Published in final edited form as:

Methods Mol Biol. 2014 ; 1101: . doi:10.1007/978-1-62703-721-1_11.

Integrating *In Silico* Resources to Map a Signaling Network

Hanqing Liu¹, Tim N. Beck^{1,2}, Erica A. Golemis¹, and Ilya G. Serebriiskii^{1,*}

¹Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111

²Drexel University College of Medicine, 2900 Queen Lane, Philadelphia, PA 19129

Abstract

The abundance of publicly available life science databases offer a wealth of information that can support interpretation of experimentally derived data and greatly enhance hypothesis generation. Protein interaction and functional networks are not simply new renditions of existing data: they provide the opportunity to gain insights into the specific physical and functional role a protein plays as part of the biological system. In this chapter, we describe different *in silico* tools that can quickly and conveniently retrieve data from existing data repositories and discuss how the available tools are best utilized for different purposes. While emphasizing protein-protein interaction databases (e.g., BioGrid and IntAct), we also introduce metasearch platforms such as STRING and GeneMANIA, pathway databases (e.g., BioCarta and Pathway Commons), text mining approaches (e.g., PubMed and Chilibot), and resources for drug-protein interactions, genetic information for model organisms and gene expression information based on microarray data mining. Furthermore, we provide a simple step-by-step protocol to building customized protein-protein interaction networks in Cytoscape, a powerful network assembly and visualization program, integrating data retrieved from these various databases. As we illustrate, generation of composite interaction networks enables investigators to extract significantly more information about a given biological system than utilization of a single database or sole reliance on primary literature.

Keywords

in silico; protein-protein interaction (PPI); network; signaling; Cytoscape; metasearch; database; pathway; model organism; text mining

1. Introduction

Until recently, any investigator wishing to identify the proteins physically or functionally interacting with a protein of interest would have had little recourse but to perform a laborious series of biochemical or genetic techniques. However, the astonishing advances that have been made in high-throughput experimentation techniques have produced a vast amount of increasingly diverse data pertaining to the functional states of biological systems. In sync with the experimental advances, many powerful informatics tools, allowing metasearches of databases for signaling pathways and analysis of signaling networks, have emerged, which have allowed the scientific community to begin to take greater advantage of the wealth of data available. These *in silico* tools for organizing, integrating, analyzing and querying biological interactions provide an invaluable resource, with the potential to save laboratory-based investigators time and money; yet many members of the scientific

*Corresponding author: Program in Developmental Therapeutics Fox Chase Cancer Center 333 Cottman Ave. Philadelphia, PA 19111 (215) 728-3885 Ilya.Serebriiskii@fccc.edu.

community are not fully aware of current capabilities. This chapter provides a step-by-step illustration of how to navigate different open-access resources and how to develop a protein-targeted network that can be used to generate and test hypotheses (a simple to follow generic protocol for in-depth analysis is provided in section 3. Methods; Figure 1 provides a broader overview of network building).

In the next section we will outline how an investigator may decide which available resources and tools are the most appropriate choices for a variety of different project goals. For example, some investigators may have identified specific proteins in a mid-throughput experiment and simply wish to elucidate interactive commonalities and network hubs, in which case a simple metasearch is sufficient. Other investigators may be interested in building a more inclusive network, with detailed descriptions, analyses and graphical displays of protein associations, with the goal of generating biomarkers for functional analyses.

We emphasize that generation of a composite network of interactions provides significantly more information about a given biological system than the sole consideration of primary literature. A gene/protein network not only presents an alternative iteration of existing data, it can also highlight how each data point precisely fits into the physical and/or functional cellular milieu. Much of this chapter focuses on *in silico* resources and how to maximize the utility of retrieved information, to generate hypotheses and to design focused experiments. We also describe how some of the tools we introduce can be used to build customized networks around groups of proteins directly identified through the techniques described in other chapters of this book.

1.1 Choice of Network Building Modalities and the Corresponding Analysis Tools

As mentioned in the introduction, before any project is initiated, it is imperative to first determine the degree of data analysis that is appropriate. We can envision three different likely scenarios:

Scenario One—If the goal is to quickly survey the biological connections of a rather small group of genes (e.g., hits selected based on some investigator-nominated criterion from a low- or mid-throughput screen), then the best choice for information retrieval would be metasearch platforms, such as STRING [1] or GeneMANIA [2] (*see 1.2 Online Metasearches*). These online-only tools allow retrieval of interactions from a variety of sources, including multiple protein-protein interaction (PPI) databases, expression databases, signal transduction maps and others. Each item of information provided by the metasearch platforms is traceable to the original publication and some sort of interaction reliability index is frequently provided. Metasearch engines can also provide information on protein structures and functions and can perform basic network analysis to identify, for example, overrepresented Gene Ontology (GO) [3] categories. Most metasearch programs have visualization features that produce attractive and modifiable network maps, which can be saved as publication-quality images. Metasearch programs do have some disadvantages. For example, when a larger number of entities are displayed (approximately 50 or more), these maps quickly become incomprehensible. Therefore, to visualize and analyze larger networks or sub-networks, specialized programs are needed (as described in Scenario three). The other drawback of relying solely on metasearch engines is the incomplete coverage of identifiable interactors (*see 3.1 Network Assembly and Analysis*).

Scenario two—In this scenario, an investigator needs to investigate a small-to-midsized, defined group of proteins to assess functional interactions. All that is essentially desired in this scenario is a composite list of genes that interact in some way with the proteins of

particular interest, without the need to visualize the network. Our recommendation is to perform a more detailed search than suggested for Scenario one, and retrieving information from a number of unique protein interaction databases either via metasearch programs, or through direct query for PPIs. Gene names for proteins of interest must be standardized to a consistent nomenclature before and after searching different databases (*see 1.3 Gene/Protein ID Conversion*), which allows information retrieved from multiple PPI databases to be stored and cross-compared for redundancy in a spreadsheet-based program such as Excel (*see 1.4 Databases describing Protein-Protein Interactions*).

In **Scenario three**, an investigator may want to not only collect the identities of interacting proteins, but also build a complex network in order to derive additional information from the network's architecture (topology). The network can also be built as a reference tool for a larger project, with relevant information embedded in the network or hyperlinked to specific **nodes** (genes or proteins) and/or **edges** (lines representing interactions). Additionally, advanced tools can be used to carefully analyze and logically visualize the properties of individual entities (such as domain architecture and evolutionary sequence conservation) as well as the whole network (functional enrichment). Moreover, investigator-generated experimental data can be reiteratively incorporated into an existing network, aiding in the interpretation of the results. For this particular scenario, more in depth data analysis and advanced network building is required. Several sophisticated yet user-friendly programs exist, which assemble all the information retrieved during the database searches. Even with the more sophisticated programs, Excel is still an invaluable tool for ID comparison and conversion, as noted for Scenario two.

1.1.1 Getting started: selecting a platform: The bulk of this chapter represents a logical, step-wise approach for selecting data sources that eventually allow the investigator to customize assembly and analysis of complex networks (Figure 1). Installing a user-friendly software platform to visualize and assemble datasets from different sources is a very important step. A periodically updated list of programs available for network visualization and analysis can be found at the website http://wiki.cytoscape.org/Network_analysis_links. However, many of the listed programs, resemble butterflies in at least two ways: they are colorful and they have very similar life spans. The few reliable programs available have many similar features, like the ability to retrieve protein-interaction data from external sources or import networks from existing files. Additional essential features include the capacity for network topology analysis, functional enrichment of a network and generation of publication quality figures. Interactive graph layout, statistical network analysis and other features are available in some but not all options (a complete list of features can be found at http://wiki.cytoscape.org/Network_analysis_links).

For novice users, we believe that the most suitable program to use is either Cytoscape [4] or VisANT [5]. One exception is if 3D visualization capabilities are required, in which case BioLayout *Express*^{3D} [6] may be the best choice. Each investigator should consider a series of important questions regarding the ultimate goals of the project when choosing a particular program. The questions we deem most relevant when comparing programs are the following (we decided to compare Cytoscape and VisANT as an example – Figure 2 lists several other available software programs):

1. Is ease of use more important, or does the number of tools available for data analysis take precedence? For ease of use, VisANT is preferred; on the other hand, Cytoscape supports more tools for analysis.
2. For visualization, how important are display aspects (e.g., multiple color tones and shapes)? (Cytoscape > VisANT)

3. Is it necessary to have multiple networks open at the same time? (Cytoscape only)
4. Is it important for the resources to be open-source (open-source can encourage crowdsourcing and innovation)? (Cytoscape and VisANT are both open-source)
5. Is it important for the resources to be free of charge (e.g., Cytoscape and VisANT), or is a fee acceptable (e.g., Ingenuity Pathway Analysis)?
6. How relevant is the availability of plugins to expand the functions of the software program (Cytoscape > VisANT).

In this chapter, we use Cytoscape as our main platform for network building and will only briefly elaborate on VisANT and BioLayout Express^{3D}. BioLayout *Express*^{3D} provides a fast and powerful interface for the visualization of networks in three dimensions, which may help to interpret the oftentimes complex relationships among nodes in an extensive network. While the 3D capacity of BioLayout Express^{3D} is a unique benefit, the deficit of this program is that some features, such as manual rearrangement of nodes and the clarity of node labeling, have not been optimized. However, the option to either run the program online or use it as a downloaded application is, like the 3D capability, another attractive feature.

VisANT is an integrative software platform for the visualization, mining, analysis and modeling of biological networks. VisANT can either be run as a web-based application or as a local Java application. Of particular convenience is VisANT's ability to continuously retrieve information from the parent database, Predictome [7], which allows uninterrupted one-click expansion of any node in the network. The VisANT name resolution feature, which is automatically applied, is also quite useful. If interactions with different protein isoforms were retrieved, the name resolution feature recognizes all of them as descendants of the same gene, and maps them as a single entity (node) in the network. Other very convenient features include the ability to effortlessly select nodes or edges based on several criteria, such as: the number of different classes of experimental methods that together support the existence of a specific interaction, nodes in which the indicated protein has a human ortholog, specific GO annotations or the specified number of interactions within a given network. Unresolved nuisances include the absence of the "undo" function and the inability to select multiple edges at once. Nevertheless, VisANT is a viable alternative to Cytoscape and is well worth looking into.

Cytoscape is one of the most popular tools for visualizing, integrating, modeling, and analyzing molecular and genetic interaction networks (Figure 2) and we highly recommend this platform [4, 8, 9]. Instead of being intrinsically linked to a single, unique biological database, it allows users to install their own favorite search tools (i.e., plugins) to access multiple other databases (Figure 3). It is a free access and open source platform, which can be downloaded at <http://www.cytoscape.org/>. The various Cytoscape plugins mentioned throughout this chapter can be intrinsically downloaded through the 'Manage Plugins' option. (Open Cytoscape -> Plugins -> Manage Plugins).

In Cytoscape, data are integrated within interactive nodes or edges, based on individual attributes, such as interactions among proteins or functional associations of proteins. Users have the flexibility to filter results according to specified characteristics, perform complex network searches, or simply highlight different components of a network by changing the shape, color, or size of nodes and edges. The manual network manipulations we describe in this chapter allow the user to highlight aspects and information of a particular protein of interest. Cytoscape also allows users to import protein data files and assembled networks as tab delimited or .xls files. Further details regarding Cytoscape can be found in the official

Cytoscape manual available on the provider's website (http://www.cytoscape.org/documentation_users.html).

1.2 Online Metasearches

When performing a basic search for protein interactions, it is important to recognize that much of the publicly available data is dispersed across multiple distinct databases rather than integrated into a single site [10]. Metasearch platforms simultaneously access multiple individual databases and produce not only a composite results list, but also a network with detailed information about nodes (proteins) and edges (interactions). Two examples of freely accessible metasearch platforms with emphasis on visualizing search results are STRING (Search Tool for the Retrieval of Interacting Genes) [1, 11] and GeneMANIA [2, 12, 13]. Commercially available metasearch platforms are also available, and include: Ingenuity (<http://www.ingenuity.com/>), GeneGo (<http://www.genego.com/>) and Protein Lounge (<http://www.proteinlounge.com/>), which generally advertise more extensive data curation by experts. Most of the listed metasearch providers – free as well as commercially available – also utilize different algorithms to supplement the data retrieved from the various individual databases.

Although the benefits of using metasearch tools greatly outweigh their shortcomings, it is important to be aware of some of the downsides of using metasearches. For example, when we evaluated some of the different available resources with test queries, we found that the results obtained by individually querying databases discussed in subsequent sections were not identical to the results provided by metasearch platforms that accessed the same exact databases (Figure 4; *see Note 1*).

STRING—(Search Tool for the Retrieval of Interacting Genes; <http://string-db.org/>) combines comprehensiveness, interactive elements and an intuitive user interface [1]. Unique among metasearch programs, STRING covers more than 1100 completely sequenced organisms, ranging from bacteria to Archaea to humans. Conveniently, prior to performing a search, STRING presents an interactive disambiguation page, which takes care of many of the problems with gene/protein identifiers described below (*see 1.3 Gene/Protein ID Conversion*). Interactions in STRING are assigned a confidence score describing how strong the evidence supporting reported interactions; summaries of protein functional domains, together with 3D structures are also available. Based on the integration of data from multiple databases, STRING covers around 220 million interactions [1]. As a source for interaction information, STRING uses cross-references with a number of widely used websites, including UniProt [14], SwissModelRepository [15] and SMART [16]. Up to 500 proteins can be simultaneously used as input queries, saving time on large projects (however, if large lists of protein names are to be queried using this or other metasearch programs, it is important to follow guidelines described in the 1.3 Gene/Protein ID Conversion). Aggregate results obtained from STRING can readily be exported as a text file, which can then be transferred into Cytoscape for further analysis.

As an example, we used the signaling protein SMAD1 - a member of the SMAD family of signal transducers and transcriptional modulators [17] - as the text input in the STRING dialog box, applying default search settings. This produced a SMAD1-centered 10-interactor network (Figure 5). Networks constructed by STRING are interactive and nodes are freely

Note 1 Despite the extensive number of databases accessed by the metasearch platforms, not all databases are available through every metasearch program, which may result in missed information. Because metasearch platforms access many different databases, some of them may provide the same proteins/genes multiple times, resulting in output redundancies that the investigator must subsequently purge.

movable by users, allowing the grouping of positive results to emphasize points of interest to the investigator.

GeneMANIA—(<http://www.genemania.org/>). This convenient and user-friendly metasearch platform generates a highly correlated signaling network centered on the protein(s) of interest. Up to 101 proteins can be searched at the same time in GeneMANIA [18], again facilitating large projects. Network visualization is one of GeneMANIA's strong suits: interaction types (e.g. co-expression and physical interactions) are shown in different colors and can quickly and conveniently be included or excluded by checking or unchecking the appropriate box on the side menu (Figure 6). For example, moving the mouse cursor over a gene in the "Genes" results list highlights only the selected gene and its interactors, while the rest of the network fades to grey (the same can be done by moving the cursor over functions associated with sections of the network). GeneMANIA also provides a "network weight" score for each interaction displayed, which can be viewed by clicking on a given interaction.

The seven organisms covered by GeneMANIA include: *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus*, and *Homo sapiens*. This search engine accesses data from GEO [19], BioGRID [20], Pathway Commons [21] and I2D (Interlogous Interaction Database) [22], as well as other organism-specific functional genomics datasets (for more details see <http://pages.GeneMANIA.org/help/>). The program allows users to upload their own data and select arbitrary subsets of the datasets associated with an organism or the type of interaction to perform their own analyses. Again, using SMAD1 as an example with the default setting, the query yielded a 20-interactor network (Figure 6). Data from GeneMANIA can also easily be imported into Cytoscape.

IPA—Ingenuity Pathway Analysis (IPA)® is a commercially available software program that helps researchers model, analyze and understand complex biological and chemical systems (<http://www.ingenuity.com/index.html>). It is very convenient to search the Ingenuity Knowledge Base for information on genes, proteins, chemicals, drugs and molecular relationships to build biological models. Figure 7 is an example of an IPA generated network using SMAD1 as the input protein. Unfortunately, the only way to export data from IPA is to produce a list of molecules in Excel format. However, it is not possible to generate a list of interactions, which makes it very inconvenient to import data from IPA into any other network visualization program.

As mentioned above, IPA is not freely accessible. In our opinion, the available free *in silico* resources such as STRING, GeneMANIA and others are not inferior to this commercially available product for the purposes of retrieving publicly available information; therefore, we will not describe use of commercially available resources in great detail and only use IPA for cross-database comparisons in subsequent sections.

Table 1 highlights the different results retrieved from the various databases; furthermore, a detailed comparison of the evaluated resources is provided in 3.1 Network Assembly and Analysis.

1.3 Gene/Protein ID Conversion: a first step for network construction

If screening databases independently or via metasearch platforms with large lists of genes, use of consistent gene nomenclature is essential for search programs to function effectively. Unfortunately, despite ongoing efforts, a systemic lack of standardization for nomenclature persists; the recording of gene information often varies significantly from database to database. The heterogeneity of identifiers is caused by the indiscriminate use of NCBI Gene

ID (a number), Gene symbol, Refseq, and other common names arising from the historical course of study of a gene to describe a given gene/protein. For example, for our test protein, SMAD1, the official Gene ID is 4086, the Gene Symbol is SMAD1 and the NCBI Reference Sequence is NM_005900.2, while the protein is also known by its synonyms JV4-1 and MADR1 or by its previous symbol MADH1 (mothers against decapentaplegic homolog 1) (http://www.genenames.org/data/hgnc_data.php?hgnc_id=6767). In rare cases, the identifiers are even unique to a given database and are not used by the scientific community at large. If one is preparing a list of proteins to screen, it is advisable to employ ID disambiguation, unification, and, if needed, conversion as a first step in a project, creating a homogenous dataset in order to apply tools to the entire network with greater ease. Conversely, once a list of proteins has been obtained from screening, re-application of Gene ID conversion tools will help integrate datasets and remove ambiguities and redundancies (see Note 2). In preparing for conversion, data from most resources can be downloaded as an Excel-compatible file or the data can be copied and pasted into Excel.

The gene-protein ID converters Clone/Gene ID Converter [23] and bioDBnet [24] are useful as starting points for identifier standardization. These programs essentially accept all the various gene identifiers in use as input; both of them very conveniently produce output files in Excel format.

An example of using bioDBnet to process a list of genes is shown in Figure 8. As a positive feature, every gene symbol is almost immediately converted to one of various output formats (e.g., Gene ID, PubMed ID, WormBase Gene ID, and others). However, it is necessary to be cautious (see Note 3). Manually checking the output data (reading gene descriptions and comparing the numbers of genes in the input and the output lists) and making appropriate corrections is highly recommended. However, the automated conversion systems do offer an additional advantage besides convenience, in that they are unlikely to “swap” one gene for another, whereas an investigator working manually might inadvertently substitute commonly used names for a given gene. For example, PKD1 (polycystic kidney disease 1) can easily be confused with PKD (protein kinase D) by an investigator, whereas this will not happen with a conversion program.

During our comparative analysis of searches using independent databases (discussed below), we found that when we compared SMAD1 interactors from different resources, without manual adjustments of Gene IDs, the overlap of interactors from database to database appeared to be only around 50%. However, once we examined the data more closely, we noticed that gene names in the MIND’s results list contained spaces in front of each gene symbol, while another database, BOND, included provisional entries without associated Gene ID, with names like “IMAGE 5740090.” Excel was used to address both issues (see Note 4). The outcome was a reduction from 127 originally identified interactors to 104. Once all the data had been standardized and manually adjusted, the overlap across all investigated databases was close to 90% (see Note 5).

Note 2 Cytoscape, like Clone/Gene ID Converter and bioDBnet, is case sensitive; therefore, Cytoscape interprets SMAD1, Smad1 and smad1 as three different entities.

Note 3 The gene-protein ID converter tools cannot always accurately convert every single protein if an extensive list is used as input: losing nodes to ID conversion is a common problem when merging information from databases solely using automated approaches. Considering the output data in Figure 8B, several input ‘Gene Symbol’ data could not be converted to output ‘Gene ID’ data. To find information on specific genes that were not assigned an ID, NCBI services should be utilized (see Figure 8B legend). Additionally, several input ‘Gene Symbol’ data were assigned multiple Gene IDs, in which case it is necessary to click on each of the Gene IDs (link to the NCBI webpage) and manually verify the appropriate one.

Errors may occur for several reasons; for example, many Entrez IDs are replaced or removed on a regular basis due to advances in knowledge, making it difficult for the curators of the gene-protein ID converters to stay up-to-date. Furthermore, incorrect use of upper and lower case letters can also be a factor that may distort the output results provided by a given converter (it is advisable to convert gene names into all capital letters – in the case of human genes - prior to use of gene ID converters).

1.4 Databases describing Protein-Protein Interactions

Investigators who are determined to do exhaustive data collection can augment or replace metasearches by directly mining the various dedicated databases to identify protein-protein interactions (PPIs) (Figure 1). Table 2 covers a subset of the available databases dedicated to PPIs: those listed in the table have been acknowledged in a large number of citations and have a good reputation for accuracy [25]. Several commonalities can be appreciated across the different databases; nonetheless, some are clearly more advantageous to use than others. In determining the value of each individual database, particular significance was assigned to the frequency of updates, curation of interactions, ease of use and ease of exporting results (*see Note 6*).

Figure 9 highlights the range of specific interactions likely to be identified by each approach out of the total number of identifiable interactions (established by evaluating all the databases mentioned in Table 1). The percentage of interactions identified depends on the number of sources used; taking this information into consideration may be useful when deciding how many sources are appropriate to access. For example, some investigators may only be interested in “High-Confidence” data, i.e., data that was identified in every single source accessed. Other investigators may be more interested in having the most inclusive dataset possible, capturing interactions found in only a single resource. Alternatively, investigators may choose a more balanced, “Mid-Confidence”, approach, which imposes some level of restriction (Figure 10).

Before specific databases are discussed, we would like to point out a potential source of inconsistencies: the different handling of protein complexes. In some databases, protein complexes are deconstructed and individual query proteins are paired as interactors with all of the proteins associated with the complex. Other databases, like IntAct [26], apply an algorithm that deconstructs the n-ary interactions (i.e., primary, secondary, etc.) among proteins of a complex into binary interactions. Several databases (e.g. Bond^{OA}) have separate results sections for complexes, and each database may or may not indicate if complex-derived interactions are also counted as binary interactions (*see Note 7*).

Note 4 Simply using the “Replace” function in Excel allowed us to remove all the spaces from the MINT dataset effortlessly. Removing unconventional gene identifiers without losing valuable information was slightly more challenging. The easiest most accurate approach was to export the results as an Excel file, sort the data by Gene ID and then exclude all the entries without Gene IDs.

Note 5 It is important to realize that when Excel is used, some gene names may automatically be converted to dates – SEPT7 (septin 7) becomes 7-Sep and MARCH9 (membrane-associated ring finger (C3HC4) 9) is turned into 9-Mar – further emphasizing the necessity to manually inspect one’s data (OpenOffice and LibreOffice also convert text to dates; Apple Numbers does not). Importing data that include calendar dates may cause software programs to crash, or may lead to the inadvertent loss of data points.

In our opinion, the most practical approach to resolve the conversion of gene names to dates issue is by using the ‘sort’ option in Excel (Data -> Sort -> Ascending) once gene symbols have been pasted into a column. Sorting the list in alphabetical order moves all the dates to the top of the list, which allows quick and reliable identification of unwanted changes. Manual modification is then required (to avoid having names changed to dates, enter an apostrophe (‘) followed by the gene name – e.g., ‘Sept7). An alternative approach is to open an empty worksheet, select the column into which gene symbols will be pasted and change the cell format from ‘General’ to ‘Text’ (Format -> Cells -> Number -> Text). Two problems may surface with this approach: 1) if the ‘Text to Columns’ (Data -> Text to Columns) function has to be used, all gene names associated with dates will be changed to dates again; and 2) formulas cannot be used in any of the cells in ‘Text’ format.

Though the process of producing an optimal dataset may seem tedious, we believe that it is essential for flawless data interpretation and advanced network building. In the case of extensive gene lists, manual curation becomes exceedingly impractical; it may be in the user’s interest to forgo using Excel and accept minor inaccuracy, or focus on subsections of a created network for manual curation.

Note 6 For the most part, every PPI database investigated is searchable by using a text input field; however, the optimal input format may vary, as do the advanced search options. In some cases, as for BOND^{OA} [27], the most pertinent results are obtained when using UniProtKB IDs (<http://www.uniprot.org>), Entrez IDs, and other options instead of the official gene symbol. Within most of the databases, protein interactions are annotated with detailed information presented as a “report” of a protein interaction and the relevant experimental data.

Note 7 The worst-case scenario is when no information is provided as to how protein complexes are managed by a given database. These discrepancies are unlikely to disturb most users; however, if an investigator’s protein of interest is part of a complex, it is important for the investigator to understand how the various databases handle protein complexes.

In examples below, we use SMAD1 again as a demonstration case for comparison of outputs from four useful PPI databases: BioGRID [20], IntAct [26], Bond^{OA} [27] and MINT [28]. A graph comparing the numbers of interactions for SMAD1 found in each of the four databases is shown in (Figure 11). The SMAD1 example clearly underlines why it may be advantageous to utilize more than one database and/or the metasearch platforms for any given search.

BioGRID—The Biological General Repository for Interaction Datasets (BioGRID) is a collection of protein and genetic interactions from thirty-eight organisms [20]. BioGRID searches 34,996 publications for 557,934 raw protein and genetic interactions (as of October 2012; <http://www.thebiogrid.com>). Data is compiled from high-throughput and low-throughput experiments through comprehensive curation efforts. Furthermore, this database is updated monthly, with newly identified interactions routinely added. For the SMAD1 example, BioGRID identified 68 unique interactions for *Homo sapiens* (Figure 11). These interactions were accumulated from 60 publications, totaling 126 physical interactions (a total of 126 interactions were identified, 68 of which were unique) and 5 genetic interactions (Figure 12C).

BOND^{OA}—The extensive Biomolecular Object Network Database (BOND^{OA}) [27] database includes BIND (Biomolecular Interaction Network Database), which is updated on a daily basis and contains 188,517 PPIs. All the recorded interactions searchable in BOND^{OA} have been experimentally verified and were published in at least one peer-reviewed journal (<http://bond.unleashedinformatics.com/>). In general terms, BOND^{OA} is a great resource, with only two significant inconveniences. First, it is difficult to perform organism specific searches. It is not possible to directly select an organism of interest; instead, one must use more specific input than the official symbol, such as a Gene ID, which is inherently organism specific. Second, one must register as a user, and is frequently asked to log back in. Using the Entrez Gene ID for human SMAD1 (in order to only retrieve interactions identified in *H. sapiens*) as the search input, BOND^{OA} identified 104 PPIs (Figure 11).

IntAct—The European Bioinformatics Institute's IntAct (EMBL-EBI IntAct) has 304,667 recorded interactions (including drug-protein interactions) from 5711 publications and is updated on a monthly basis (<http://www.ebi.ac.uk/intact/>). All interactions are based on literature curation or direct user submissions [26]. Interestingly enough, IntAct only identified 7 binary interactions with SMAD1 (Figure 11) with the exclusion of protein-complex expansion. This was somewhat surprising considering the search results obtained from BOND^{OA} and BioGRID.

MINT—Lastly, the Molecular INTERaction database (MINT) was explored. MINT only includes experimentally verified PPIs mined by expert curators [28]. It is not as professional appearing as the other three databases discussed, but it is easy to use and offers a unique web-based viewer (Figure 12). The main disadvantage noticed when we searched for SMAD1 interactions was that results for organisms other than *Homo sapiens* were included in the output table; even though, it had been specified to only find results for *H. sapiens*. To eliminate results for species other than *H. sapiens*, we copied and pasted the results table into an Excel spreadsheet and manually removed all unwanted interactions. The final results list included 53 interactions for SMAD1 (Table 1).

To create a maximally inclusive resource, all the retrieved results for SMAD1 interactions from the four different databases were imported into Cytoscape by downloading the complete binary interaction files or by copying and pasting results into an Excel spreadsheet

and then transferring the list to Cytoscape. These imported files were then used to create networks within Cytoscape (Figure 13). Downloading the data files into a single resource is more practical than using the web-browser interface for each isolated database, because the composite dataset allows customization of one aggregated file (e.g., application of Gene/Protein ID conversion, to eliminate redundancies), and also allows the user to generate subsets of the network based on specified attributes. The networks generated from the binary interaction files allowed us to also retrieve interactions between, for example, the first neighbors (proteins that directly interact with a specified protein), whereas the web-browser interfaces for the individual databases often only show the interactions between the protein of interest and its first neighbors, without showing interactions among proteins other than the input protein.

1.5 Resources Beyond Protein-Protein Interaction Databases

PPI datasets typically comprise the backbone of interaction networks. However, the full informative potential of a network is unlikely to be reached without consideration of other sources of information such as metabolic and signal transduction maps, expression databases, gene-disease-drug associations and reports of gene fusions. Some of the available resources do not provide meaningful information in all instances: for example, gene fusion events are more frequent in prokaryotes, and not every gene is linked to a disease or has relevant microarray data. It is thus up to each investigator to determine which sources of information are relevant for a specific gene or a specific group of genes. Below we will give a brief overview of some of these information sources as well as bring up general considerations on how to merge all retrieved data into a coherent set (Figure 10).

1.5.1 Pathway Databases—Important information about protein functionality and interactions arises from low throughput experiments analyzing signaling relationships that are published in the primary scientific literature. PubMed, the central resource for listing biomedical publications, currently contains over 22 million entries (<http://www.ncbi.nlm.nih.gov/pubmed>), making it difficult for the individual investigator to stay abreast of recent developments impacting more than a small number of proteins. Pathway databases assemble the results from individual experiments and sort them based on cellular signaling pathway associations. Cellular signaling pathway maps can greatly assist in the realization of molecular functions and they can be extremely useful in the identification of proteins upstream and/or downstream of specific pathway components.

For the purpose of this chapter, cell signaling and metabolic pathway maps can be defined as snippets of knowledge restricted to a particular process in the cell, where the entities (most frequently proteins and small molecules) are connected with focus on functionality. One caveat concerning pathway databases is that pathways are inherently difficult to demarcate (unlike the more objectively determined PPIs), with experts deciding where the “natural boundaries” of a particular process are. Consequently, even the maps of well-studied pathways, when collected from different sources, may have less than expected overlap. A good example is the EGFR (epidermal growth factor receptor) signaling pathway (Figure 14A): in preparation of an EGFR-centered interaction network [29], five different pathway maps were scrutinized. Out of 371 unique proteins collectively nominated by these maps, only 10 were common among all five pathway maps (Figure 14B). Nevertheless, cellular signaling pathway maps can help elucidate molecular functions and they can be extremely useful in the identification of proteins upstream and/or downstream of specific pathway components.

Currently, there are many popular pathway databases that can be accessed without a fee. These include the Kyoto Encyclopedia of Genes and Genomes (KEGG) [30], Pathway

Commons [21], REACTOME [31, 32], the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) [33], BioCarta (<http://www.BioCarta.com/>), Netpath [34], PID [35], Wikipathways [36] and several others.

Essentially, there are two modes of extraction relevant information from pathway databases. In some cases, every single entity of a particular pathway may be significant to the function of a protein of interest. If that is so, then the whole list of participating proteins should be exported. Most of the pathway databases include a simple to use export function: a ‘protein list’ button in BioCarta, a ‘molecule list’ in NCI/Nature pathways and ‘export all interactors’ in REACTOME (see Note 8).

A second approach to the extraction of pathway information is to extract only molecules that are in the immediate neighborhood of the “seed” protein (i.e. the input protein). This can be done either manually at the time of pathway scrutiny, or by using Cytoscape, which provides options to create sub-networks centered on a specified node and its “first-neighbors” (see Figure 15 for details).

BioCarta (<http://www.biocarta.com>) is one of the dedicated databases for cellular pathway information and its website provides open and immediate access to existing pathways. It also offers a dynamic platform for information exchange and collaboration. Parts of the open-access dynamism are the graphical templates provided by BioCarta, to encourage public participation in data expansion. So called “Gurus” (experts in the field) evaluate submissions and, if appropriate, merge pathway information. Users can view molecular interactions in ‘Guru-curated’ graphical maps, by browsing the pathway listings and then selection pathways of interest. Alternatively, pathways can also be searched by pathway name, gene name or ‘multiple-gene’ search (up to three). BioCarta provides information for over 120,000 genes from humans and some of the most often studied species (e.g., mouse, fly and yeast).

REACTOME is another freely available pathway databases with curation of data done by experts [31, 32]. One strong suit of REACTOME is its pathway annotation. Annotations are written by biologists and cross-referenced to several other well-known databases like NCBI, Entrez Gene and PubMed [31, 32]. The foundation of pathways accessible through REACTOME is “biomolecular reactions.” Reactions among entities such as nucleic acids, proteins and small molecules are used to assemble network, which in turn are grouped together to give rise to a particular pathway [32]. Also, data from the REACTOME database on *H. sapiens* is used to infer orthologous events in twenty non-human species. REACTOME makes it easy for users to upload a list of experimentally identified proteins and it can also be used to analyze expression data. For the SMAD1 example, REACTOME produced 220 results, which either included SMAD1 itself or a member of the SMAD family (Figure 16). Once a specific hyperlinked result has been selected, all interactions in the opened pathway can be exported (see Note 9).

Note 8 There are two caveats to this approach, one conceptual and one technical. Firstly, some pathways may contain hundreds, if not thousands, of molecules, not all of which are equally relevant in regard to the protein of interest. Secondly, since many components of a given pathway are frequently depicted in multiple forms (e.g., active/inactive or with specific modifications, e.g., phosphorylated and glycosylated), the dataset is likely to contain several redundant entries that need to be removed (Section 1.3 Gene/Protein ID Conversion).

Note 9 We found that the most convenient way of exporting data from REACTOME is clicking on ‘Analyze, Annotate & Upload’ once a pathway is open, then selecting the ‘Interaction Database’ and clicking on ‘Display table of all interactors for pathway.’ The provided table can be copied and pasted into an Excel spreadsheet (as long as care is taken not to highlight the title line, which will inhibit the ‘paste’ function in Excel). The ‘Export all interactors for pathway’ button opens up a new window with information arranged by UniProtKB id, without reference to the gene symbol, perhaps making it slightly less informative than the table. There are no convenient ways to export the result data from all the sources accessed by REACTOME at once. REACTOME also does not offer the ability to refine searches beyond specifying pathways, reactions, proteins or ‘others’ as stipulations.

PID (The National Cancer Institute/Nature Pathway Interaction Database) is a well-curated and peer-reviewed collection of biomolecular interactions and signaling pathways [35]. Unfortunately, it stopped updating its database in September 2012, and it has not yet been determined if this truly outstanding resources will be maintained by a separate entity. Data from three different sources are amalgamated to generate the PID database. The first set of data is NCI-Nature curated data (137 human pathways – 9248 interactions), which was created by Nature Publishing Group editors and reviewed by experts in the field. The second data source is BioCarta: this data is imported without expert review. And the third type of important data is from REACTOME. BioCarta and REACTOME together provide information on 322 human pathways, covering 7575 interactions.

PID's advanced search tools allow users to perform a “batch query” for up to two groups (uploaded lists) of genes. The user can select if pathways are to be identified for each group – the results are displayed in a table with identified pathways for each entry in each of the lists shown side-by-side - or if a complete network map is to be generated (generally fragmented, unless all input genes fit neatly into a single pathway) (*see* Note 10).

Pathway Commons is a metasearch platform, which, in addition to PPI data retrieval, also collects pathway data from multiple publicly available databases, including REACTOME [31], Systems Biology Center New York (<http://sbcny.org/data.htm>), HumanCyc (<http://humancyc.org/>), The Cancer Cell Map (cancer.cellmap.org) and PID [35]. Pathway Commons includes biochemical reactions, complex assembly, transport and catalytic events and physical interactions. The data is updated with varying frequency. Pathway Commons currently contains data on about 1,700 pathways and 442,182 interactions from 414 organisms. While REACTOME provides unique pathway information, Pathway Commons does not have its own database. Pathway Commons can easily be accessed either directly online, or through Cytoscape's built-in import “Network from Web Services” function (Figure 3C). After a keyword search for a selected organism has been initiated, result specifications can be made based on pathway involvement, data resources and interaction types (e.g. physical interactions and biochemical interactions). The ‘retrieve interactions’ command will finish importing the network.

A variety of alternative sources can also be tapped for pathway maps. Commercial software programs (e.g., Pathway Studio Visualizing the Science) all include their own versions of generated pathway maps, with expert curation and professional designs. In addition to the few excellent dedicated databases described above, there are several other resources, maintained by the science academic community (e.g., <http://biocyc.org>), various life science companies (cellsignal.com & sabiosciences.com; these frequently provide detailed maps as background for informed product searches), web sites of individual labs (best searched for using Google image search), and, finally, peer reviewed publications (*see* Note 11).

In our opinion, merging and analyzing several versions of a pathway from different sources increases the accuracy of the information, compared to solely relying on a single source, no matter how renowned the single source may be. As an example, consider the way we decided to resolve the EGFR pathway data discrepancies mentioned above. We simply

Note 10 The output styles include: jpg (interactive graphic), svg (interactive graphic that requires the SVG plugin), xml (text) and BioPax (text).

Note 11 It should be mentioned; however, that the differences among pathway databases is somewhat blurred, with the commercially produced pathway maps frequently used and posted by third parties; maps created by life science companies adopted and/or disseminated by academic institutions (BioCarta maps available through National Cancer Institute/Nature Pathways Interaction Database (PID)) [35]; Roche's map adopted by *ExPASy* - a bioinformatics resource portal operated by the Swiss Institute of Bioinformatics; and other examples. Hence, a certain degree of diligence is necessary: if two maps imported from different sources match exactly, it may be that different incarnations of the same parent map are being evaluated.

selected a “core” subset of proteins, which were present in at least two of the five pathway databases evaluated (103 of 371 unique proteins, Figure 14B), which is in line with the ‘Mid-Confidence’ vetting approach mentioned in Figure 10. This selection process ensures that our network only includes proteins that were determined to be part of the EGFR pathway by two (or more) independent curators.

1.5.2 Microarray Databases—The increasing availability of gene expression technologies (such as microarray and RNA sequencing) has resulted in the publication of thousands of gene expression datasets, exploring a variety of biological conditions. Integrating data derived from PPI databases and other sources of information with predictions derived from expression data can greatly improve the overall predictive value of a generated network. Essentially, there are two major ways to utilize expression technology-derived data. First, it is often possible to infer important functional relationships by assessing which genes show changes in mRNA expression as a consequence of manipulating a protein of interest. Such manipulations may include pharmacological inhibition, gene deletion or siRNA inactivation or overexpression. The second way is focused on co-expression (i.e., comparing expression profiles) of genes, which is particularly important for large-scale screening experiments. Relevant experimental data can be found in a number of public databases devoted to high-throughput gene expression and genomic hybridization. For the purpose of constructing a signaling network, Gene Expression Omnibus (GEO) [37, 38] and ArrayExpress (EBI) [39–41] are especially helpful.

GEO (Gene Expression Omnibus) can be searched directly from the NCBI main web page (<http://www.ncbi.nlm.nih.gov/geo/>) using gene symbols [19]. Retrieved information is presented as a list of experiments, which a user can then browse to identify those deemed relevant. Once all appropriate results are identified, GEO provides a flexible and user-friendly online tool (GEO2R), which allows comparisons between two or more sets of results. GEO2R thus provides the means to identify varying gene expression levels across different experimental conditions. The resulting table of genes can then be exported and subsequently integrated into a network (using Cytoscape for example) or into a gene list in Excel.

ArrayExpress serves a similar function and also provides a public database of microarray gene expression data (<http://www.ebi.ac.uk/arrayexpress/>). While GEO lists individual experiments and their conditions, ArrayExpress groups different genes first based on type of organism, and then subdivides each group into categories of cell lines, cell types, compound treatment, developmental stages, disease states, infections, organism parts and phenotypes. ArrayExpress makes it conveniently easy to access the exact expression data an investigator is interested in [41].

We would like to point out that expression technology-derived data can not only be utilized by looking at specific changes caused by a unique treatment, but also by using large microarray compendia to create an expression profile for any given gene across the entire range of experimental conditions tested [42]. This allows investigators to perform searches for genes with similar transcriptional profiles (i.e. genes that display similar transcriptional responses across a variety of experimental conditions). In both, yeast and humans (SPELL – Serial Pattern of Expression Levels Locator, <http://spell.yeastgenome.org/search/>; Connectivity Map, <http://www.broadinstitute.org/cmap/>, [43]) this approach has been shown to have a high predictive value. STRING (see 1.2 Online Metasearches) has an option to search for genes with high co-expression values (based on expression technology-derived data) (Figure 6). Similar patterns of mRNA expression measured by DNA arrays and similar technologies across databases increase the co-expression score assigned by STRING to

interactions between genes. GeneMANIA also provides co-expression data, with more detailed links to individual publications (*see* Note 12).

1.5.3 Drug-Protein Interactions—In a number of approaches, proteins are pulled out of a dataset if they share an interactor (be it a third protein, a small molecule inhibitor or the like) or participate in a common well-defined process (i.e., a particular disease). The identified proteins are then used to generate a network (interactions amongst identified proteins are not necessarily related to the input protein, small molecule or disease process). This section will focus on drug-protein interactions and target-protein network generation.

The DrugBank database is a freely available drug focused dataset and is a unique bioinformatics and cheminformatics resource that contains 6711 drug entries (including ~1600 FDA-approved drugs and ~5100 experimental drugs) and ~4200 non-redundant sequences linked to these drug entries. Entering a protein name into the search box results in a list of names of available drugs or the name of drugs targeting proteins associated with the input protein, as long as such data is available in the dataset. Each drug given in the results section is hyperlinked to additional information, such as chemical structure, type, pharmacological information and targets. Although several proteins can be entered into the search box at once (comma-delimited), the output is not quite network ready and requires (as is the case for several other databases) manual transfer into tabular (Excel or tab-delimited .txt) format for subsequent import in Cytoscape. A somewhat smaller subset of DrugBank (version 2006, with ~1200 FDA-approved drugs and ~3000 experimental drugs) can be downloaded as an Excel file from the supplementary materials of Yildirim et al. [44]. Yildirim et al. generated a dataset based on US Food and Drug Administration-approved drugs, in which two proteins are clustered together if they are targeted by the same drug (<http://www.nature.com/nbt/journal/v25/n10/extref/nbt1338-S2.xls>). The information provided as an Excel spreadsheet can then be used in Cytoscape or other network visualization programs.

STITCH: Once a drug of interest has been identified in The DrugBank database, it can be helpful to then use Search Tool for Interactions of Chemicals (STITCH) - which is a user-friendly drug-protein interaction database derived from STRING – to query the drug name. STITCH integrates information about interactions from metabolic pathways, crystal structures, binding experiments and drug–target relationships, and generates a network [45]. It also uses information from phenotype evaluations, text mining and chemical structure similarities to predict interactions between chemicals. The ability to explore drug interactions with an entire network of associated proteins is another useful feature available, as is the ease of locating the original data source for each provided interaction. This database essentially helps users gain a better picture of the biochemistry of metabolites, drugs and other compounds (*see* Note 13).

Note 12 Despite the deceptive simplicity of extracting lists of genes from gene expression experiments, we strongly recommend that retrieved expression technology-derived data is analyzed by a bioinformatics professional before any conclusions are made. The quality of the underlying array experiments needs to be verified to ensure reliability.

Note 13 An important caveat regarding STITCH is in order due to the fact that STITCH uses the same ‘confidence score’ system (interpretation of the evidence supporting an interaction) as STRING. If a given protein of interest has many high confidence protein interactors, these proteins are often deemed more relevant by the program than any pharmacological agents that target the input protein (pharmacological agents are sometimes assigned lower confidence scores). Therefore, if the default settings (network only includes the ten interactors with the highest confidence score) are used, relevant pharmacological agents are often not included in the network. For example, entering EGFR into the search box results in a network of the ten highest scored interactors, which does not include Erlotinib (Tyrosine Kinase Inhibitor); even though, Erlotinib is in the database. One option to address this issue is to use the pharmacological agent (e.g., identified in DrugBank) as input text instead of using a gene/protein. Another option is to change the number of interactors shown from ten (default setting) to twenty or higher (*see* Figure 5 – the “Info & Parameter” panel for STRING and STITCH are identical).

The over 300,000 small molecules (in addition to 2.6 million proteins) included in the STITCH database are derived from PubChem [46]. Interactions among chemicals are based on similarities of activity profiles in the NCI-60 cell lines; activity profiles were established using pharmacological actions assigned to chemicals in Medical Subject Headings (MeSH) and from the literature. Drug-drug and drug-protein associations are integrated from pathway-focused databases, experimental databases and the literature.

Information sets from DrugBank and STITCH are non-redundant and should be merged either in Excel or in Cytoscape using uniform gene/drug names. Although the resources listed in this section are very convenient for network building, both, STITCH and The DrugBank database are updated irregularly. Therefore for exhaustive and most up-to-date coverage, researchers should still resort to manual searches using PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>), Google Scholar (<http://scholar.google.com/>), Ingenuity (<http://www.ingenuity.com/>) or other resources.

1.5.4 Text Mining—The term “text mining” refers to computer-assisted extraction of valuable bits of information from electronically available literary sources [47]. Text mining is a useful method to identify additional functional or physical interactions not sufficiently covered by some of the databases described in previous sections. Perhaps the best-known text mining tool is PubMed, which offers access to over twenty-two million citations (<http://www.ncbi.nlm.nih.gov/books/NBK3827/>). Chilibot (<http://www.chilibot.net/>), Google Scholar (<http://scholar.google.com/>), iHop (information hyperlinked over proteins) [48] and Microsoft Academic Search (<http://academic.research.microsoft.com/>) are additional useful text mining platforms. As an example of text mining, consider AURKA (aurora kinase A) and one of its activators in mitosis, NEDD9 (neural precursor cell expressed, developmentally down-regulated 9). We found that GeneMANIA (*see 1.3 Online Metasearches*) does not identify the established interaction between NEDD9 and AURKA [49]; on the other hand, Chilibot was able to identify several publications describing the interaction between NEDD9 and AURKA (Figure 17). It should be obvious that text mining cannot only be used to add information to an existing network; it can also be a starting point for collecting data (Figure 1).

Unfortunately, a necessary prerequisite for information extraction is automated recognition of biological entities, such as diseases, genes, proteins or drugs, which is often challenging because of the ambiguities created by jargon and terminological inconsistencies. Therefore, information retrieved via text mining requires careful scrutiny, to avoid acceptance of inappropriate data. Also, not all text mining tools result in an easily identified set of interactors of a protein of interest. For example searching ‘NEDD9 interactions’ in PubMed results in a list of 67 identified articles, whereas STRING generates a network with directly identifiable interactors.

Finally, some text mining tools attempt to connect seemingly disparate proteins/genes not mentioned together in any one paper, by finding intermediate links between them. For example, BITOLA (<http://ibmi3.mf.uni-lj.si/bitola/>), when given the names of two apparently disconnected genes/proteins, attempts to identify concepts (e.g., gene names or diseases), which are associated with either of the two entries. Output data is presented in tabular form, with the numbers of papers linking each concept to either of the two input genes/proteins. Concepts that are associated with both proteins/genes can then easily be identified.

1.5.5 Advanced Metasearches—Three additional metasearch platforms are available for advanced metasearches best suited for investigators who want to take full advantage of

available resources: iRefScape, Michigan Molecular Interactions and ‘Web Services Import’. These platforms are conveniently accessed through Cytoscape plugins (Figure 3).

“Web Services Import” is a built-in function in Cytoscape and does not require any additional plugins. It can be accessed from within Cytoscape: File -> Import -> Network from Web Services. Four options are available in the pull down “Data Source” menu: PSICQUIC, BioCyc, NCBI Entrez and Pathway Commons. PSICQUIC (PSI Common QUery InterfaCe [50] access 27 databases (see <http://www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml> for a complete list) and indicates from which database information for a specific input protein was retrieved. BioCyc (<http://websvc.biocyc.org/>) is a pathway specific database (contains 1962 pathways); however, in general, it is more strongly geared towards integrating genomic and cellular network information and will not be discussed in further detail here [51]. NCBI Entrez retrieves interactions from the Gene section of NCBI, which, in turn, collects entries from multiple databases (BOND [27], BioGRID [13], EcoCyc [52], HPRD [53], KEGG [54] and REACTOME [31]). Entering a Gene ID into the Cytoscape Entrez search box retrieves interactions that are organism specific, while entering a gene symbol is likely to retrieve interactions for multiple organisms (not recommended; very indiscriminate). More extensive refinement of the search parameters is not possible.

Pathway Commons previously described in the “Pathway Databases” section, is the default selection in the Cytoscape ‘Data Source’ menu. It is our preferred option, and offers multiple filters to refine search criteria: it can retrieve either whole pathways or gene-centered interactions, which can be further filtered by database source or interaction type (participation in a complex or binary PPI, or biochemical reaction). In the imported network, entries are connected by color-coded edges, and an additional panel, ‘Node details,’ appears, providing live links to Uniprot, Enrez Gene and iHOP. Alas, once the session was stored and is reopened at a later point, all the links are rendered non-functional. The biggest nuisance is, however, the way the Pathway Commons handles protein complexes: if any one component of the complex is interacting with the target protein, the whole complex is imported. This results in unexpected bloating of the network, which is hard to anticipate and even harder to control. For example, if AURKA (aurora kinase A) is used as input, Pathway Commons initially identifies 123 interactors; yet the generated network including 700 nodes, 555 of which are part of a large complex.

iRefScape is generally geared towards advanced users. This Cytoscape plugin allows visualization and data mining of protein interaction databases from the interaction Reference Index (iRefIndex) [25]. As a key difference with this resource versus many of the other databases mentioned in this chapter, in order to run iRefScape, the entire database must be downloaded, with searches run on the downloaded dataset (*see* Note 14). iRefIndex consolidates protein interaction data from ten individual databases, which include: BIND [27], BioGRID [20], CORUM [55], DIP [56], HPRD [53], IntAct [26], MINT [28], MPact [57], MPPI [58] and I2D [59]. iRefIndex version 8.0 covers 480,368 distinct interactions involving 86,757 unique protein interactors (*see* Note 15).

We used ‘SMAD1’ again as the input protein to evaluate iRefScape; only first order neighbors – proteins that directly interact with SMAD1 – were provided, as stipulated by the

Note 14^BBecause data retrieval with iRefScape is performed on a local computer (researcher would have to download the database once), searches tend to be faster than web-based sources and without the possibility of database downtime. No weights or confidence score are assigned to the interactions in the iRefScape-generated network, which could be considered a slight disadvantage compared to STRING and GeneMANIA.

Note 15^BThe amino acid sequence of each interactor and the taxonomy identifier are used by iRefIndex to assign a globally accessible hash-key to each distinct interactor and interaction.

default settings. To obtain interactions between first order neighbors the user must specifically confirm that this is how the search is to be executed. In order to sort out predicted and text mining results, the user can click on the panel of ‘advanced feature’ and perform the ‘filter’ function. By choosing the ‘type’ to be ‘edge’, ‘attribute’ to be ‘i.method_name’, ‘value’ to be ‘contains’, and ‘condition’ to either be ‘predicted’ or ‘text mining’, and clicking ‘set selected’, the selected group of interactions will be highlighted in the graph view area (Figure 18).

MiMI (Michigan Molecular Interactions) integrates and normalizes data from multiple sources using alignment of hierarchical data structures [60]. MiMI integrates BIND [27], BioGRID [20, 56], DIP [56], HPRD [53], IntAct [26], the datasets from Max Delbrueck Center [61] and data from the Center for Cancer Systems Biology at Harvard [62]; this data is supplemented with protein information from GO [5], InterPro [63], miBLAST [64], OrganelleDB [65], OrthoMCL [66], PFam [67], and ProtoNet [68]. In addition, MiMI annotates interactions using MeSH, OMIM and the Gene Ontology. MiMI tracks the provenance of data, allowing users to decide which data to trust and which to ignore. Users can pre-select specific sources as a search criterion to limit results to desired sources. The multitude of search parameters that can be manipulated in MiMI allows users to, for example, only search the nearest (first order) neighbors, interactions between neighbor proteins, or neighbors shared by more than one specified gene. To fully evaluate MiMI, we used SMAD1 again; the results can be seen in Figure 11 (see Note 16).

1.5.6 Insights from Model Organisms—Model organisms such as *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, and *Rattus norvegicus* have been extensively studied to determine functional PPIs. Most of these species have been used extensively for genetic studies, providing a critically important measure of functional interdependence of two proteins. Additionally, the genome sequences of the model organisms have been available for many years [69]. Therefore, a much higher percentage of proteins are functionally annotated in model organisms than in *H. sapiens*. Partially due to parsimony, many important physical and functional relationships are known to or are very likely to be conserved between lower and higher eukaryotes; thus, the introduction of interactions across multiple species can increase the diversity and novelty of nodes within a focused network.

Walhout [70] introduced the concept of “interologs”: orthologous pairs of interacting proteins in different organisms. It was proposed that if two proteins interact in one species, identifiable human orthologs are likely to interact as well. This concept has been proven valid in several key studies [71-73], which underlines the potential benefits of including interactions from multiple species in a customized network.

Essentially, there are two ways of extraction relevant information from the model organism data. The easiest one is to rely on metasearch engines. Both STRING and GeneMANIA are

Note 16 The advanced metasearch programs accessible through Cytoscape all allow the investigator to accrue a large amount of data in a convenient format. Networks generated from the different databases can be merged, sub-networks can be selected and nodes can be rearranged. If Cytoscape has already been installed, the advanced metasearch programs may be the fastest and easiest option to get a general idea about the interactions associated with a specific protein. Furthermore, Cytoscape allows a greater degree of network-customization (node arrangement, edge colors, sub-networks, network merging, etc.) than, for example, STRING and GeneMANIA. Cytoscape also makes it easy to save information for later use, which is quite convenient. However, we do not recommend performing multiple metasearches using different plugins within the same Cytoscape session, if it is important to the investigator to have more control over the data included in a given network. Importing lots of data in one swoop can prompt several suboptimal conditions: it will create multiple networks which cannot be simply merged because different protein identifiers are used; and it will create visually unwieldy node attributes table, with multiple (and mostly empty) attributes for the same protein. We thus recommend that the advanced metasearch options are either used one per Cytoscape session, with the subsequent ID conversion, or to augment other networks, once the retrieved data has been manually checked (Figure 1).

convenient sources for the retrieval of model organism focused data. The convenience does come at the cost of somewhat lower coverage; therefore, researchers wanting to maximize information inclusion should manually find homologs in other species and retrieve information from the databases devoted to individual organisms. The most prominent dedicated model organism databases are listed in this section (see Note 17).

DroID: *Drosophila* Interactions Database (<http://www.droidb.org>) is currently the most comprehensive and frequently updated interaction database dedicated to *Drosophila*, covering 514,325 unique interactions and 15,254 genes [74, 75]. The interactions covered by DroID include physical protein interactions, genetic interactions and predicted interactions based on data from other model organisms (e.g. yeast, worm, zebrafish, frog, mouse and others) and humans. DroID can be searched using text input (e.g., gene symbol, gene name, molecular function, biological process, cellular component, protein domain, phenotype and gene expression) and the information output can be stipulated in several ways (e.g., experiment type, predictions, genetic interactions, transcription factor-gene interaction and miRNA-gene interaction) [75]. Interaction annotations or gene attributes can be accessed via a provided hyperlink to FlyBase, [76], a dedicated *Drosophila* database that can also be directly accessed at <http://flybase.org/>. Gene expression data are also provided.

All interactions accessible through DroID can be explored and downloaded either through a text based web site, through IM (Interaction Map) Browser (a web based Java application with visualization capabilities) [77], or by using a Cytoscape plugin. DroID's search function is quite powerful and can recognize many different inputs and even accepts ortholog names from other model organisms. All details, which include different forms of IDs, genomic locations, a general summary, recent updates, mapping, detailed expression data, protein-protein interactions, pertaining to a given text input are provided with the appropriate references. The IM browser is also very convenient to use and allows users to select different input formats and highlight various interaction types (experimental, predicted, etc.). The different interaction types can then easily be filtered or can be assigned different colors within the network. The functions of IM browser are quite simple and user-friendly and provide a quick way for network visualization.

SGD: (*Saccharomyces* Genome Database; <http://www.yeastgenome.org>) provides access to manually curated information on *Saccharomyces cerevisiae* from peer-reviewed literature, combined with high throughput datasets [78]. A gene name, systematic ORF name or a SGDID can be used as the query input, and pertinent interactions, as well as a summary of the gene, location, related literature, gene ontology and more is subsequently displayed. In the interaction section, the reaction type, sources and references are all listed. *S. cerevisiae* is one of the mostly widely utilized organisms in the fields of biochemistry, molecular biology, cell biology and systems biology, because it can be grown and manipulated with relative ease. Interactions between yeast genes and/or proteins are established by an impressive array of techniques that evaluate physical and genetic interactions, co-occurrence in complexes, co-expression profiles and chemogenomic similarities. Furthermore, many of *S. cerevisiae*'s

Note 17 Two general words of caution: First, finding orthologous genes in other species is not straightforward. Using different theoretical approaches, some of which are still hotly debated, may produce different results. Lists produced by different online converters (general resources like bioDBnet, <http://www.ensembl.org>, <http://www.ncbi.nlm.nih.gov/homologene>, or specialized ones like <http://flight.icr.ac.uk/>) should be cross-compared for reliability.

Second, many of the specialized databases also have an option to transfer interactions from other model organisms. We strongly recommend that this option NOT be checked; otherwise, a deceptive feedback loop is created. For example, if a tentative interaction between human proteins is suggested, additional support for such an interaction can be generated from data on interactions between fly orthologs of human genes. However, the database entries on interactions between *Drosophila* proteins, in this example, only exist because the checked option "transfer data from other model organisms" enabled the *Drosophila* database to import the very same tentative human interaction the investigator is trying to find stronger evidence for.

genes and pathways are conserved in higher organisms, presenting the opportunity to use data on *S. cerevisiae* to gain insights into gene interactions in less well studied organisms. BioGrid, which is the database for all the protein-protein and genetic interactions retrievable from SGD, is one of the most popular databases in the ‘yeast community’.

WormBase: (<http://www.wormbase.org>) is one of the central data repositories for nematode biology and provides information of the genetics, genomics and biology of *Caenorhabditis elegans* and sixteen related nematodes [79, 80]. *C. elegans* is a well-established genetic model system because it offers several advantages to other species, such as: small size, short generation time (3.5 days), compact genome (100 MB) and invariant cell lineage. It has been and is being used extensively in fields ranging from development biology to neurobiology, including single gene to whole genome studies. Because of the plethora of data on *C. elegans* it is an extremely useful model organism to predict PPIs in humans. WormBase has, like most search-platforms, a text input box, which can be used to enter any gene/protein of interest. First, several hyperlinked results are presented, which will guide the user to the main results section, where information on the input gene/protein is provided. The information ranges from expression, to phenotype, to gene ontology, to human diseases as well as interactions. Clicking ‘Interactions’ on the left most menu bar moves into view a generated network and/or table with annotations for each interaction (a link to GeneMANIA is also provided).

MIPS: (Mammalian Protein-Protein Interaction Database; <http://mips.helmholtz-muenchen.de/proj/ppi/>) [58] is a simple, user-friendly website to query mouse genes. *Mus musculus* is by far the most commonly used laboratory mammal studied, because of the ease of maintenance and handling, the high degree of homology with humans and the high reproduction rate. From the MIPS opening page, clicking on ‘PPI search form’ leads to a new page for the MPPI search text input. Text input can be any combination of one protein, two proteins and/or an interaction. The results are given as a table of paired proteins (interactors) with the option to expand the section to include details on evidence, interaction site and function (see Note 18).

MouseNet: (<http://mousenet.princeton.edu/>) is a genome-wide functional network based on data from laboratory mice [81]. It allows users to predict functional assignments and network components. The network includes probabilistic functional linkages and over 20,000 protein-coding genes that can be accessed through an interactive web-based interface. The queried network is ranked by assigned confidence scores and can be exported and saved without much hassle.

The general takeaway message from this section is the augmentative benefit of many of the available model organism based databases and which databases are particularly useful. Model organisms have been used extensively in experimentation and in some regard the wealth of information available for some of them easily exceed the information available for *H. sapiens*. Also, considering the fact that many genes are highly conserved; the benefit of utilizing the information available for model organisms to strength the informative capacity of a network should be obvious. This may be particularly useful if an investigator is focusing on novel interactions not yet fully established or not previously observed in *H. sapiens*.

Note 18 As a caveat, although the information provided by MIPS may be quite useful; we found that some of the steps are, though not complicated, not as intuitive as they are for some of the other databases, which can be a source of frustration.

2. Materials

Software Platforms

Cytoscape (<http://www.cytoscape.org/>)

VisANT (<http://visant.bu.edu>)

BioLayout *Express*^{3D} (<http://www.biolayout.org/>)

Online Metasearches

Free—STRING (<http://string-db.org/>)

GeneMANIA (<http://www.genemania.org/>)

Not Free—Ingenuity Pathway Analysis (<http://www.ingenuity.com/>)

GeneGo (<http://www.genego.com/>)

Protein Lounge (<http://www.proteinlounge.com/>)

Gene/Protein ID Conversion

Clone/Gene ID Converter (<http://idconverter.bioinfo.cnio.es/>)

bioDBnet (<http://biodbnet.abcc.ncifcrf.gov/>)

Protein-Protein Interaction Databases

BioGrid (www.thebiogrid.com)

BOND (<http://bond.unleashedinformatics.com/>).

IntAct (<http://www.ebi.ac.uk/intact/>).

MINT (<http://mint.bio.uniroma2.it/mint/>)

Pathway Databases

KEGG (<http://www.genome.jp/kegg/>)

Pathway Commons (<http://www.pathwaycommons.org/pc/>)

REACTOME (<http://www.reactome.org/ReactomeGWT/entrypoint.html>)

PharmGKB (<http://www.pharmgkb.org/>)

BioCarta (<http://www.biocarta.com/>)

Netpath (<http://www.netpath.org/>)

PID (<http://pid.nci.nih.gov/>)

Wikipathways (<http://www.wikipathways.org/index.php/WikiPathways>)

Microarray Databases

GEO (<http://www.ncbi.nlm.nih.gov/geo>)

ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>).

Drug-Protein Interactions

The DrugBank database (<http://www.drugbank.ca/>)

STITCH (<http://stitch.embl.de/>)

Text Mining

PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>)

Chilibot (<http://www.chilibot.net/>)

Google Scholar (<http://scholar.google.com/>)

iHop (<http://www.ihop-net.org/UniPub/iHOP/>)

Microsoft Academic Search (<http://academic.research.microsoft.com/>)

BITOLA (<http://ibmi3.mf.uni-lj.si/bitola/>)

Advanced Metasearches

PSICQUIC (<http://www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml>)

BioCyc (<http://biocyc.org/>)

iRefScape (<http://irefindex.uio.no/wiki/iRefIndex>)

MiMI (<http://mimi.ncibi.org/MimiWeb/main-page.jsp>)

Insights from Model Organisms

DroID (<http://www.droidb.org/>)

SGD (<http://www.yeastgenome.org/>)

WormBase (<http://www.wormbase.org>)

MIPS (<http://mips.helmholtz-muenchen.de/proj/ppi/>)

MouseNet (<http://mousenet.princeton.edu/>)

General

NCBI (<http://www.ncbi.nlm.nih.gov/>)

ENTREZ (<http://www.ncbi.nlm.nih.gov/gquery>)

UniProt (<http://www.uniprot.org/>)

GeneCards (<http://www.genecards.org/>)

3. Methods

3.1 Network Assembly and Analysis

In order to highlight the differences regarding network assembly and analysis between individual databases and metasearch platforms, we used several of the sources mentioned above to assemble datasets, build networks and then analyze and compare the results. Based on information provided throughout this chapter and the network assembly instructions given in this section, users should be able to quickly and reliably determine which approaches are most suitable for their purposes and how to execute the necessary steps (Figures 1 & 10).

For a first composite network for comparisons, we only included data obtained experimentally (no predictions were included) and exclusively retrieved using metasearch platforms. The purpose of this example is to illustrate the difference between data retrieved via metasearch platforms and data directly retrieved from some of the individual databases the metasearch platforms access. Before we assembled any composite network, we standardized all the data retrieved as described in 1.3 Gene/Protein ID Conversion.

SMAD1 was used as the input protein and the following results were obtained for our metasearch network: GeneMANIA identified a network of 101 interactors; STRING identified 77 interactors; MiMI identified 153 interactors; iRefScape identified 156 interactors; and Cytoscape's built-in Pathway Commons database search identified 149 interactors. After using Excel to sort our gene list, we ended up with a composite list of 213 unique interactors (Figure 11).

To compare results from the metasearch platforms mentioned in the previous paragraph, we directly accessed four unique databases that are also directly accessed by the metasearch programs and then compared the results. The following was observed: all 7 interactors from IntAct were also in the metasearch list; out of 104 interactors retrieved directly from BOND, 100 had also been identified by the metasearch platforms; 51 out of 53 interactors retrieved from MINT were also included in our metasearch composite list; and for BioGrid, 58 interactors were identified by the metasearch platform out of 68 directly retrieved interactors. In conclusion, for SMAD1, 96% of interactors that had been retrieved by directly accessing individual databases were also identified by the combined five-metasearch methods. Therefore, combined, the metasearch methods can provide a dataset nearly as complete as a dataset based solely on data directly retrieved from individual databases (Table 1).

Our next network model demonstrates the overall coverage of four PPI databases by each of the above-mentioned metasearch platforms – which were evaluated on an individual basis. We first integrated SMAD1 associated interactors identified by directly accessing the four dedicated databases; this approach led to a list of 181 interactors (Table 1). We then compared each of the metasearch results to this list. iRefScape and Pathway Commons searches covered 76% and 73% of PPIs directly retrieved from dedicated databases respectively. MiMI was found to cover 71%, and GeneMANIA and STRING covered 48% and 27%, respectively. Based on these results, which may vary slightly depending on the input protein, we feel comfortable making one of two recommendations to a user who is interested in a quick and easy way of capturing data: 1) If Cytoscape is available, iRefScape or Pathway Commons (Cytoscape plugins) are the best options for a one-stop search; 2) If Cytoscape is not an option, GeneMANIA, which is web-based, is a reasonable alternative.

If a user would also like to use one of the individual databases directly, we recommend using Pathway Commons together with BOND, which in our case covered 90% of all the

identified interactions. Figure 9 approximates the percentage of possible interactors identified (those that could be retrieved by directly accessing individual databases) covered by combinations of sources, irrespective of which specific sources are actually used (Table 1).

Lastly, we compared the results included in our model networks to the commercially available Ingenuity (IPA®) database. IPA identified 203 interactors for SMAD1. It covered 65% of the interactors from the merged list of the four previously mentioned direct PPI searches (Table 1). The results provided similar coverage as the other metasearch platforms did; the additional unique interactors found by Ingenuity were retrieved from databases other than the four we evaluated. IPA may provide additional services and features; however, we did not find that IPA outperforms the free metasearch platforms as far as data retrieval is concerned.

Figure 1 outlines how different tools and databases can be assembled in order to serve a specific purpose. The following step-by-step protocol lays-out a generic approach (various databases introduced throughout this chapter can be substituted for each other) that leads to a fairly in-depth network:

1. Installation of visualization software
 - a. Install Cytoscape (<http://www.cytoscape.org/>)
 - b. Install iRefScape and MiMI (use Cytoscape's 'Manage Plugins' option)
2. Query your gene of interest in STRING (<http://string-db.org/>)
 - a. Once the default network is generated, uncheck everything except the appropriate 'Active Prediction Methods', select 'required confidence (score)' and the number of 'interactors shown' (Figure 5)
 - b. Click 'Update Parameters'
 - c. Click 'Save' (right below the network) and select the 'Network proteins description' .txt file
 - d. Copy and paste the list into Excel (or open directly)
 - i. If necessary, split contents into columns (Data -> Text to Columns)
 - ii. The first column will have gene names (symbols)
 - iii. Save the Excel file
3. Query MiMI (<http://mimi.ncibi.org/>) using your gene of interest (do not use the Cytoscape plugin for this step)
 - a. Find the 'Download full table as:' line (below results table, once a given results section has been expanded) and click on the Excel icon
 - b. Column B (Gene2) contains the gene names (symbols) of the interactors for the gene of interest.
4. With both Excel files (from STRING and MiMI) open:
 - a. Copy and paste both lists into one spreadsheet
 - b. Compare lists of interactors retrieved from the two sources ('sort' and 'vlookup' functions are very useful for this; it is also helpful to change the text or cell color for one of the two lists, in order to identify which

database each individual gene came from once the lists have been merged and sorted)

- c. Decide if the overlap (number of genes identified by both sources) is satisfactory (depends on the goals of your project)
5. If the degree of overlap is not satisfactory, query additional sources
 - a. E.g., from within Cytoscape use Pathway Commons (File -> Import -> Network from Web Services...) and/or download the iRefScape plugin (use 'Manage Plugins')
 - b. Exporting gene symbols
 - a. Highlight all the nodes in the
 - b. Information for all the nodes should be displayed in the 'Data Panel' (if not, click on the first icon of the 'Data Panel' and select more attributes)
 - c. Right-click on any node listed in the 'Data Panel' -> Export -> Entire Table
 - d. Save Table
 - e. Copy and paste gene list into Excel
 - c. If even more exhaustive information desired, query individual PPI databases (Figure 1), saving and adding the gene lists into the composite Excel spreadsheet
 - a. If needed, export the composite Excel list into bioDBNet (<http://biodbnet.abcc.ncifcrf.gov/db/db2db.php>) and convert all entries to gene symbols
6. Once satisfied with the overlap and/or coverage, create a new Excel file that is going to be import back into Cytoscape. Column A should be the names of all the retrieved interactors; Column B should be filled all the way down with the name of the gene used for the queries.
7. Import this file into the Cytoscape using: File -> Import -> Network from Table (Text/MS Excel)...
 - a. Use the two columns in Excel to specify interactors
 - i. E.g., 'Source Interaction' -> Column 1 and 'Target Interaction' -> Column 2
 - b. For the 'Interaction Type', leave "Default Interaction"
 - c. Under 'Advanced'
 - i. Check 'Show Text File Import Options'
 1. Make sure that 'Network Import Options Default Interaction' is pp or ppi (for protein interaction)
 - d. Click on Import
8. Select another type of interaction (e.g., pathways)
 - a. Go through steps 2-7, choosing the databases accordingly (Figure 1)

- b. For step 7c, change the name of the default interaction (“pathways”, “co-expression”, “disease” or any other interaction description)
 9. Once all the desired types of information have been collected and imported into Cytoscape, one network per interaction type (“pathway”, “co-expression”, etc.) should be open.
 - a. Merge all networks in Cytoscape using: Plugins -> Advanced Network Merge. Some proteins may now have multiple edges connecting them to the target protein – one edge for each interaction type.
 - b. VizMapper (View -> Open VizMapper; or click on the menu icon) can be used to color the edges based on interaction type.
 - c. The merged network will still have a “star” configuration, with the target protein at the center and all interactors surrounding it.
 10. If interactions between all the proteins are desired, the STRING web-platform or the MiMI plugin can be used.
 - a. Use the list generated in step 9a as input
 - b. STRING
 - i. Highlight all the nodes in the network from 9a (default name is ‘Union’)
 - ii. Information for all the nodes should be displayed on the ‘Data Panel’ (if not, click on the first icon of the ‘Data Panel’ and select more attributes)
 - iii. Right-click on any node listed in the ‘Data Panel’ -> Export -> Entire Table
 - iv. Save Table
 - v. Follow steps 2 and 6 to import the final network back into Cytoscape
 - c. MiMI plugin
 - i. Generate a list of genes for the network from step 9a (follow steps 10b(i-iv))
 - ii. Copy and paste the list into the MiMI query box
 - iii. Click ‘Search’
 11. Use the customized-network to generate hypotheses, to better understand the role of the target protein and to help interpret in-house experimental data.

The assembled network, be it one generated following the cited protocol or a more customized approach following Figure 1, retains much useful information. For one, the network stores all the identified interactions and attributes in an organized fashion. More importantly, the assembled network offers insight into the role a specific protein or a group of proteins play(s) in the molecular landscape. The ability to visualize data and appreciate the interplay among proteins can yield crucial information. Not only are all the interactors of a given protein identified (as far as supported by available data), the assembled network also helps the investigator appreciate the interactome as a whole and perhaps prompts realization of previously neglected intricacies.

Having a solid understanding of the interactions of proteins can enable investigators design experiments focused on associations highlighted by the generated network. Additionally, previous experimental observations may be supported by the data on display. In the case of SMAD1, the assembled network emphasized that SMAD1 is not simply part of the bone morphogenetic protein-signaling pathway. The network highlighted that SMAD1 plays an important role in immunological processes (via STAT3); furthermore, our network underlined the fact that SMAD1 not only interacts with SMAD5 and SMAD8, but also SMAD2, which is known for its role in TGF-beta signaling [82]. These two findings regarding SMAD1 would not have been as obvious and may have been missed without the construction of an interaction-based network. Additionally, the network presents the opportunity to nominate protein-partners that can be explored in the context of newly hypothesized biological functions.

Acknowledgments

The authors were supported by U54 CA149147, R01 CA63366 and P50 CA083638 from the NIH (to EAG), postdoctoral fellowship from SASS Foundation for Medical Research and Ann Schreiber Program of Excellence Grant from the Ovarian Cancer Research Fund (to HL), Drexel University College of Medicine MD-PhD Program (to TNB) and NIH core grant CA06927 (to Fox Chase Cancer Center).

5. References

1. Szklarczyk D, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 2011; 39(Database issue):D561–8. [PubMed: 21045058]
2. Mostafavi S, et al. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* 2008; 9(Suppl 1):S4. [PubMed: 18613948]
3. Dempsey K, et al. Functional identification in correlation networks using gene ontology edge annotation. *Int J Comput Biol Drug Des.* 2012; 5(3-4):222–44. [PubMed: 23013651]
4. Smoot ME, et al. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics.* 2011; 27(3):431–2. [PubMed: 21149340]
5. Hu Z, et al. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.* 2009; 37(Web Server issue):W115–21. [PubMed: 19465394]
6. Theocharidis A, et al. Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nat Protoc.* 2009; 4(10):1535–50. [PubMed: 19798086]
7. Mellor JC, et al. Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.* 2002; 30(1):306–9. [PubMed: 11752322]
8. Cline MS, et al. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc.* 2007; 2(10):2366–82. [PubMed: 17947979]
9. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003; 13(11):2498–504. [PubMed: 14597658]
10. Lehne B, Schlitt T. Protein-protein interaction databases: keeping up with growing interactomes. *Hum Genomics.* 2009; 3(3):291–7. [PubMed: 19403463]
11. Snel B, et al. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* 2000; 28(18):3442–4. [PubMed: 10982861]
12. Pena-Castillo L, et al. A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol.* 2008; 9(Suppl 1):S2. [PubMed: 18613946]
13. Warde-Farley D, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 2010; 38(Web Server issue):W214–20. [PubMed: 20576703]
14. Apweiler R, et al. A large-scale protein-function database. *Nat Chem Biol.* 2010; 6(11):785. [PubMed: 20956966]
15. Kiefer F, et al. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.* 2009; 37(Database issue):D387–92. [PubMed: 18931379]

16. Letunic I, Doerks T, Bork P. SMART 6: recent updates and new developments. *Nucleic Acids Res.* 2009; 37(Database issue):D229–32. [PubMed: 18978020]
17. Heldin C-H, Miyazono K, ten Dijke P. TGF- β signalling from cell membrane to nucleus through SMAD proteins. *Nature.* 1997; 390(6659):465–471. [PubMed: 9393997]
18. Montojo J, et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics.* 2010; 26(22):2927–8. [PubMed: 20926419]
19. Barrett T, et al. NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res.* 2011; 39(Database issue):D1005–10. [PubMed: 21097893]
20. Stark C, et al. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.* 2011; 39(Database issue):D698–704. [PubMed: 21071413]
21. Cerami EG, et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* 2011; 39(Database issue):D685–90. [PubMed: 21071392]
22. Brown KR, Jurisica I. Online predicted human interaction database. *Bioinformatics.* 2005; 21(9): 2076–82. [PubMed: 15657099]
23. Alibes A, et al. IDconverter and IDClight: conversion and annotation of gene and protein IDs. *BMC Bioinformatics.* 2007; 8:9. [PubMed: 17214880]
24. Mudunuri U, et al. bioDBnet: the biological database network. *Bioinformatics.* 2009; 25(4):555–6. [PubMed: 19129209]
25. Razick S, et al. iRefScape. A Cytoscape plug-in for visualization and data mining of protein interaction data from iRefIndex. *BMC Bioinformatics.* 2011; 12:388. [PubMed: 21975162]
26. Kerrien S, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 2012; 40(Database issue):D841–6. [PubMed: 22121220]
27. Alfano C, et al. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.* 2005; 33(Database issue):D418–24. [PubMed: 15608229]
28. Licata L, et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 2012; 40(Database issue):D857–61. [PubMed: 22096227]
29. Astsaturov I, et al. Synthetic lethal screen of an EGFR-centered network to improve targeted therapies. *Sci Signal.* 2010; 3(140):ra67. [PubMed: 20858866]
30. Kanehisa M, et al. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 2004; 32(Database issue):D277–80. [PubMed: 14681412]
31. Croft D, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 2011; 39(Database issue):D691–7. [PubMed: 21067998]
32. Joshi-Tope G, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 2005; 33(Database issue):D428–32. [PubMed: 15608231]
33. Hernandez-Boussard T, et al. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.* 2008; 36(Database issue):D913–8. [PubMed: 18032438]
34. Kandasamy K, et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol.* 2010; 11(1):R3. [PubMed: 20067622]
35. Schaefer CF, et al. PID: the Pathway Interaction Database. *Nucleic Acids Res.* 2009; 37(Database issue):D674–9. [PubMed: 18832364]
36. Kelder T, et al. WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* 2012; 40(Database issue):D1301–7. [PubMed: 22096230]
37. Barrett T, et al. NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res.* 2007; 35(Database issue):D760–5. [PubMed: 17099226]
38. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002; 30(1):207–10. [PubMed: 11752295]
39. Brazma A, et al. ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 2003; 31(1):68–71. [PubMed: 12519949]
40. Parkinson H, et al. ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* 2007; 35(Database issue):D747–50. [PubMed: 17132828]

41. Parkinson H, et al. ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* 2011; 39(Database issue):D1002–4. [PubMed: 21071405]
42. Hibbs MA, et al. Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics.* 2007; 23(20):2692–9. [PubMed: 17724061]
43. Lamb J, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science.* 2006; 313(5795):1929–35. [PubMed: 17008526]
44. Yildirim MA, et al. Drug-target network. *Nat Biotechnol.* 2007; 25(10):1119–26. [PubMed: 17921997]
45. Kuhn M, et al. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.* 2008; 36(Database issue):D684–8. [PubMed: 18084021]
46. Wang Y, et al. PubChem's BioAssay Database. *Nucleic Acids Res.* 2012; 40(Database issue):D400–12. [PubMed: 22140110]
47. Cohen KB, Hunter L. Getting Started in Text Mining. *PLoS Comput Biol.* 2008; 4(1):e20. [PubMed: 18225946]
48. Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics.* 2005; 21(Suppl 2):ii252–8. [PubMed: 16204114]
49. Plotnikova OV, et al. Calmodulin activation of Aurora-A kinase (AURKA) is required during ciliary disassembly and in mitosis. *Mol Biol Cell.* 2012; 23(14):2658–70. [PubMed: 22621899]
50. Orchard S. Molecular interaction databases. *Proteomics.* 2012; 12(10):1656–62. [PubMed: 22611057]
51. Latendresse M, Paley S, Karp PD. Browsing metabolic and regulatory networks with BioCyc. *Methods Mol Biol.* 2012; 804:197–216. [PubMed: 22144155]
52. Keseler IM, et al. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res.* 2011; 39(Database issue):D583–90. [PubMed: 21097882]
53. Mathivanan S, et al. Human Proteinpedia enables sharing of human protein data. *Nat Biotechnol.* 2008; 26(2):164–7. [PubMed: 18259167]
54. Kanehisa M, et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 2006; 34(Database issue):D354–7. [PubMed: 16381885]
55. Ruepp A, et al. CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Res.* 2010; 38(Database issue):D497–501. [PubMed: 19884131]
56. Salwinski L, et al. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 2004; 32(Database issue):D449–51. [PubMed: 14681454]
57. Guldener U, et al. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.* 2006; 34(Database issue):D436–41. [PubMed: 16381906]
58. Pagel P, et al. The MIPS mammalian protein-protein interaction database. *Bioinformatics.* 2005; 21(6):832–4. [PubMed: 15531608]
59. Brown KR, Jurisica I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.* 2007; 8(5):R95. [PubMed: 17535438]
60. Jayapandian M, et al. Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. *Nucleic Acids Res.* 2007; 35(Database issue):D566–71. [PubMed: 17130145]
61. Stelzl U, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell.* 2005; 122(6):957–68. [PubMed: 16169070]
62. Han JD, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature.* 2004; 430(6995):88–93. [PubMed: 15190252]
63. Hunter S, et al. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 2012; 40(Database issue):D306–12. [PubMed: 22096229]
64. Kim YJ, et al. miBLAST: scalable evaluation of a batch of nucleotide sequence queries with BLAST. *Nucleic Acids Res.* 2005; 33(13):4335–44. [PubMed: 16061938]
65. Wiwatwattana N, et al. Organelle DB: an updated resource of eukaryotic protein localization and function. *Nucleic Acids Res.* 2007; 35(Database issue):D810–4. [PubMed: 17130152]
66. Fischer S, et al. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics.* 2011 Chapter 6:Unit 6 12 1-19.

67. Punta M, et al. The Pfam protein families database. *Nucleic Acids Res.* 2012; 40(Database issue):D290–301. [PubMed: 22127870]
68. Rappoport N, et al. ProtoNet 6.0: organizing 10 million protein sequences in a compact hierarchical family tree. *Nucleic Acids Res.* 2012; 40(Database issue):D313–20. [PubMed: 22121228]
69. Adams MD, et al. The genome sequence of *Drosophila melanogaster*. *Science.* 2000; 287(5461): 2185–95. [PubMed: 10731132]
70. Walhout AJ, Vidal M. Protein interaction maps for model organisms. *Nat Rev Mol Cell Biol.* 2001; 2(1):55–62. [PubMed: 11413466]
71. Echeverria PC, et al. An interaction network predicted from public data as a discovery tool: application to the Hsp90 molecular chaperone machine. *PLoS One.* 2011; 6(10):e26044. [PubMed: 22022502]
72. Sharan R, et al. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J Comput Biol.* 2005; 12(6):835–46. [PubMed: 16108720]
73. Ulitsky I, Shamir R. Pathway redundancy and protein essentiality revealed in the *Saccharomyces cerevisiae* interaction networks. *Mol Syst Biol.* 2007; 3:104. [PubMed: 17437029]
74. Murali T, et al. DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Res.* 2011; 39(Database issue):D736–43. [PubMed: 21036869]
75. Yu J, et al. DroID: the *Drosophila* Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics.* 2008; 9:461. [PubMed: 18840285]
76. McQuilton P, St Pierre SE, Thurmond J. FlyBase 101--the basics of navigating FlyBase. *Nucleic Acids Res.* 2012; 40(Database issue):D706–14. [PubMed: 22127867]
77. Pacifico S, et al. A database and tool, IM Browser, for exploring and integrating emerging gene and protein interaction data for *Drosophila*. *BMC Bioinformatics.* 2006; 7:195. [PubMed: 16603075]
78. Cherry JM, et al. *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 2012; 40(Database issue):D700–5. [PubMed: 22110037]
79. Stein L, et al. WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* 2001; 29(1):82–6. [PubMed: 11125056]
80. Harris TW, et al. WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* 2010; 38(Database issue):D463–7. [PubMed: 19910365]
81. Guan Y, et al. A genomewide functional network for the laboratory mouse. *PLoS Comput Biol.* 2008; 4(9):e1000165. [PubMed: 18818725]
82. ten Dijke P, Arthur HM. Extracellular control of TGFbeta signalling in vascular development and disease. *Nat Rev Mol Cell Biol.* 2007; 8(11):857–69. [PubMed: 17895899]

Figure 1.

Flow Chart for Building a Signaling Network. The flow chart initially diverges to indicate three possible approaches to network building (2.). Next, the chart lists different databases or types of databases that are great sources for information mining (3. and 4.). The bottom half of the chart maps out steps necessary to process retrieved data (5., 6. and 7.) and reach various outcomes (8.).

Figure 2.
Number of citations per listed software platform (as of October, 2012).

Figure 3. Advanced metasearch based networks generated using Cytoscape. (A) MiMI (plugin); (B) iRefScape (plugin); and (C) Pathway Commons (build-in Cytoscape feature).

Combined PPI databases : 181 unique interactors

Figure 4.

Comparison of the retrieved results from four different unique databases (BioGrid, MINT, IntAct and BOND) to the results retrieved using several metasearch platforms (See Table 1 for details). 'SMAD1' was used as the input gene for all cases. This example highlights the fact that solely relying on metasearch platforms will not retrieve an exhaustive list of interactors. The reason why the metasearch platforms appear to have retrieved many more interactors than the unique databases is because several unique databases not shown in this figure are also accessed by the metasearch platforms (such as HPRD, OPHID, InnateDB and CORUM).

Figure 5.

The network in this figure was generated using STRING ('SMAD1' was used as the input gene and the default settings were applied). Differently colored 'edges' indicate the type of evidence supporting each interaction: cyan = curated STRING database; light green = text mining; purple = experiments; and lavender = homology (the different types of evidence are not mutually exclusive). The 'Info & Parameters' panel shows the different options available to filter a generated network, using criteria such as: co-expression, neighborhood, gene fusion, co-occurrence (in PubMed abstracts), experiments, databases, textmining, required confidence and number of interactors (See <http://string-db.org/> for more details).

Figure 6. GeneMANIA results for 'SMAD1'. **(A)** GeneMANIA's 'advanced options' menu provides details about a given network, including links to specific data sources, as well as the possibility to quickly and efficiently alter several parameters of a generated network. **(B)** The illustrated network was altered using the provided side menu **(C)**, to only include 'Physical interactions' (nodes were manually rearranged to obtain the pattern shown).

Figure 7. Ingenuity (IPA) generated network. This figure is an illustration of a ‘SMAD1’ centered network only using ‘Ingenuity Expert Information’ (a randomly selected subset of only 10 retrieved genes is displayed for clarity). Information about each individual node can be visualized by moving the mouse cursor over the node of interest (in this example NR3C1). The ‘Relationship Summary’ window can be summoned by clicking on the ‘edge’ connection two nodes (in this example the line between ACVR1 and SMAD1). (Colors seen in this network were added manually).

Figure 8. bioDBnet ID conversion. **(A)** bioDBnet ID conversion allows conversions of single as well as multiple gene names. Numerous options for input and output formats are available (See <http://biodbnet.abcc.ncifcrf.gov/db/db2db.php> for details). Taxon ID (9606 is for *H. sapiens*) can be used to refine the output list. **(B)** Demonstration of a typical bioDBnet ID conversion output list. Red boxes highlight genes that could not be converted and genes with multiple assigned Gene IDs. In the case of genes that could not be converted (first red box), the best option is to copy and paste the Gene Symbol into <http://www.ncbi.nlm.nih.gov/> to search the NCBI database (often times unidentified genes are ‘hypothetical’ genes or proteins). In the case of multiple Gene IDs having been assigned to one Gene Symbol, it is easiest to click on each Gene ID and use the then provided information to determine which Gene ID is most accurate. Inaccurate Gene IDs and ‘hypothetical’ genes should be removed manually once the list of genes has been exported.

Figure 9.

Estimated range of specific interactions likely to be identified by each search approach (listed on horizontal axis) out of the total number of identifiable interactions. The dashed lines indicate the range of percentages covered by either a single database or various combinations of databases (the database(s) used as well as the specific input protein can greatly influence the percentage of identifiable interactions actually identified). The solid horizontal line is the most probable outcome irrespective of the specific database(s) used or of the input protein(s). Information retrieved from the databases listed in Table 1 was used to compose this figure.

Figure 10.

Approaches to combining data from several sources. The red, yellow and blue inner circles represent ‘core’ datasets; by this term we denote the most valuable information, which is context dependent. For example, the most valuable information could be based on: proteins directly linked to the network seed (input protein); proteins nominated by several (opposed to a single) pathway maps; or proteins with more prominent mRNA expression level changes. The outer circles (surrounding each colored circle) indicate “context” datasets: proteins more distantly connected to the seed (e.g., interactors of interactors) or proteins nominated by a single pathway map. The composite dataset represents a merged list (e.g., in Excel) of all the retrieved data. The ‘High-Confidence’ vetting approach only includes data points identified in every single source utilized. The ‘Mid-Confidence’ vetting approach includes all the data from ‘core’ datasets, plus the intersections between any two “context” datasets. The ‘Most Inclusive’ vetting approach includes every single data point retrieved. The purpose of a given project dictates which approach is most suitable.

Figure 11.

Protein-Protein Interaction databases compared to merged metasearch platforms. PSICQUIC View (<http://www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml>) and IMEx (<http://www.imexconsortium.org/>) are two metasearch platforms that access multiple databases at once (27 in the case of PSICQUIC and 11 in the case of IMEx). 'SMAD1' was used as the input entry for all the databases mentioned here. This figure highlights the potential advantage of accessing individual databases directly. '#N/A' = database not accessed; '0' = database accessed but no interactors identified. IN = Ingenuity; iRS = iRefScape; PC = Pathway Commons; BN = BOND^{OA}; GM = GeneMANIA; ST = STRING; BG = BioGRID; MIN = MINT; IA = IntAct.

Figure 12. Examples of PPI databases. ‘SMAD1’ was used to search for interactors. **(A)** MINT (PPI database) **(B)** ‘MINT Viewer’ (web-based visualization tool) **(C)** BioGrid (PPI database).

Figure 13.

Examples of datasets for SMAD1 retrieved from individual PPI databases and displayed using Cytoscape. SMAD1 first neighbors (direct interactors) retrieved from MINT, BOND, BIOGRID and IntAct are shown. The differently colored nodes indicate the overlap among PPI databases: red = interactor was identified in all four databases; blue = interactor was identified in three of the four databases; yellow = interactor was identified in two of the four databases; and pink = interactor was only identified in one database.

Figure 14. Pathway databases. (A) EGFR pathway map from BioCarta (http://www.biocarta.com/pathfiles/h_egfPathway.asp; permission to use this image was granted by BioCarta, LLC). (B) Overlap of pathway entities for the EGFR signaling pathway among pathway datasets retrieved from Netpath, BioCarta, Protein Lounge, STKE and Cell Designer (CellDes). 371 unique interactors were collectively identified using all five pathway databases. Out of these 371 interactors, only 103, 46, 32, 15 or 10 interactors were mentioned in at least two, three, four or all five databases, respectively. For investigators who are interested in only ‘High-Confidence’ datasets (Figure 10), the 10 interactors identified in all five databases mentioned in this example, would be of particular interest.

Figure 15. Pathway Commons (accessed through Cytoscape). **(A)** BMP (bone morphogenetic protein) pathway – ‘File’ -> ‘Import’ -> ‘Network from Web Services...’ -> use ‘Smad1_human’ as input -> select ‘BMP receptor signaling’ from generated list -> double click. **(B)** SMAD1-centered Sub-network – type ‘smad1_human’ into toolbar search box -> ‘Select’ -> ‘Nodes’ -> ‘First Neighbors of Selected Nodes’ -> click ‘Create new network from selected nodes, all edges’ icon on toolbar. (Some nodes were manually rearranged).

Figure 16. SMAD1 search in REACTOME. The best results were obtained when ‘pathways’, ‘reactions’, ‘proteins’ and ‘others’ were included as search options. Pathway maps are opened up by clicking on the hyperlink of a given search result (‘Reaction: The phspho-R-Smad1/5/8:Co-Smad transfers to the nucleus (*Homo sapiens*)’ was clicked on in this example). All entities of a selected pathway can be exported once the pathway map has been opened (See Section VI. Resources Beyond PPIs).

Figure 17. Chilibot text mining. Chilibot was used to search PubMed abstracts for specific relationships between NEDD9 and AURKA. Eleven interaction relationships (only eight are shown here) were identified, strongly suggesting direct interactions between NEDD9 and AURKA. This information could have been missed, had, for example, only GeneMANIA (was used to generate the network shown) been used (no direct interaction between NEDD9 and AURKA is indicated in the GeneMANIA network).

Figure 18.

SMAD1 search in iRefScape (Cytoscape plugin). **(A)** The iRefScape search panel allows users to choose species, load information from files, set advanced filters and adjust several other network parameters. **(B)** SMAD1 network generated after filtering out interactions from text mining and predicted interactions.

Table 1

SMAD1 search results from multiple databases and metasearch platforms. The discrepancies of nodes identified between metasearch platforms and dedicated databases are due to the fact that the metasearch platforms cover addition PPI databases besides the four used in his comparison (e.g., iRefScape retrieves many interactors and interactions from HPRD, OPHID, InnateDB and CORUM).

		Number of Nodes Identified	Number of Nodes also Identified in Metasearch	Number of Nodes also identified in PPI	Percentage
PPIDatabases	BioGrid	68	58		85%
	BOND	104	100		96%
	IntAct	7	7		100%
	MINT	53	51		96%
	Total Unique PPI Interactors	181	167		92%
Metasearch-free	GeneMania	101		87	48%
	STRING	77		49	27%
	MiMI	153		128	71%
	IRefScape	156		138	76%
	Pathway Commons	149		132	73%
	Total Unique Metasearch Interactors	213		173	96%
Metasearch-free	Ingenuity	203	118	65%	1

Table 2

Recommended unique protein-protein Interaction databases.

	BioGRID http://thebiogrid.org/	IntAct http://www.ebi.ac.uk/intact/	BondOA http://bond.unleashedinformatics.com/	MINT http://mint.bio.uniroma2.it/mint/	HPRO www.hprd.org/	I2D-Ophid http://ophid.utoronto.ca/ophidv2.201/index.jsp
	Highly Recommended	Highly Recommended	Highly Recommended	Recommended	Recommended	Recommended
General Comments	Data from high throughput and low throughput experiments compiled through comprehensive curation efforts	All Interactions are derived from literature curation or direct user submissions	Records are created for Interactions which have been shown experimentally and published In at least one peer-reviewed journal	Experimentally verified Interactions mined by expert curators	Expert curators read and Interpret each paper related to the molecule of Interest	Built by mapping high-throughput data (308,402 source Interactions and 386,847 predicted Interactions)
Graphical representation of results	No	Yes	No	Yes	No	No
Updates	Monthly	Monthly	Daily	Aug-11	Apr-10	Apr-10
Number of Interactions	527,569	304,667	188,517	241,264	39,194	681,404
Ease of Data Export	Cytoscape plug-In; data can be downloaded in different formats	Export data as Mi-Tab or XGMML (Cytoscape)	Drop-down lists to export data, Including files for Cytoscape	Cannot directly export data; have to manually copy and paste results; no plug-ins	Files available for download - must sign up; no plug-ins	Request required to download I2D data; results cannot be exported