knowledge can be extracted from the combination of existing information (Fig. 1). We shortly describe the characteristics of some of the many pathway and protein

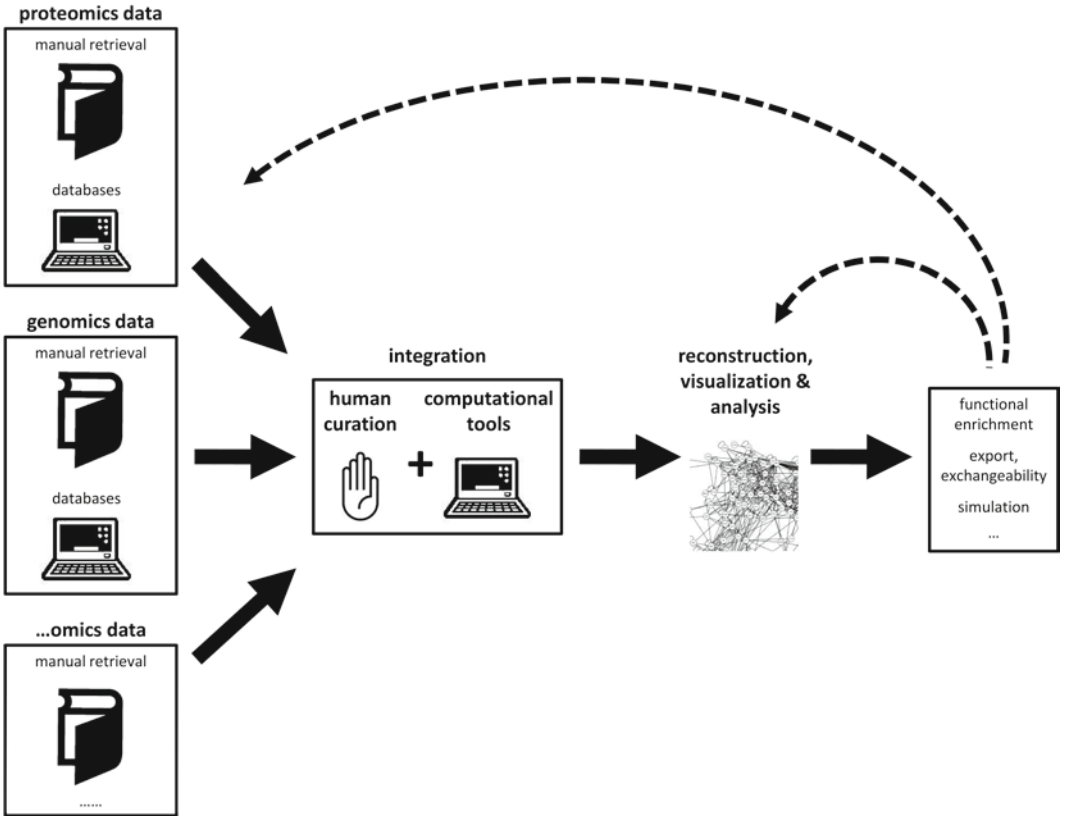

Fig. 1. Schematic representation of the analysis workflow. From manual and automated data retrieval, through human curation and software platforms, data are integrated to reconstruct coherent objects able to undergo mathematical analysis. Results can feed back in the pipeline for further enrichment, analysis, simulations, or improvement of existing models and representations.

interactions online resources and databases, and how the Cytoscape software platform and other analysis tools can be applied to reconstruct and analyze some exemplar pathways and interactomes.

2. Materials

2.1. Overview of Databases and Online Data Sources

Signaling cascade and pathway information is more and more systematically collected and organized into publicly available databases. Such kind of resources lay the foundations for the systems level approach, allowing a workflow consisting in the reconstructive process of the pathway/interactome network, that generally consists in the manual or automated retrieval of pathway data, their integration, merging, comparison and enrichment with other forms of data, and then the analytical process (simulation, mathematical modeling, statistical analysis). Iterative cycles of

such procedures, modeling, and prediction, combined with experimental validation, can result in the improvement of the knowledge of cell signaling and responses.

Online dedicated databases usually store cell signaling data in exchangeable formats (often BioPAX – Biological Pathway Exchange-, or SBML – Systems Biology Markup Language; see Note 1) accessible by diverse software platforms and tools, allowing for their retrieval, visualization, and analysis. The following list should by no means be considered as exhaustive; links and URLs can be found in the Notes section.

The Pathguide (the Pathway Resource List, see Note 2) (19) is a useful resource serving as starting point for biological pathway analysis, since it is a content aggregator for integrated biological information systems. Pathguide is a meta-database that provides an overview of current pathway and other systems biology-oriented databases. Pathguide currently lists and provides details and links to more than 300 web-accessible biological pathway and network databases. These include databases on metabolic pathways, signaling pathways, transcription factor targets, gene regulatory networks, genetic interactions, protein–compound interactions, and protein–protein interactions. The listed databases are curated and maintained by diverse scientific groups in different worldwide locations, and the information represented is derived either from the scientific literature or from systematic, high-throughput experiments.

Reactome ((20), see Note 2) is a pathway database covering a wide set of biological processes, organized in a hierarchical manner: Lower levels for smaller reactions, higher levels for pathways and extended processes. Data are extracted from literature and biomedical experiments, are human-curated and are represented as chains of chemical reactions (including transcription, catalysis, binding). Data can be physical entities (DNA, RNA, protein complexes, phosphorylated and unphosphorylated proteins, small molecules…), or events (reaction-like event for smaller reactions, or pathway-like event clustering a set of reaction-like events). The tool allows remote search and browsing, but also to download data in the most common formats or in graphical representation. The Web site also provides some useful statistical and graphical tools and can be accessed through a Simple Object Access Protocol (SOAP, http://www.w3.org/TR/soap) Web service for automated data queries.

KEGG ((21, 22), see Note 2) consists of a number of interlinked databases devoted to several domains in the cell and beyond (genes, genomes, proteins, chemical compounds, pathways, diseases, drugs, ontologies). The pathways section covers many organisms, including human. Data are categorized into the different processes (metabolic, genetic information, signaling, etc.) and are coded in a special XML format (KGML), but also in BioPAX and SBML through the use of additionally available coding tools.

The Nature Pathway Interaction Database (PID) ((23), see Note 2) is hierarchically organized in a way similar to Reactome and hosts pathway data (available in BioPAX or XML) obtained from peer reviewed literature or imported from other databases, such as Reactome or BioCarta (a supplier of reagents and assays for biopharmaceutical and academic research; see Note 2). DNA and RNA are not part of the PID pathways but active/inactive, phosphorylated/unphosphorylated states are annotated. The pathways can be browsed starting from UniProt, Entrez Gene (see Note 2), or other identifiers, and query as well as statistical tools are provided.

Pathway Commons is based on already existing databases, such as Reactome, PID, and other protein interactions databases, and provides an integrated access point and a compilation of such databases, thus conserving their structure and data hierarchies. However, this kind of integration is not only a simple task and may result in overlapping, but also discordant and/or redundant information. A useful feature is the complete accessibility through the dedicated Pathway Commons plugin from the Cytoscape platform (see later in the chapter).

WikiPathways ((24), see Note 2) is an open source and collaborative platform for biological pathway information, storage, and curation, in the wake of the Wikipedia style. Data are categorized by species and processes (e.g., metabolic process, molecular function, etc.) and are coded in the GenMAPP (an application designed to visualize gene expression and other genomic data on maps representing biological pathways and groupings of genes, *see* Note 2) Pathway Markup Language (GPML), being compatible with applications, such as PathVisio (a visualization tool, see Note 2), Cytoscape, and GenMAPP.

Agile Protein Interaction DataAnalyzer (APID) (25) is an interactive Web-based platform devoted to the exploration and analysis of diverse information about protein interactions, integrated and unified in a common and comparative environment. APID provides an open access frame where all experimentally validated protein–protein interactions (obtained from protein interactions databases, such as BIND, BioGRID, DIP, HPRD, IntAct, and MINT, see Note 2) are unified in a unique Web application that allows the exploration and analysis of networks and interactomes. APID provides some embedded online tools to query and browse data and, most useful, a Cytoscape plugin (APID2NET, (26)) that allows to extract, visualize, and analyze unified interactome data by directly quering APID servers, including all the annotations and attributes associated to the retrieved PPIs.

Transcriptional Regulatory Element Database (TRED) ((27, 28), see Note 2) is a manually curated database of regulatory elements (promoters, transcription factor binding sites, both *cis* and *trans*) with experimental evidence in mammalian genomes.

Currently, it enlists a total of 36 transcription factors families (most of which are involved in cancer), more than 7,000 target genes and around 15,000 target promoters, with the goal to assist detailed functional studies and to help in obtaining a panoramic view of gene regulatory networks in a cancer research perspective.

TRANSPATH (29), together with the more famous TRANSFAC ((30), see Note 2), that stores transcription factors and their DNA binding sites, is a widely used and powerful knowledge base on gene regulatory networks that comprises and integrates information on signal transduction and tools for visualization and analysis. It allows obtaining complete signaling pathways from ligand to target genes and their products. Its access requires a license purchase, even if a version dating back to years ago can be accessed for free.

NetPath ((31), see Note 2) is a curated compendium of human signaling pathways which currently contains annotations for several cancer and immune signaling pathways. Pathway data are available for browsing and download in the most common formats (including the Proteomics Standards Initiative-Molecular Interaction – PSI-MI format), and listing of up- and downregulated genes for each pathway is provided based on experimental data and literature.

Notwithstanding the quantity and quality of the publicly available resources, information automatically extracted from pathway databases is usually not yet exhaustive. Given the often complementary nature of data in different databases, they should be retrieved, integrated, and combined, and we feel the quality of the result strongly relies upon a sharp manual curation effort (16–18). The integration process itself, however, can present several problems, not least those of interchangeability of the different formats and data models, but also in terms of reaction annotation, or of significant differences in other key biological factors, such as cellular state and type (16). Thus, the process of literature extraction of data (also possibly aided with text mining techniques) combined with information from databases under expert supervision and curation probably remains a good choice in order to get an accurate pathway reconstruction. A complete and deep curation process can last months and employ many experts, and yet yield controversial results. Conversely, manual integration of data extracted from online pathway resources – under expert review – can be decently performed in days, allowing to create a sufficiently accurate (also depending on the scope) representation of a given pathway, or part of it, ready for further functional enrichment and analysis.

2.2. Computational Analysis Software

2.2.1. Main Platforms and Tools

Since the purpose of the interactome or pathway reconstruction process is to have an "object" which can be further elaborated, enriched, and analyzed step by step, we need to access and store data in local machines, in contrast to browsing them online. As described before, most of the databases allow downloading the

relevant data in diverse formats (BioPAX, SBML, PSI-MI, among others). At this point, the choice of one or more tools for network editing and analysis is up to the user. Some of these are directly embedded or available inside the different databases, such as Reactome, WikiPathways, BioCarta, and GenMAPP. Others are commercial suites, such as Ingenuity or Pathway Studio, with special visualization features (see Note 2).

Among the open source applications, Cytoscape ((32), see Note 2) is a very powerful software platform, available for all the major operating systems, designed for biological research, but versatile enough to be used in many other fields where network editing, visualization, and analysis are key features. The core tool has been developed to visualize molecular interaction networks and biological pathways, and to integrate these networks with annotations, gene expression profiles, and other state data. Many more additional features, such as advanced network and molecular profiling analyses, new layouts, additional file format support, scripting, and connection with databases, are available as plugins. Cytoscape supports many different standard network and annotation file formats, including Simple Interaction Format (SIF), BioPAX, PSI-MI, SBML, tab-delimited text files, and MS Excel.

BiologicalNetworks ((33), see Note 2) is an integrated research environment for biological sciences that allows querying and integrating molecular interaction networks, metabolic and signaling pathways with a large number of biological features related to transcriptional regulation, microarray and proteomics experiments, 3D structures ontologies, taxonomies, and other types of data. The tool is based on a database currently integrating over 100 curated and publicly contributed data sources for thousands of eukaryotic, prokaryotic, and viral genomes.

CellDesigner ((34), see Note 2) is a structured diagram editor for drawing gene-regulatory and biochemical networks. Networks are drawn based on a process diagram, with a dedicated graphical notation system, and are stored using the SBML format. Networks can be linked with simulation and other analysis packages through a wider software platform named Systems Biology Workbench (SBW).

We in the Methods section focus on a workflow mainly based on the Cytoscape platform given its free availability, diffusion in biology research, upgradeability, and versatility.

2.2.2. Other Specific Analysis Tools and Plugins

Powerful standalone packages specific for network analysis are freely available. Pajek (35) ("spider" in Slovene, the nationality of the developers, see Note 2), for instance, is able to visualize and analyze networks of millions of nodes. Specific add-on modules can be used inside the well-known R statistical package (http://www.r-project.org).

Other packages have direct Web-based functionality: GraphWeb (36) is a public Web server for graph-based analysis

that has been designed for extensive analyses of directed and undirected, weighted and unweighted, heterogeneous networks of genes, proteins and microarray probesets for many eukaryotic genomes, and is able to integrate multiple, diverse datasets for constructing extended networks.

Among the many available Cytoscape plugins (for an exhaustive list and references see the Cytoscape.org Web site), NetworkAnalyzer (37) requires no expert knowledge in graph theory. The tool provides functionality to compute and display charts for a quite complete set of topological parameters for undirected and directed networks, which includes the number of nodes, edges, and connected components, the network diameter, radius, density, centralization, heterogeneity, clustering coefficient, and the characteristic path length.

ClusterMaker (Cytoscape plugin) unifies different clustering techniques and displays into a single interface. It uses specific algorithms for clustering expression or genetic data, and similarity networks to look for protein families and putative functional similarities.

The Hub Objects Analyzer (Hubba) (38) is both a Web-based service and a Cytoscape plugin for exploring networks for the discovery of hubs in an interactome network generated from specific small- or large-scale experimental methods.

3. Methods

3.1. General Retrieval and Reconstruction Procedures

3.1.1. Data Retrieval

The process of manual literature mining for data extraction is labor-intensive and time consuming but typically gives back high-quality data and models. It is evident that, given the broadness and importance of this topic, it cannot be exhaustively treated here and we refer to Jensen and colleagues (39) for a comprehensive review on the field of manual and machine-aided extraction of biomedical facts from scientific literature.

In the first step of retrieving the pathway data of interest through Cytoscape, the user can utilize one of the many existing plugins, each one designed to query and retrieve data from many different databases. It is evidently advisable that the user has in a first step browsed the candidate databases to understand which type, model, and format is used for data representation.

Among the many Cytoscape plugins, BioNetBuilder (40) can be used to build networks for many different species, including most common model organisms and human, retrieving data from currently supported databases that include DIP, BIND, KEGG, HPRD, BioGrid, among others. The interface offers different options to specify a set of initial genes/gene products for which to find molecular interactions (including loading a text file, finding

genes with specified Gene Ontology annotations, and finding genes whose common name match a given string). Biological networks for whole organisms can also be created and displayed.

Another very useful plugin is the aforementioned APID2NET, linked to the APID database. This tool allows to specify a list of proteins for retrieving the network of their interactions at the desired connection level (level 0 considers only the interactions among the listed proteins, level 1 considers all their neighbors in APID, level 2 considers also the neighbors of the neighbors, etc.) and validated by a number of different experimental methods to be selected. The system also displays additional information on node, edge, and network attributes.

The user can also start a Cytoscape session with the embedded "import network from web service" function to connect directly to the Pathway Commons or WikiPathways servers and obtain the data. It is also possible to retrieve the data from each single database simply by downloading a formatted file and then upload and open it in the Cytoscape client for visualizing the network.

It is not always possible to retrieve data following a plugin-automated or semi-automated process as described above. For some databases, not specifically designed for systems biology but containing useful and well-arranged information, as for instance the TRED, no workflow is provided. Here, it may be necessary to formulate a query, to extract the data with copy/paste operations in text format, and to perform further adaptation to import and incorporate them into a network in a very manual fashion.

3.1.2. Data Merging and Combination

As said, combination of data from different pathway databases is highly desirable. The user can, for instance, download the same pathway data as provided by two or more databases and try to combine them in order to make it as complete as possible. For this purpose, again, suitable Cytoscape plugins (e.g., AdvancedNetworkMerge) or embedded functions can be used. This is a critical point, since frequently molecular and reaction data are encoded and modeled in different manner according to the originating database so that the network resulting from the merging of such two or more networks can disappointingly result as a simple sum of the originating objects, or anyway an inconsistent network, even without any partial overlap, or any other shared information or link. There is no trivial solution to this kind of issues, since from database to database there are no uniquely defined identifiers for each of the entities that compose the pathways or the networks. Accurate filtering and expert curation performed before the merging process could purge the data from undesired or redundant information. This also usually makes it quite easy to build improved versions of the networks based on additional and different types of data.

3.1.3. Functional Enrichment

Obtained networks can be functionally enriched, i.e., can be integrated and superimposed with data of different type, such as gene expression data or Gene Ontology (GO) categories to verify if statistically overrepresented features are linked to topological characteristics. Some plugins are available for Cytoscape and many others are accessible on the internet. Among them, we just mention BiNGO (41) and ClueGO (42) as plugins enabling to determine which Gene Ontology categories are overrepresented in sets of genes, (in the present context corresponding to subgraphs of a given biological network), allowing to map the predominant functional themes of a given gene set on the GO hierarchy as a graph, and to perform cluster analysis and comparison of clusters.

3.2. Network Analysis

Once that the user has performed the reconstruction steps and considers the "object" pathway or interactome in some way complete and stable (for the subjective purpose of the study to be carried out), it is time to proceed with the subsequent network analysis. All cited computational platforms are precisely designed to perform such analyses that can be easily implemented through embedded or add-on features.

3.2.1. Topological Measures

The goal of topological analysis of protein networks is to discern organizational "design" principles, relate those to dynamical properties, and establish connections to biological functions. The detection of interesting topological properties occurs by comparing the network under study with a "null model"; that is, a set of networks that reflect what is expected by random chance. If a network under study possesses certain characteristics different from what is expected by chance alone, then these might be related to the specific function of the network: they could have been selected by evolution for their advantageous properties.

Topological measures have demonstrated their usefulness in uncovering the organizing principles that rule the development and the evolution of networks of different nature (8). Several observations led to the conclusion that the classical degree distribution, and the well-investigated scale-free characteristic of nodes in PINs, for instance, correlates with biological meaningful features, such as importance, lethality, robustness, and dynamics of perturbations. Hierarchical topology, subgraphs, modular structures, clusters are, among others, strongly characterizing features of networks that a focused analysis can reveal (10). In some fields, such as cancer research, extensive and deep meta-analyses have shown how some specific measures, such as betweenness and stress centrality, among others, are particularly relevant in characterizing pathological states and malignant tissues (43).

3.2.2. Dynamical Models

Owing to the intricacy of signal transduction, computational analysis is necessary to obtain the understanding of dynamical properties of PSNs. Even for very small, relatively simple PSNs,

it has been shown that a wide range of complex dynamical properties could be attained (13, 44–46), and parallels were drawn between signaling circuits and man-made control systems for explaining important biological properties, such as amplification, robustness, homeostasis, and adaptation, particularly highlighting the importance of feedbacks in PSNs (45, 47–51). Several larger mathematical models based on Ordinary Differential Equations have been formulated for signal PSNs, and their parameters were optimized in order to fit experimental observations (52–55). Although studies with such models provide detailed insights into the dynamics and function of signaling pathways, formulating such models is a difficult problem that requires a huge amount of specific and quantitative experimental data, which are not expected to be available on proteome-wide scale in the near future. Dynamical models of proteome-wide PSNs, although lacking precise quantitative information of the kinetic dependencies, can still be used to discover principles of global dynamical organization. For example, a qualitative approach to the dynamic modeling of PSNs is the use of Boolean logic, in which each protein is "off" or "on" at a given time-step depending on the states of its inputs. Recently, it was shown that a PSN formalized with Boolean logic can classify sets of inputs into distinct output patterns – an ability that arises through the complex wiring pattern among the proteins in the PSNs (56). This ability is an emergent dynamical property determined by the structure of the PSN, as the authors showed that randomizing the network results in loss of this ability. Interesting metaphors have been drawn between PSNs and computational networks (14). Back in 1990, Dennis Bray highlighted the similarity between PSNs and "artificial neural networks" (57). Rather than signal transduction as just a mechanism to transmit information from the cell surface to the nucleus and other functions, this analogy suggests a process of turning complex input signals (environments) into complex output signals (biological responses). Similar to artificial neural networks, where the parameters are adjusted through mathematical optimization to obtain required input–output relationships, evolution has tweaked the parameters in PSNs to obtain the ability to generate appropriate responses to the wide variety of complex environmental signals that organisms are subjected to (57).

It has become clear that the proteome forms a complex system with many emergent properties yet to be discovered and understood (10, 14). Topological and dynamical studies of PSNs that take explicitly the INPUT→CENTRAL NETWORK→OUTPUT structure (56, 58–61) into account most certainly yields many insights into the functional organization these intricate protein networks (Fig. 2).

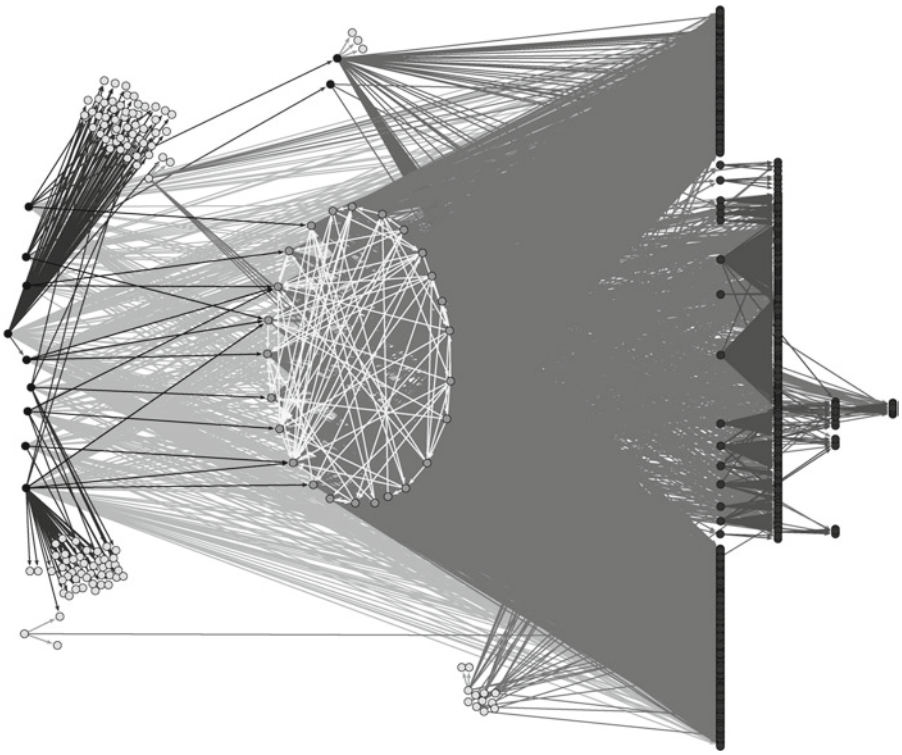

Fig. 2. The "bow-tie" layout of a Human Protein Signaling Network clearly shows the network's main information flow from the input nodes to the central core, which processes and passes the information to the output nodes, in turn establishing the physiological responses (59). This is one example among the many relevant results that can be obtained by network analysis.

3.3. Practical Applications

3.3.1. Examples

We present here the main steps of the workflow followed for the transcription factor Nuclear Factor-κB (NF-κB) interactome reconstruction and analysis (62). NF-κB is a central transcription factor, involved in inflammation as well as in many other normal and pathophysiological processes. Given the intricacy of the signaling system and the number of genes directly regulated, it is interesting to study the main characteristics of its interactome.

1. We start from manual literature mining and review: such approach, that in this particular case lasted about 2 months, guarantees a quite complete list of proteins that take part in the signaling cascade with different roles and importance. This basic list can be expanded, enriched, and refined step by step confronting and complementing preliminary results with data browsed and downloaded (manually or utilizing tools as Cytoscape) from several pathway and PIN databases. The result at the end of this manual process is a "core list" consisting of 140 proteins.

2. Protein interactions data are added to build the first version of the "core interactome." The main tool used in this step is the APID database, automatically accessed through the dedicated plugin in Cytoscape. The result is a network consisting of 140 nodes and 829 nondirectional interactions.
3. By using an automated retrieval tool and databases (APID2NET, BioNetBuilder in Cytoscape, main PIN databases) a "wider interactome" is built, taking into account all the proteins with the evidence of interaction with at least one protein present in the "core interactome." At the end of the process, the whole "wider interactome" consists of more than 3,100 proteins accounting for a total of more than 42,600 interactions.
4. Data elaborated from a manually curated list of NF-κB-downstream genes (63), from the TRED database (manually extracted), and integrated with results from TRANSFAC allow to constitute a relatively comprehensive list of about 400 genes that result to be up- or downregulated via NF-κB. Gene products and relative UniProt identifiers are obtained directly through the ID mapping functions available on the UniProt Web interface, allowing to compile the list of proteins whose expression can be regulated by NF-κB.
5. The whole interactome now consisting of core proteins (those that directly participate in signaling cascades activating NF-κB), wider interactome proteins (their direct interactors), regulated genes and relative expressed proteins, now undergoes functional enrichment and analysis: topological characterization, GO enrichment and clustering are all easily performed thanks to the availability of several standalone analysis tools as Cytoscape as well as Web-based services. Results from the analysis include, among others, a wider, integrated overlook of the NF-κB signaling system and its main topological characteristics, the detection of specific hubs or central proteins, the discovery of feedback loops and cross controls among proteins, and genes that can be candidates for further in-depth studies.

3.3.2. Pitfalls

We take into account here some pitfalls in the procedure shown, as well as some general considerations on the proposed workflow and relative problems encountered.

One of the major concerns in pathway and PSN reconstruction is the lack of clear and comprehensive data about reactions and subsequent directionality. Directional information is still a rare quality. As said, the user unlikely finds that the same pathway is represented in at least similar ways in different databases. This poses the necessity to choose one out of different data models and content, or to engage in the nontrivial effort of integrating

and complementing the various data and data types. The lack of undisputable data about a number of reactions and proteins in the mentioned NF-κB interactome reconstruction and the existence of normal time constraints persuaded us – at least provisionally – to omit the relative dynamical information in our representation. Without directional information, it is impossible to implement dynamical models and simulation, even a simple model based on Boolean dynamics, unless willing to make the assumptions that each edge A-B is bidirectional, i.e., A→B and B→A, which is very unrealistic indeed. Automation of procedures able to integrate different pathways in a coherent and biological meaningful way is a critical point. Currently, there is no practical, coherent, and effective way to integrate data from multiple sources into a single object other than manual intervention. Even if data from single pathways in the different databases are often very close to be precise, comprehensive, and satisfying to serve as a starting base, it is the integration process and subsequent elaboration to hopefully bring valuable information and new knowledge. Actually, in this regard, data models, representations, and annotation are key points in the discussion about these hot topics (64–67).

4. Notes

1. Standards for representation of information about pathways are necessary for integration and analysis of data from various sources.

   BioPAX (Biological Pathway Exchange, http://www.biopax.org) is a biological pathway data exchange format. It enables the integration of diverse pathway resources by defining an open file format specification for the exchange of biological pathway data. Widespread adoption of BioPAX for data exchange facilitates access to uniformity of pathway data from different sources, thereby increasing the efficiency of computational pathway research.

   The Systems Biology Markup Language (SBML, http://www.sbml.org) is a computer-readable format for representing models of biological processes. It is mostly used for dynamical models of metabolism, cell-signaling, and many other topics.

   PSI-MI (http://www.psidev.info) is a standard proposed for improving the annotation and representation of molecular interaction data wherever it is published, i.e., in journal articles, authors' Web sites, or public domain databases, and for improving the general accessibility of molecular interaction data.

The Gene Ontology project (http://www.geneontology.org) is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data.

2. The following is an alphabetical and nonexhaustive list of the resources cited and used in the described reconstruction and analysis process.

- APID Agile Protein Interaction DataAnalyzer – http://bioinfow.dep.usal.es/apid/index.htm
- Ariadne Genomics Pathway Studio – http://www.ariadnegenomics.com/products/pathway-studio
- BIND Biomolecular Interaction Network Database – http://www.bind.ca
- BioCarta Pathways – http://www.biocarta.com/genes/index.asp
- BioGRID The Biological General Repository for Interaction Datasets – http://www.thebiogrid.org
- BiologicalNetworks – http://biologicalnetworks.net
- CellDesigner – http://www.celldesigner.org
- ClusterMaker – http://www.cgl.ucsf.edu/cytoscape/cluster/clusterMaker.html
- Cytoscape – http://www.cytoscape.org
- DIP Database of Interacting Proteins – http://dip.doe-mbi.ucla.edu/dip/Main.cgi
- Entrez Gene – http://www.ncbi.nlm.nih.gov/gene
- GenMAPP Gene Map Annotator and Pathway Profiler – http://www.genmapp.org
- GraphWeb – http://biit.cs.ut.ee/graphweb
- HPRD Human Protein Reference Database – http://www.hprd.org
- HUBBA Hub objects analyzer – http://hub.iis.sinica.edu.tw/Hubba
- Ingenuity Systems – http://www.ingenuity.com
- IntAct – http://www.ebi.ac.uk/intact
- KEGG Kyoto Encyclopedia of Genes and Genomes – http://www.genome.jp/kegg
- MINT the Molecular INTeraction database – http://mint.bio.uniroma2.it/mint
- NCI-Nature Pathway Interaction Database – http://pid.nci.nih.gov

- NetPath – http://www.netpath.org
- NetworkAnalyzer – http://med.bioinf.mpi-inf.mpg.de/netanalyzer
- Pajek – http://vlado.fmf.uni-lj.si/pub/networks/pajek
- Pathguide: the pathway resource list – http://www.pathguide.org
- PathVisio – http://www.pathvisio.org
- Pathway Commons – http://www.pathwaycommons.org
- R Project for Statistical Computing – http://www.r-project.org
- Reactome – http://www.reactome.org
- SBW Systems Biology Workbench – http://sbw.sourceforge.net
- TRANSFAC & TRANSPATH – http://www.gene-regulation.com
- TRED Transcriptional Regulatory Element Database – http://rulai.cshl.edu/cgi-bin/TRED
- UniProt – http://www.uniprot.org
- WikiPathways – http://www.wikipathways.org/index.php/WikiPathways

Acknowledgments

This work has been partially funded by Emilia-Romagna Region BioPharmaNet High Technology Network (http://www.biopharmanet.eu) and by the Regional Authorities of Sardinia.

References

1. Travers J., and Milgram S. (1969) An experimental study of the small world problem. *Sociometry* **32**, 425–43.
2. Alderson D.L., Li L., Willinger W., and Doyle J.C. (2005) Understanding internet topology: principles, models, and validation. *IEEE/ACM Trans Netw* **13**, 1205–18.
3. Watts D.J., and Strogatz S.H. (1998) Collective dynamics of 'small-world' networks. *Nature* **393**, 440–42.
4. Albert R., Jeong H., and Barabasi A.L. (2000) Error and attack tolerance of complex networks. *Nature* **406**, 378–82.
5. Jeong H., Tombor B., Albert R., Oltvai Z.N., and Barabasi A.L. (2000) The large-scale organization of metabolic networks. *Nature* **407**, 651–4.
6. Newman M.E.J. (2000) Models of the small world. *J Stat Phys* **101**, 819–41.
7. Jeong H., Mason S.P., Barabasi A.L., and Oltvai Z.N. (2001) Lethality and centrality in protein networks. *Nature* **411**, 41–2.
8. Barabasi A.L., and Oltvai Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–15.
9. Goh K.I., Cusick M.E., Valle D., Childs B., and Vidal M., et al. (2007) The human disease network. *Proc Natl Acad Sci USA* **104**, 8685–90.

10. Pieroni E., de la Fuente van Bentem S., Mancosu G., Capobianco E., Hirt H., and de la Fuente A. (2008) Protein networking: insights into global functional organization of proteomes. *Proteomics* **8**, 799–816.
11. Boccaletti S., Latora V., Moreno Y., Chavez M., and Hwang D.U. (2006) Complex networks: structure and dynamics. *Phys Rep* **424**, 175–308.
12. Tieri P., Valensin S., Latora V., Castellani G.C., Marchiori M., Remondini D., and Franceschi C. (2005) Quantifying the relevance of different mediators in the human immune cell network. *Bioinformatics* **21**, 1639–43.
13. Bhalla U.S., and Iyengar R. (1999) Emergent properties of networks of biological signaling pathways. *Science* **283**, 381–7.
14. Bhalla U.S. (2003) Understanding complex signaling networks through models and metaphors. *Prog Biophys Mol Biol* **81**, 45–65.
15. Ivakhno S., and Armstrong J.D. (2007) Non-linear dimensionality reduction of signaling networks. *BMC Sys Biol* **1**, 27.
16. Adriaens M.E., Jaillard M., Waagmeester A., Coort S.L.M., Pico A.R., and Evelo C.T.A. (2008) The public road to high-quality curated biological pathways. *Drug Discov Today* **13**, 856–62.
17. Bauer-Mehren A., Furlong L.I., and Sanz F. (2009) Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol Sys Biol* **5**, 290.
18. Gardy J.L., Lynn D.J., Brinkman F.S., and Hancock R.E. (2009) Enabling a systems biology approach to immunology: focus on innate immunity. *Trends Immunol* **30**, 249–62.
19. Bader G.D., Cary M.P., and Sander C. (2006) Pathguide: a pathway resource list. *Nucleic Acids Res* **34**, D504–6.
20. Matthews L., Gopinath G., Gillespie M., Caudy M., Croft D., de Bono B., Garapati P., Hemish J., Hermjakob H., Jassal B., Kanapin A., Lewis S., Mahajan S., May B., Schmidt E., Vastrik I., Wu G., Birney E., Stein L., and D'Eustachio P. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* **37**, D619–22.
21. Kanehisa M., and Goto S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27–30.
22. Kanehisa M., Goto S., Furumichi M., Tanabe M., and Hirakawa M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**, D355–60.
23. Schaefer C.F., Anthony K., Krupa S., Buchoff J., Day M., Hannay T., and Buetow K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res* **37**, D674–9.
24. Pico A.R., Kelder T., van Iersel M.P., Hanspers K., Conklin B.R., and Evelo C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol* **6**, e184.
25. Prieto C., and De Las Rivas J. (2006) APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res* **34**, W298–302.
26. Hernandez-Toro J., Prieto C., and De las Rivas J. (2007) APID2NET: unified interactome graphic analyzer. *Bioinformatics* **23**, 2495–7.
27. Zhao F., Xuan Z., Liu L., and Zhang M.Q. (2005) TRED: a transcriptional regulatory element database and a platform for in silico gene regulation studies. *Nucleic Acids Res* **33**, D103–7.
28. Jiang C., Xuan Z., Zhao F., and Zhang M.Q. (2007) TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res* **35**, D137–40.
29. Choi C., Krull M., Kel A., Kel-Margoulis O., Pistor S., Potapov A., Voss N., and Wingender E. (2004) TRANSPATH-A high quality database focused on signal transduction. *Comp Funct Genom* **2**, 163–8.
30. Matys V., Fricke E., Geffers R., Gössling E., Haubrock M., Hehl R., Hornischer K., Karas D., Kel A.E., Kel-Margoulis O.V., Kloos D.U., Land S., Lewicki-Potapov B., Michael H., Münch R., Reuter I., Rotert S., Saxel H., Scheer M., Thiele S., and Wingender E. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374–8.
31. Keshava Prasad T.S., Goel R., Kandasamy K., Keerthikumar S., Kumar S., Mathivanan S., Telikicherla D., Raju R., Shafreen B., Venugopal A., Balakrishnan L., Marimuthu A., Banerjee S., Somanathan D.S., Sebastian A., Rani S., Ray S., Harrys Kishore C.J., Kanth S., Ahmed M., Kashyap M.K., Mohmood R., Ramachandra Y.L., Krishna V., Rahiman B.A., Mohan S., Ranganathan P., Ramabadran S., Chaerkady R., and Pandey A. (2009) Human Protein Reference Database – 2009 update. *Nucleic Acids Res* **37**, D767–72.
32. Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., and Ideker T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–504.

33. Baitaluk M., Sedova M., Ray A., and Gupta A. (2006) Biological Networks: visualization and analysis tool for systems biology. *Nucleic Acids Res* **34**, W466–71.
34. Funahashi A., Tanimura N., Morohashi M., and Kitano H. (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO* **1**, 159–62.
35. Batagelj V., and Mrvar A. (2003) Pajek – analysis and visualization of large networks. In Jünger M., Mutzel P., (Eds.) *Graph drawing software*. Springer, Berlin. 77–103.
36. Reimand J., Tooming L., Peterson H., Adler P., and Vilo J. (2008) GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res* **36**, W452–9.
37. Assenov Y., Ramírez F., Schelhorn S.E., Lengauer T., and Albrecht M. (2008) Computing topological parameters of biological networks. *Bioinformatics* **24**, 282–4.
38. Lin C.Y., Chin C.H., Wu H.H., Chen S.H., Ho C.W., and Ko M.T. (2008) Hubba: hub objects analyzer – a framework of interactome hubs identification for network biology. *Nucleic Acids Res* **36**, W438–43.
39. Jensen L.J., Saric J., and Bork P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* **7**, 119–29.
40. Avila-Campillo I., Drew K., Lin J., Reiss D.J., and Bonneau R. (2007) BioNetBuilder: automatic integration of biological networks. *Bioinformatics* **23**, 392–3.
41. Maere S., Heymans K., and Kuiper M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448–9.
42. Bindea G., Mlecnik B., Hackl H., Charoentong P., Tosolini M., Kirilovsky A., Fridman W.H., Pagès F., Trajanoski Z., Galon J. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–3.
43. Platzer A., Perco P., Lukas A., and Mayer B. (2007) Characterization of protein-interaction networks in tumors. *BMC Bioinformatics* **8**, 224.
44. Bray D. (1995) Protein molecules as computational elements in living cells. *Nature* **376**, 307–12.
45. Sauro H.M., and Kholodenko B.N. (2004) Quantitative analysis of signaling networks. *Prog Biophys Mol Biol* **86**, 5–43.
46. Tyson J.J., Chen K.C., and Novak B. (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr Opin Cell Biol* **15**, 221–31.
47. Alon U., Surette M.G., Barkai N., and Leibler S. (1999) Robustness in bacterial chemotaxis. *Nature* **397**, 168–71.
48. Ferrell J.E., Jr. (1996) Tripping the switch fantastic: how a protein kinase cascade can convert graded inputs into switch-like outputs. *Trends Biochem Sci* **21**, 460–6.
49. Goldbeter A., and Koshland D.E., Jr. (1981) An amplified sensitivity arising from covalent modification in biological systems. *Proc Natl Acad Sci USA* **78**, 6840–4.
50. Levin M.D., Morton-Firth C.J., Abouhamad W.N., Bourret R.B., and Bray D. (1998) Origins of individual swimming behavior in bacteria. *Biophys J* **74**, 175–81.
51. Yi T.M., Huang Y., Simon M.I., and Doyle J. (2000) Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc Natl Acad Sci USA* **97**, 4649–53.
52. Chen K.C., Calzone L., Csikasz-Nagy A., Cross F.R., Novak B., and Tyson J.J. (2004) Integrative analysis of cell cycle control in budding yeast. *Mol Biol Cell* **15**, 3841–62.
53. Chen K.C., Csikasz-Nagy A., Gyorffy B., Val J., Novak B., and Tyson J.J. (2000) Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol Biol Cell* **11**, 369–91.
54. Kholodenko B.N. (2006) Cell-signalling dynamics in time and space. *Nat Rev Mol Cell Biol* **7**, 165–76.
55. Tyson J.J., Chen K., and Novak B. (2001) Network dynamics and cell physiology. *Nat Rev Mol Cell Biol* **2**, 908–16.
56. Helikar T., Konvalina J., Heidel J., and Rogers J.A. (2008) Emergent decision-making in biological signal transduction networks. *Proc Natl Acad Sci USA* **105**, 1913–8.
57. Bray D. (1990) Intracellular signalling as a parallel distributed process. *J Theor Biol* **143**, 215–31.
58. Cui Q., Yu Z., Purisima E.O., and Wang E. (2006) Principles of microRNA regulation of a human cellular signaling network. *Mol Syst Biol* **2**, 46.
59. de la Fuente A., Fotia G., Maggio F., Mancosu G., and Pieroni E. (2008) Insights into biological information processing: structural and dynamical analysis of a Human Protein Signalling Network. *J Phys A* **41**, 224013.
60. Liu W., Li D., Zhang J., Zhu Y., He F. (2006) SigFlux: a novel network feature to evaluate the importance of proteins in signal transduction networks. *BMC Bioinformatics* **7**, 515.
61. Ma'ayan A., Jenkins S.L., Neves S., Hasseldine A., Grace E., Dubin-Thaler B., Eungdamrong

N.J., Weng G., Ram P.T., Rice J.J., Kershenbaum A., Stolovitzky G.A., Blitzer R.D., and Iyengar R. (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network. *Science* **309**, 1078–83.

62. Tieri P. (2009) Reconstruction and analysis of the NF-kB pathway interactome, communication to NetSci 2010, International Conference on Complex Network Science, 10–14 May 2010, M.I.T. Boston, USA (http://www.netsci2010.net/abstracts/Tieri.htm), and RECOMBSAT 2010, 16-20 November 2010, Columbia Univ., New York, USA (available from Nature Precedings, http://dx.doi.org/10.1038/npre.2010.5266.1.).
63. Gilmore T.D. Rel/NF-kB Transcription Factors website, http://www.nf-kb.org.
64. Ceol A., Chatr-Aryamontri A., Licata L., and Cesareni G. (2008) Linking entries in protein interaction database to structured text: the FEBS Letters experiment. *FEBS Lett* **582**, 1171–7.
65. Leitner F., and Valencia A. (2008) A text-mining perspective on the requirements for electronically annotated abstracts. *FEBS Lett* **582**, 1178–81.
66. Gerstein M., Seringhaus M., and Fields S. (2007) Structured digital abstract makes text mining easy. *Nature* **447**, 142.
67. Termanini A., Tieri P., Franceschi C. (2010) Encoding the states of interacting proteins to facilitate biological pathways reconstruction. *Biology Direct* **13**, 5:52.