Journal of proteome research

Tutorial

pubs.acs.org/jpr

Analyzing Protein−Protein Interaction Networks†

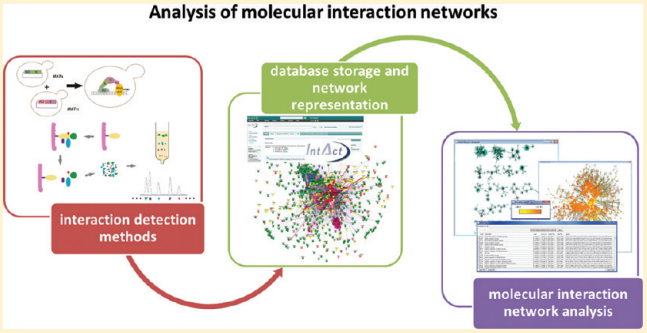
Gavin C. K. W. Koh,[‡,§,∥] Pablo Porras,[‡,∥] Bruno Aranda,[‡] Henning Hermjakob,[‡] and Sandra E. Orchard*[,‡]

[‡]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

[§]Department of Medicine, University of Cambridge, Cambridge, United Kingdom

Supporting Information

ABSTRACT: The advent of the "omics" era in biology research has brought new challenges and requires the development of novel strategies to answer previously intractable questions. Molecular interaction networks provide a framework to visualize cellular processes, but their complexity often makes their interpretation an overwhelming task. The inherently artificial nature of interaction detection methods and the incompleteness of currently available interaction maps call for a careful and well-informed utilization of this valuable data. In this tutorial, we aim to give an overview of the key aspects that any researcher needs to consider when working with molecular interaction data sets and we outline an example for interactome analysis. Using the molecular interaction database IntAct, the software platform Cytoscape, and its plugins BiNGO and clusterMaker, and taking as a starting point a list of proteins identified in a mass spectrometry-based proteomics experiment, we show how to build, visualize, and analyze a protein−protein interaction network.



KEYWORDS: *molecular interactions, network analysis, enrichment analysis, clustering, interaction databases, proteomics, network representation, database curation, high-throughput screening, HUPO tutorial program*

HISTORICAL BACKGROUND

The Interactome

The totality of molecular interactions that take place within a cell is known as the "interactome".[1] The study of the interactome may be grouped with other "-omic" disciplines, all of which share an interest in developing strategies to understand biological processes from an integrated point of view, while eschewing the reductionist approach that focuses on information fragmentation. The cell is viewed as a complex, heavily organized system in which diverse elements (proteins, genes, metabolites, etc.) interact and influence each other at multiple levels.[2*,3*]

Most—if not all—cellular processes are driven by molecular interactions and the study of the interactome provides us with a scaffold on which cellular events may be organized.[4−6*] Although the term "molecular interactions" encompasses the interplay between all biomolecules within a cell, the largest amount of data available is for protein−protein interactions, and the term interactome is therefore often used interchangeably with protein−protein interaction networks (PPINs) (Figure 1), which are the subject of this tutorial.

The Advent of High Throughput Interaction Data

Small-scale biochemical or genetic experiments were, for many years, the only source of interaction data. However, low-throughput data presented in multiple individual publications are not amenable to bioinformatic or systems biology analyses unless collated. It is not until recently that this heterogeneous data set has been made available online through initiatives that actively curate and store interaction information in databases using standardized procedures. The creation and development of such databases has run parallel with the production of the first high-throughput interaction data sets.

High-throughput methods for interaction detection became generally available around 18 years ago (reviewed in ref 2*), resulting in an explosion of protein−protein interaction data and motivating a number of serious attempts to produce large and semicomprehensive maps of protein−protein interactions.[7−10**] The first high-throughput data sets made no attempt to map the entire interactome. Instead, they focused on specific biological processes (e.g., pathogenesis of specific disease) or specific sources of data (e.g., cDNA libraries).[11−16**] As computing resources improved and more experimental data became available, objectives became more ambitious and data sets aiming to record the interactome for single species were produced, for example, *Escherichia coli*,[17**,18**] *Saccharomyces cerevisiae*,[5*,19−23**] *Drosophila melanogaster*,[24**] *Caenorhabditis elegans*[25−27**] and human.[28−30**] Although these data sets often lack the biological context of their low-throughput counterparts, high-throughput PPI data derive their importance from their size and the fact that the interactions curated have been

† This Tutorial is part of the International Proteomics Tutorial Program (IPTP11).

Published: March 2, 2012

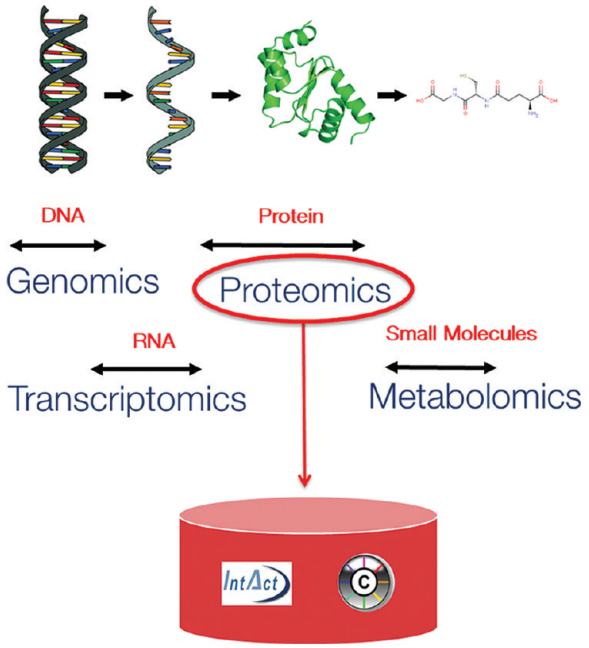


Figure 1. Interactome and interactomics. DNA is transcribed to RNA and RNA is translated to proteins. The proteome is the totality of proteins in a cell, and its discipline of study is called proteomics. The advances in proteomics have allowed for the study of the interactome, which is the totality of molecular interactions with the cell (which may include interactions with small molecules and nucleic acids). Credits for the upper figures: DNA-RNA, Adapted from http://en.wikipedia.org/wiki/File:Difference_DNA_RNA-EN.svg; Protein, yeast Grx2 3D structure model, PDB entry 3d4m, taken from the PDBe Web site (www.ebi.ac.uk/pdbe/); CHEBI, glutathione, CHEBI:16856, taken from the ChEBI Web site (www.ebi.ac.uk/chebi/).

obtained from comparable experimental methods, which allows for confidence scoring and sensitivity estimations.[31**]

There are now several actively curated interaction databases hosting both low- and high-throughput data and some of these are described in Table 1. This list is by no means comprehensive and there are important differences in curation policy and coverage.

Given the large number of interactions that take place in an organism[31**,32**] and the small proportion of these interactions currently covered by experimental data, computational approaches may provide meaningful predictions about uncharted regions of the interactome. There are multiple predictive methods that cover both the mapping of uncharted interactions and the characterization of existing ones, for example, the identification of binding interfaces may be inferred from sequence or structural homology with known binding sites on other proteins. PPIs may also be inferred from interactions found in orthologous, coexpression, colocalization or phylogenetic studies.[33*] Finally, there are computational approaches that rely on text-mining tools used on scientific literature in order to find relationships and interactions between biological molecules, such as iHOP (www.ihop-net.org)[34] or PESCADOR (http://cbdm.mdc-berlin.de/tools/pescador/).[35]

It was recognized early on that the exponential growth in protein−protein interaction data necessitated the creation of new tools to facilitate their analysis.[36*] Early researchers used software originally designed to analyze networks arising from other fields (social networks, transportation networks or organizational graphs): these included Pajek,[37] Graphlet (no longer under development or supported, but still available from the Computer Science Department of the Universtität Passau: http://www.fim.uni-passau.de/en/fim/faculty/chairs/theoretische-informatik/projects.html) and uDraw(Graph) (www.informatik.uni-bremen.de/uDrawGraph). Techniques originally used for analyzing social networks have also been applied to the analysis of PPINs,[38] and specific methods have been created to solve PPIN problems.[39] The open source software platform, Cytoscape, made these tools generally available to biological researchers in 2003.[40,41]

BASIC CONCEPTS

Methods for Identifying Protein−Protein Interactions

There is a wide variety of methods for identifying protein−protein interactions (Figure 2). The growth in published protein−protein interaction data is almost entirely due to the development of two methodologies which can be used in high-throughput mode: yeast two-hybrid (Y2H) and affinity purification mass spectrometry (AP-MS).

Y2H is based on the complementation of two halves of a transcription factor that are bound to two proteins whose interaction is to be tested. The method was first described by Fields and Song in 1989.[42**] The DNA-binding domain (BD) of the transcription factor is bound to the molecule of interest (called the *bait*), while the potential interaction partner (the *prey*) is bound to the activation domain (AD) of the transcription factor. If the bait and prey bind to each other when expressed in a yeast cell (i.e., they interact), then this reconstitutes the transcription machinery and a reporter gene is activated. The original method used β-galactosidase activity as the reporter, but the method has been adapted to use other reporters and to other organisms such as *E. coli*. Y2H is fast, inexpensive and scalable: it is an *in vivo* system and binding sites may be accurately mapped.

Table 1. Summary of Some Currently Available Interaction Databases

database	curation	species coverage	molecule types	reference	comments
IntAct	Deep	Broad	All	91	
MINT	Deep	Broad	Proteins	105	
DIP	Deep	Broad	Proteins	106	
MatrixDB	Deep	Limited	Extracellular matrix molecules only	107	
InnateDB	Deep	Limited	Proteins	108	
MIPS	Deep	Mammalian	Proteins	109	
MPACT	Deep	Yeast	Proteins	110	MPACT is a part of MIPS
BIND	Deep	Broad	All	111	Ceased curation 2006/7
BioGRID	Shallow	Limited	Proteins	112	
HPRD	Shallow	Human	Proteins	113	
MPIDB	Shallow	Microbial	Proteins	66	

dx.doi.org/10.1021/pr201211w | J. Proteome Res. 2012, 11, 2014−2031

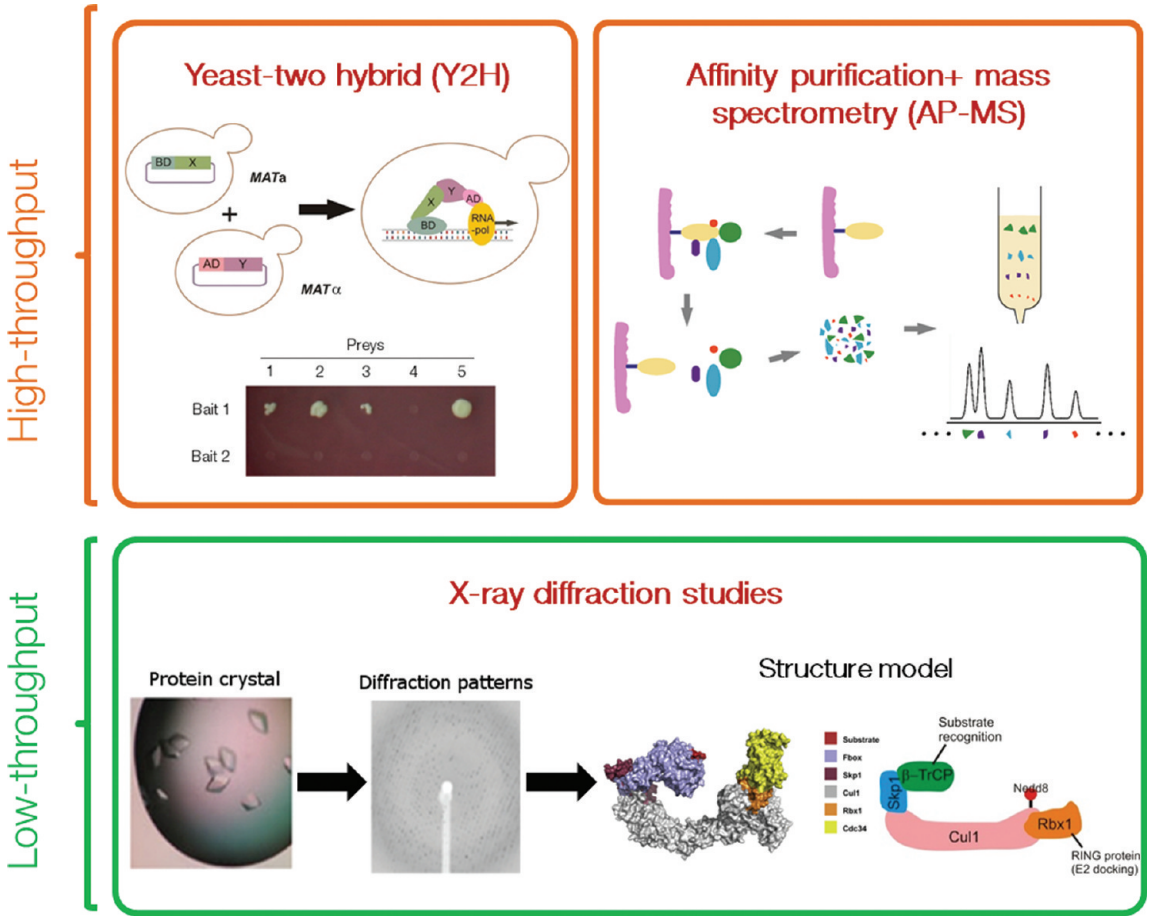


Figure 2. Protein–protein interaction experiments. The most widely used two high-throughput methods for obtaining protein–protein interaction data are yeast two-hybrid (top left) and affinity purification mass spectrometry (top right). Low-throughput methods such as X-ray crystallography (bottom) may be used to obtain detailed structural and chemical insights for selected interactions. Cullin-complex model from Ref 107.

Y2H will quickly screen large numbers of binary interactions, but false positives do occur. Early experiments failed to recognize that if either bait or prey are in fact a transcription factor, the reporter gene may be activated even if there is no interaction between the proteins of interest. Users often regard data from Y2H experiments as proof of a direct, binary interaction, but there are cases when yeast proteins may act as a bridge. When the bait interacts with the yeast protein and the yeast protein binds the prey, then there is in fact no direct interaction between bait and prey.

Finally, it must be remembered that Y2H may show interactions between two proteins that would not normally occur in the same cellular compartment (or perhaps not even in the same tissue). Although the two proteins may interact in the yeast cell, this interaction would never actually occur under physiological conditions. Y2H will also miss interacting pairs when prey proteins fail to express, are toxic to the cell, or require post-translational modifications (e.g., glycosylation or methylation) that do not take place in the yeast cells. An extensive review about the strengths and weaknesses of this technique and its variants can be found in ref 43*.

In AP-MS, a single protein or molecule of interest is affinity-captured in a matrix as *bait*.[44**] A protein mixture (often a lysate from the cell or tissue of interest) is passed through the matrix, and interacting partners (*prey*) are retained by interaction with the bait. Proteins that do not interact pass through the matrix and are discarded. There are multiple variations in the affinity purification step, including immunoprecipitation and pull-down of epitope tagged molecules.[45,46*] The resolving power of AP-MS comes from the analysis of the captured sample by mass spectrometry, identifying interaction participants from their peptide signatures. Mass spectrometry is capable of identifying hundreds of potential interactors simultaneously at subpicomole concentrations. The major advantage of AP-MS is therefore the ability to examine interactions between multiple proteins (*n*-ary interactions). Another advantage of AP-MS is that prey proteins are present in their native state and concentration.

AP-MS also has its limitations. Prey proteins without a prototypic peptide signature (due to obscure post-translational modifications, for example), or that are present only in very low amounts, will not be identified unless very specific approaches are applied (or not at all). Transient but biologically important interactions (e.g., enzyme–substrate interactions) will probably be missed, and the same goes for weak interactions, which will be disrupted by washing procedures. The method requires cell or tissue lysis, so proteins from different cellular compartments may interact, which never encounter each other under physiological conditions. A detailed review about the method and several variations can be found in ref 46*.

Large-scale interaction data sets obtained with the high-throughput methods outlined above are often retested using alternative approaches that improve the confidence for an interaction.[6*,31**] Medium or low-throughput methods such as luminescence-based mammalian interactome mapping (LUMIER),[47]

dx.doi.org/10.1021/pr201211w | *J. Proteome Res.* 2012, 11, 2014–2031

mammalian protein−protein interaction trap (MAPPIT)[48] or fluorescence/bioluminescence-resonance energy transfer (FRET/BRET)[49,50] complement data obtained from high-throughput experiments. These methods have been reviewed in detail by Petschnigg et al.[51*]

Other experimental methods for determining protein−protein interactions are not amenable to large- or even medium-scale experiments, but provide deeper insight into certain characteristics of the interaction. X-ray crystallography[52] has been said to represent the gold standard in evidence for protein−protein interactions, because it provides extremely detailed, high quality data about binding surfaces, with detailed mapping of binding sites to the level of individual atoms and chemical bonds. The main limitation of crystallography is that it is extremely challenging technically, and the technique is extremely low throughput, because it may take years to successfully crystallize a pair of proteins. Crystallography also requires large quantities (in the order of milligrams) of extremely pure protein and is primarily suited to studying water-soluble proteins. Crystallography can give only limited information about trans-membrane proteins (and hydrophobic proteins in general): this data is usually confined to the extracellular or cytosolic domains of the protein of interest.

In order to capture one class of transient interaction data, enzymatic reactions, such as substrate cleavage, are taken as evidence for an interaction between the enzyme and substrate. Such analyses may only be performed *in vitro* and not in whole cells, where direct interactions cannot be distinguished from enzyme pathways. A major concern about data obtained through these methods is that many enzymes that are specific *in vivo* are promiscuous *in vitro*; thus, evidence that an enzyme cleaves a substrate *in vitro* does not mean that the reaction occurs *in vivo*.

In summary, all interaction data are to an extent potentially artifactual and the limitations of each method must be borne in mind by the researcher and may bias the apparent interactome generated from a database of interactions. The confidence with which one may say an interaction occurs grows with the evidence base, but many interactions identified by high throughput experiments have never been independently confirmed by other methods. No method is perfect and false positives and negatives will occur regardless of methodology,[31**] but generally speaking, interactions detected by more than one methodology are more likely to be "real". Comprehensive mapping of the interactome will necessarily involve combinations of methods.

Protein Identifiers

An interaction database is queried by giving the database a list of potential interaction participants. This is usually a list of proteins, or more specifically, protein identifiers. A protein is uniquely specified by its amino acid sequence, but behind this simple concept hides a panoply of complications.

The first two complications may be predicted from knowledge of genetics. First, allelic differences occur between individuals within the same species. These may be single nucleotide polymorphisms (SNPs), frame shifts, insertions or deletions. With the exception of silent SNPs, these mutations will all result in alterations to the protein sequence. Should each allele be assigned a different identifier or should all alleles of a single gene share the same identifier?

The second complication arises from the fact that eukaryotic proteins are frequently composed of multiple exons and introns. Exons code for protein, introns do not. Exons from the same gene may be assembled in various ways (splicing) so as to produce different proteins (splice variants).[53**] Should each splice variant be given a separate identifier, or should all the products of a single gene share the same identifier?

The third complication is less obvious. Proteins are not usually sequenced directly: instead, the corresponding gene or cDNA is sequenced and the protein sequence then predicted from the DNA sequence. Protein sequences therefore vary according to gene model (where is the initiation site? where are the intron/exon boundaries?). For this reason, multiple published sequences may exist for the same protein, each of which varies by only a few of amino acids. Proteins predicted from DNA sequences may even 'disappear' or be withdrawn: this may occur, for example, when the draft sequence for a genome assembled from throughput sequencing methods is later shown to have been assembled incorrectly. Whole genes may then disappear from the corrected genome of the organism, leading to the disappearance of putative "proteins" that never actually existed.

The identifier used in this tutorial will be UniProt Knowledgebase[54**] (www.uniprot.org) accession numbers (UniProtKB AC, e.g., P12345). UniProtKB has two strategies for handling these complications, one protocol for UniProtKB/TrEMBL (the automatically curated section of UniProtKB) and another for UniProtKB/Swiss-Prot (the manually curated section of UniProtKB). UniProtKB/TrEMBL contains a high level of redundancy and often contains multiple identifiers for the same gene transcript, each differing by only a few amino acids. This is because entries in TrEMBL are never combined unless the sequences they describe are 100% identical. UniProtKB/SwissProt, on the other hand, contains a single nonredundant entry for each gene. All protein transcripts for the gene, including splice isoforms and post-translationally cleaved peptides are described within that one entry.

Other protein sequence databases include the Reference Sequence database (RefSeq) and International Protein Index (IPI). RefSeq is an open access, manually curated database of publicly available DNA, RNA and protein sequences operated by the National Center for Biotechnology Information.[55] RefSeq identifiers follow the format NP_123456.1.

IPI was a complete, nonredundant data set of human, mouse and rat proteins built by combining data from SwissProt, TrEMBL, Ensembl and RefSeq.[56] IPI closed in September 2011 and its functions have been taken over by UniProtKB. IPI identifiers (format IPI1234567890) are therefore deprecated and, as of September 2011, are no longer updated.[56,57]

Lastly, there also exist important species-specific databases, each with its own identifiers and curation policy. Two of the largest are WormBase (*Caenorhabditis elegans,* format WBGene12345678) and FlyBase (*Drosophila melanogaster,* format FBgn1234567).

It is frequently (but not always) possible to translate between different types of identifiers. However, there may not be a one-to-one correspondence between identifiers. One microarray probe may map to multiple genes and one gene may map to multiple isoforms. These complications are frequently disregarded in analyses, but it is important to be aware that the initial choice of identifier may significantly influence the results obtained.

dx.doi.org/10.1021/pr201211w | *J. Proteome Res.* 2012, 11, 2014−2031

Interaction Databases

To make effective use of interactome data, it is fundamental to understand where and how data is collected and stored. These considerations will determine what database(s) the researcher will use.

The fundamental character of a database is determined by its data acquisition policies. Broadly speaking, data acquisition may be described in two dimensions: shallow versus deep (reflecting the amount of data captured about the experiment in the database), and manual versus "automatic" (whether the data acquisition is done by a human being or whether the information is obtained by computational approaches, for example, text-mining algorithms or predictive methods based on sequence or domain analysis). The process by which manuscripts are selected from the literature and how the interaction data provided in a manuscript is extracted and entered into a database is known as curation, so from now on we will use it as a synonym for manual data acquisition.

Differences in curation policy mean that the representation of PPI data is extremely heterogeneous and the data captured by different repositories may not overlap, even when the same publications are curated.[58*] Shallow curation is not necessarily better or worse than deep curation: for network analysis, the researcher may only be interested in database coverage ("I want as many proteins as possible") rather than curation depth ("I want as much detail about each interaction as possible"). Curated data is considered to be of higher quality than data obtained by computational prediction or automatic text-mining, although errors in data capture will still occur and false positives may be captured as well as true interactions.[59*,60*]

Databases vary also according to their source of data (primary or secondary). Primary databases collect data exclusively from peer-reviewed publications (e.g., IntAct, MatrixDB).[61*] Secondary databases (or meta-databases) seek to integrate the data from multiple primary databases (e.g., APID[62] and PINA[63]). To increase coverage, databases may go further and include computationally predicted interactions for which there are no experimental evidence (e.g., MIMI,[64] PIPs,[65] MPIDB,[66] STRING,[67] I2D/OPHID[68] or UniHI[69]).

Efforts such as the International Molecular Exchange consortium (IMEx) guidelines[70**,71**] aim to standardize the level of detail that needs to be captured in order to accurately represent an interaction. The basic level of detail required was outlined in the Minimum Information required to report a Molecular Interaction Experiment (MIMIx) standard,[72**] which is followed by many of the databases listed above. The Human Proteome Organisation Proteomics Standards Initiative Molecular Interactions work track (HUPO PSI-MI)[73**] has produced unified XML and tab-delineated formats for the consistent representation of interaction data, so that data provided by different databases may be easily combined. The PSICQUIC interface is an online service that combines data from multiple PSI-MI compliant databases (www.ebi.ac.uk/Tools/webservices/psicquic/view).

Visualizing the Interactome as a Network

Interaction data is usually depicted in the form of a graph or network. The interacting partners (proteins or other molecules) are represented as nodes, while the interactions themselves are represented as lines or edges connecting the nodes. Graphs provide a powerful, global overview of the interactions. However, the representation of large PPINs usually results in an incomprehensible 'hairball' from which a human reader may find it difficult to extract useful information. Various software packages provide tools and visualization strategies to extract meaningful information from such graphs. For example, the addition of expression data to the nodes may help organize and clarify the networks.[74*]

There are several publicly available tools that are extensively used to visualize PPINs and these have been reviewed elsewhere.[74*] Arguably the most popular of these is Cytoscape, the tool that will be used for this tutorial.[41]

The popularity of Cytoscape may be attributed to its open-source, modular design, which affords great flexibility and extensibility. Aside from the suite of tools that comes packaged with Cytoscape, additional functionality may be gained from a library of third-party plugins that provide complementary tools for network analysis and functional enrichment.[75] The open-source model adopted for Cytoscape means that interested developers may build their own plugins to address almost any network analysis problem.

Network Clustering

The study of the protein interactome is essentially the study of how proteins work together. The strategies that aim to interpret PPINs generally try to find common attributes within members of the network. Useful attributes include similarities in expression pattern, function or subcellular localization. Nodes may also be grouped on the basis of network topology: groups of highly interconnected nodes may form clusters. Although clusters are identified solely on the basis of the topology, the assumption underlying this approach is that clusters will identify groups of proteins that share a similar function. One algorithm designed to identify such clusters is the Molecular Complex Detection (MCODE).[39] MCODE finds small clusters of densely connected nodes, but does not aim to assign every protein within the network to a cluster. The main weakness of the MCODE algorithm is that it is sensitive to noise in the network (specifically, false-positive interactions).[76*] The other issue is that MCODE does not aim to assign every node to a cluster, which means the resulting clusters are often small and it is unclear how users should interpret nodes that are not clustered.

Community analysis was originally developed for the study of social networks. The algorithm begins by simplifying the network to give it a community-like structure by removing duplicate edges (you count each friend only once) and self-looping (you cannot be friends with yourself). Community analysis was introduced to the Cytoscape community via the GLay plugin, which included a suite of these methods.[38] The clusterMaker plugin has incorporated the GLay implementation of the Newman-Girvan fast greedy algorithm.[77] The algorithm identifies clusters by iteratively removing edges from the network and then checking to see which nodes are still connected.

Brohée and van Helden evaluated four methods for the detection of previously annotated complexes.[76] The four methods evaluated were Markov clustering (MCL), Restricted Neighborhood Search Clustering (RNSC), Super Paramagnetic Clustering (SPC) and MCODE, each one based on a different search strategy. MCL (which uses flow simulation to find clusters) was found to be the algorithm that identified annotated complexes most robustly despite background noise (false-positives) and random network permutation.

Whichever clustering strategy is chosen, the question of greatest interest to the biologist is how well a particular algorithm identifies biologically meaningful protein complexes. Cluster-detection algorithms remain an active area of research

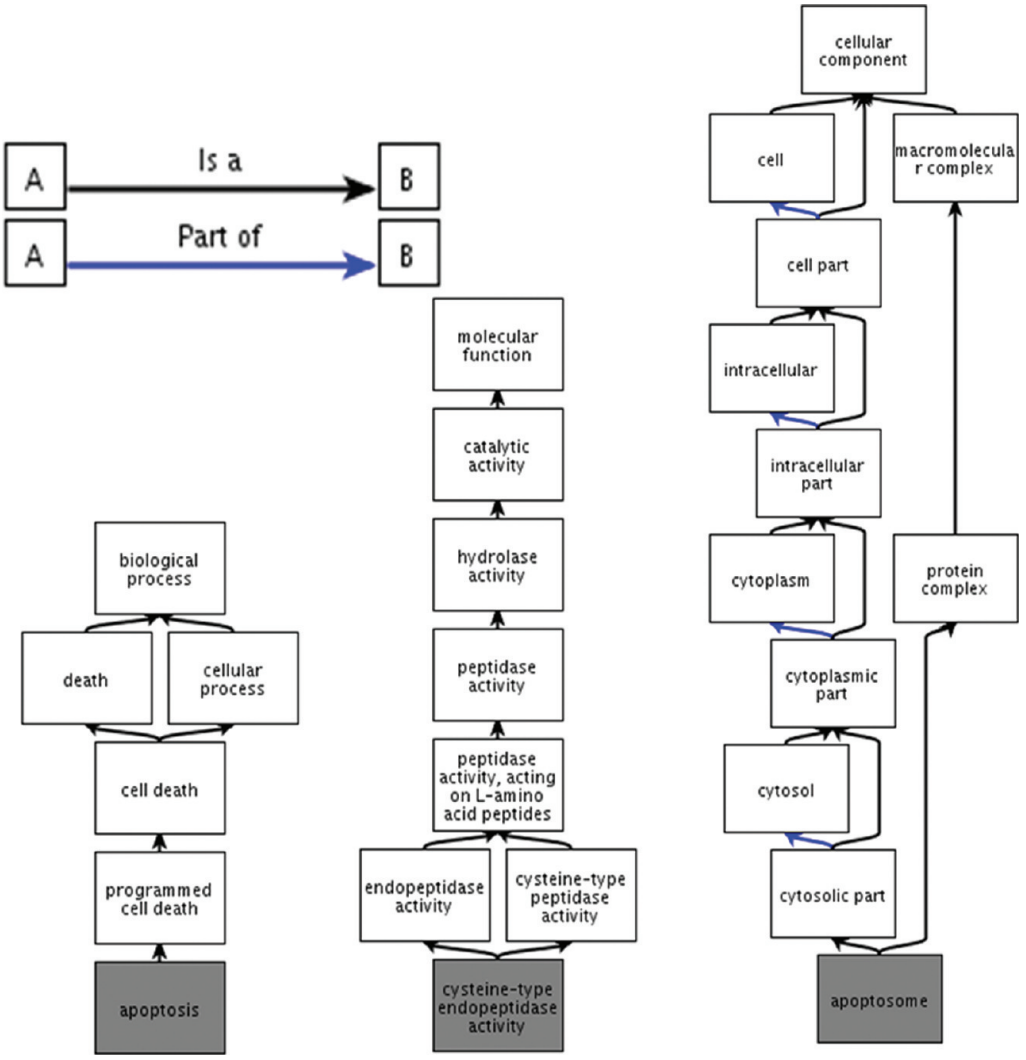


Figure 3. Gene ontology: three controlled vocabularies. This figure displays the GO terms from each of the three ontologies used to annotate the protein, caspase-9. The most specific terms are at the bottom: these are "apoptosis" (which is a Biological Process), "cysteine-type endopeptidase activity" (Molecular Function) and "apoptosome" (Cellular Component). Arrows point from children to parents: black arrows are "is a" relationships, blue arrows indicate "part of" relationships. The representation of the ontologies has been taken from the QuickGO Web site for navigating and representing the gene ontology (www.ebi.ac.uk/QuickGO).

and the interested reader is referred to the recent review by Wang et al.[78*]

Network Annotation

Annotation of nodes and edges is often required to make use of the information present in PPINs.

Edge annotation may include information about the method by which the interaction depicted by the edge was detected, associated quantitative parameters such as values obtained from experimental observation, or confidence scores calculated *a posteriori*. These may be used to highlight or to filter out interactions of a specific type or confidence level.

Nodes may be annotated with different types of identifiers for different purposes. Taking hemoglobin subunit beta as an example: the UniProtKB accession P68871 is useful for searching databases, but the gene symbol HBA1 is more human-readable and therefore useful for presentation and publication. Additional annotation such as quantitative data about expression levels, cellular location (e.g., cytosolic or ribosomal) or molecular function (e.g., serine protease or DNA binding) may also be useful depending on the needs of the individual researcher.

There are different ways to provide node or edge annotation, depending on the software platform. In Cytoscape, this information may be embedded in a network file (formats such as XGMML support this) or provided in the form of a text table that is loaded into Cytoscape after the network has been drawn.

Although it is most convenient for annotation to be embedded in the network file itself, it is more often the case that information has to be integrated from another source in order to provide biological context that allows for a meaningful analysis.

The most widely used source of additional information in network analysis is the Gene Ontology (GO) project.[79**] The GO project is an international initiative that aims to provide consistent descriptions of gene products (i.e., proteins). These descriptions are taken from controlled, hierarchically organized vocabularies called "ontologies" (Figure 3). GO uses three ontologies covering three independent biological domains.

dx.doi.org/10.1021/pr201211w | *J. Proteome Res.* 2012, 11, 2014−2031

These are (1) Cellular Component, or the location of the protein within the cell (e.g., cytosol or mitochondrion); (2) Biological Process, or a series of events accomplished by one or more ordered assemblies of molecular functions (e.g., glycolysis or apoptosis); and (3) Molecular Function, which is the activity proteins possess at a molecular level (e.g., catalytic activity or transmembrane transport). The interested reader is directed to the GO Web site for a detailed description of definitions and policies (geneontology.org).

GO is used to annotate large sets of proteins like the ones found in PPINs and strategies have been developed to make effective use of such annotation. Enrichment analysis looks to see which GO annotations are over- or under-represented in a protein list. Of the three available ontologies, the most commonly used for this purpose is the Biological Process ontology. This way, modules of proteins that are part of the same process are highlighted, allowing a functional interpretation of the network to emerge. This type of analysis is often combined with cluster detection tools and will be described in greater detail later in this tutorial.

There are a number of other initiatives that describe proteins in terms of one or more pathways that underlay different biological events (e.g., protein translation or mitotic division). An example of this approach is the Reactome project,[80,81] an open-source pathway database based at the European Bioinformatics Institute. Reactome is curated and cross-referenced to other major bioinformatic resources, such as GO, NCBI Entrez Gene, Ensembl and UniProt. Other similar resources include the Kyoto Encyclopedia of Genes and Genomes (KEGG)[82] and the WikiPathways initiative.[83]

EXAMPLES OF PPIN ANALYSES

Every interactome analysis strategy is different, and will need to be adapted to a specific biological question. Every method has its strengths and weaknesses. As this field of biology is still under development, it is difficult to be dogmatic about analysis strategies; nevertheless, there are several commonly used tools that may be adapted to address specific needs. These tools will be described in greater detail later in this tutorial.

In order to give an impression of the diversity of interactome studies, we will now describe briefly two publications taken from the recent literature. Both data sets are available in the IntAct molecular interaction database.

Soler-Lopez et al., a Network for Alzheimer's Disease

In the first of our two examples, Soler-Lopez et al.[84] generate an Alzheimer's disease (AD) centered protein interaction network.

AD is the most prevalent neurodegenerative process affecting human population. It is strongly associated with old age and characterized by the accumulation of extracellular neurofibrillary tangles in the central nervous system. Even though enormous effort has been made to unravel the molecular mechanisms that underlie the origin of the disease, most of them remain elusive and there is no curative treatment currently available. The characterization of the fraction of the interactome related with the proteins involved is a tool of enormous potential in the study of Alzheimer's disease and related processes.

The authors take a short list of manually selected AD-related genes and perform Y2H library screenings to find interacting partners. They also add to that list genes located in susceptibility-related loci in human chromosomes and filter them on a variety of criteria, for example, they include brain-specific proteins but remove transcription factors (due to the promiscuity of transcription factors in the Y2H assay) and transmembrane proteins (due to the fact that often these proteins cannot be properly expressed or imported to the nucleus in a yeast cell). The identified network (71 nodes and 246 edges) is evaluated and selected with stringent criteria to generate a high-confidence set (HC). A large proportion of these interactions (around 87%) are then validated with alternative methods such as GST-pull down or coimmunoprecipitation.

Compared against known interaction data sets, the data set showed a low degree of overlap with the known fraction of the interactome. This is not surprising, because a common characteristic of PPI data sets is that strategies that aim to maximize accuracy often result in decreased interactome coverage. A GO enrichment analysis was performed on the HC network, highlighting overrepresented terms in the three branches of GO that were both consistent with knowledge present in the literature and bringing up new functions previously unknown.

The PPIN was then broadened with known PPIs from publicly available databases. This increased its size to 1704 nodes and 5881 edges. Cluster analysis using the MCL algorithm was then performed and the clusters found were again submitted for GO enrichment analysis using the Biological Process branch of the ontology. The homogeneity of the functions assigned to the proteins in the clusters was also tested. Among the clusters identified, the authors describe in greater detail two of special interest.

PDCD4, a nuclear protein associated with cell death and with protein translation-inhibition abilities, was found to be present in a module homogeneous for "translation elongation" and which was up-regulated in AD human brain tissue. These observations point to a potential role for this protein in AD-related neurotoxicity in conjunction with proteins known to have a role in the disease such as APOE and PSEN2.

A second module of interest was the one that includes ECSIT, an adapter protein involved in nuclear factor kappa-B activation and the BMP signaling pathway. The module also shows enrichment for redox signaling processes and immune responses. The study uncovered a previously unknown interaction between ECSIT and the AD-related protein APOE, which is found in AD-affected neurons, associated with neurofibrillary tangles and with mitochondria. The authors proposed a model in which ECSIT links oxidative stress, mitochondrial dysfunction and inflammation.

Sowa et al., Deubiquitinating Enzyme Interaction Landscape

The second example, taken from Sowa et al., explores a significant portion of the interactome of human deubiquitinating enzymes (DUBs).[85]

Ubiquitin is a small polypeptide that tags protein targets for degradation by the proteasome when it is covalently attached to proteins in the form of poly ubiquitin chains, a process known as poly ubiquitination. This process is performed via an enzymatic cascade of ubiquitin-activating enzymes (E1s), ubiquitin conjugating enzymes (E2s) and ubiquitin ligases (E3s). DUBs have the double role of antagonizing this enzymatic cascade and renewing the pool of free ubiquitin in the cell. Little is known about the mechanisms that drive substrate recognition for DUBs or their role in the regulation of biological processes.

The strategy used in this publication makes use of a software tool called CompPASS that the authors developed to analyze the results of the AP-MS experiments they used to explore the human DUBs interactome. Using a green fluorescent protein construct as a negative reference set and performing repetitive

dx.doi.org/10.1021/pr201211w | J. Proteome Res. 2012, 11, 2014−2031

AP-MS experiments with 75 human DUBs as baits, they found 774 high-confidence interacting partners selected by CompPASS.

CompPASS uses ion counts of specific peptides (or total spectral counts, TSCs) as a quantitative measurement of interaction strength. Once this high-confidence list of interactions was built into a network, its accuracy was confirmed by demonstrating that several of the interactions found were already documented in the literature by searching BioGrid (thebiogrid.org). Further validation was achieved by performing coinmunoprecipitation experiments with a fraction of interactions, with a high rate of success (45 out of 66 selected interactions were validated). A detailed comparison with DUB interactions found in the literature revealed a large overlap with interactions found using affinity-based techniques (50 to 70%), but much lower overlap compared to other techniques such as Y2H or *in vitro* methods (4%). This is an example on how different techniques may map different interactions within the same interactome.

The proteins in the PPIN were annotated using the GO Biological Process and Cellular Component ontologies. A map was then built in which the DUBs were assigned to functional modules located in specific compartments of the cell. Notably, one-third of the participants were located in the nucleus of the cell. Among the biological processes represented in the network were terms linking the DUBs to chromatin processing, DNA damage repair, autophagy or Endoplasmic Reticulum-Associated protein Degradation (ERAD). Additional interactions were added from various databases and topological clustering analysis was performed, allowing for a classification of DUBs into seven topological classes. Such classification helped in identifying functional modules, both known (e.g., the signalosome) and previously unknown.

The authors provide detailed analyses of four previously studied DUB complexes and integrate RNAi screening data to previously unknown complexes. As a final experimental validation of this analysis, the role of one of the DUBs (predicted to be a modulator of VCP-mediated ERAD) was confirmed by functional assays. Among the most interesting functional modules found by this study was a putative regulatory network they named the DPW (DUB, phosphatase, WD40) network. In the DPW network, DUBs are activated by WD40 domain-containing proteins in a process involving phosphatases.

■ PRELIMINARY CONCEPTS FOR THE TUTORIAL

The tools presented here are accessed online and you will therefore require access to a web browser and an Internet connection to use them.

UniProt Knowledgebase Accessions

This tutorial will take UniProtKB/Swiss-Prot accessions (UniProtKB AC) as its starting point. The choice of UniProtKB ACs allows us to easily integrate external information from both IntAct and GO resources into our analysis.

Your list may have been obtained from a variety of sources (*e.g.,* a high-throughput mass spectrometry data set or a microarray) and the identifiers you obtain may be RefSeq or IPI. There are multiple online services available to convert other identifiers to UniProtKB accessions, one of which is the UniProt ID mapping service (www.uniprot.org, go to "ID Mapping"). An alternative is the Protein Identifier Cross Reference (www.ebi.ac.uk/Tools/picr).[86]

Using IntAct to Generate a Network of PPIs

Of the multiple interaction databases available, this tutorial will focus on IntAct, a publicly available repository for curated PPI data hosted by the European Bioinformatics Institute. All interaction with IntAct occurs online via a web browser (www.ebi.ac.uk/intact).

The most accurate results will be obtained from IntAct if the search is performed using UniProtKB accessions. Gene symbols will work as inputs, but species specificity will be lost unless the species or NCBI taxid is included in the search. IntAct returns a list of interactions (in the form of a table) for all the proteins in the list you give it, along with their interaction partners. You may direct IntAct to return interaction data annotated with information about the publication in which that interaction is described as well as the type of experiment and level of evidence.

Using Cytoscape to Visualize and Analyze the Network

Cytoscape 2.8.2 is an open source, publicly available, network visualization and analysis tool (www.cytoscape.org).[41] It is written in Java and will work on any machine running a Java Virtual Machine, including Windows, Mac OSX and Linux.

We will also use two plugins to extend the Cytoscape's functionality. The first is clusterMaker 1.9, a suite of tools for identifying clusters in networks (www.cgl.ucsf.edu/cytoscape/cluster/clusterMaker.html[89]). The second is BiNGO 2.44, a tool for annotating proteins with their corresponding GO terms.[88] We will use BiNGO to annotate our network and help us interpret it. The GO file used was an OBO file version 1.2 (gene_ontology.1_2.obo) downloaded on the 21 Oct 2011. The ontologies are updated weekly, so you should download the most recent version from the GO website (geneontology.org/GO.downloads.ontology.shtml).

■ WORKED EXAMPLE

Introduction to the Data Set

The example data set is taken from a chronic myeloid leukemia-derived cell line (K562) treated with a candidate drug, belinostat. Belinostat is a histone deacetylase inhibitor that in this experiment was being evaluated for its use in the treatment of hematological malignancies and solid tumors.

The data set may be downloaded from the PRoteomics IDEntifications database (PRIDE)[89] (www.ebi.ac.uk/pride, experiment reference 15401[90]). This is a mass spectrometry data set consisting of a list of 1809 peptides obtained from an AP-MS experiment using the drug belinostat as bait. The peptides were mapped to 326 proteins identified with International Protein Index identifiers (IPI IDs). This list of 326 IPI IDs is what we are going to use as a starting data set in the tutorial and it can be found in the supplementary file "HUPO tutorial data set IPI.txt". From what is known of the function of this class of drugs, we expect to find protein complexes and pathways associated with DNA metabolism.

Mapping Identifiers

We will use UniProtKB accessions in this tutorial. The advantages of using these ACs are that (i) they are stable (they are not changed or updated once assigned); (ii) they can reflect isoform information, if provided; and (iii) they are recognized by many interaction and annotation databases (in this instance, the two databases we will be using: IntAct and GO).

Figure 4. Converting identifiers to UniProtKB ACs. (A) ID Mapping tool on the UniProt Web site may be used to translate multiple types of identifiers to UniProt ACs. (B) Example of UniProt ID Mapping tool results. This screenshot shows only the manually curated UniProtKB/Swiss-Prot entries, as indicated by the gold star in the "Status" column. Entries that have not been reviewed are indicated by a silver star.

The first step is to convert the IPI IDs to UniProtKB ACs using the UniProtKB ID Mapping tool found as a tab on the UniProtKB Web site (www.uniprot.org) (Figure 4A). The IPI IDs may be copied from the list into the "Database identifiers" box, or the file may be uploaded using the "Choose File" button. The type of identifier being upload needs to be chosen from the "From" drop-down menu (in this case, "IPI"). From the "To" drop-down menu, select "UniProtKB AC".

Once a list of UniProtKB ACs has been produced, the results may be filtered by status: the search will be restricted to UniProtKB/Swiss-Prot entries to ensure that a nonredundant list of identifiers is used as a starting point. Select the "UniProtKB" link just below the search dialogue area to open the page shown in Figure 4B. Just above the results table, click the "Show only reviewed" link to obtain a list of 265 fully reviewed and annotated UniProt ACs. Clicking in the "Download" button delivers a file that you should now save to your local computer.

Retrieving Protein Interactions from IntAct

With the list of UniProtKB ACs we have just obtained, we will now use IntAct (www.ebi.ac.uk/intact)[91] to find curated interactions. Enter the Web site (Figure 5A). At the top of the page is a "Search" dialogue box, with a series of tabs below it, listing options to browse or search the interactions stored in IntAct.

dx.doi.org/10.1021/pr201211w | J. Proteome Res. 2012, 11, 2014–2031

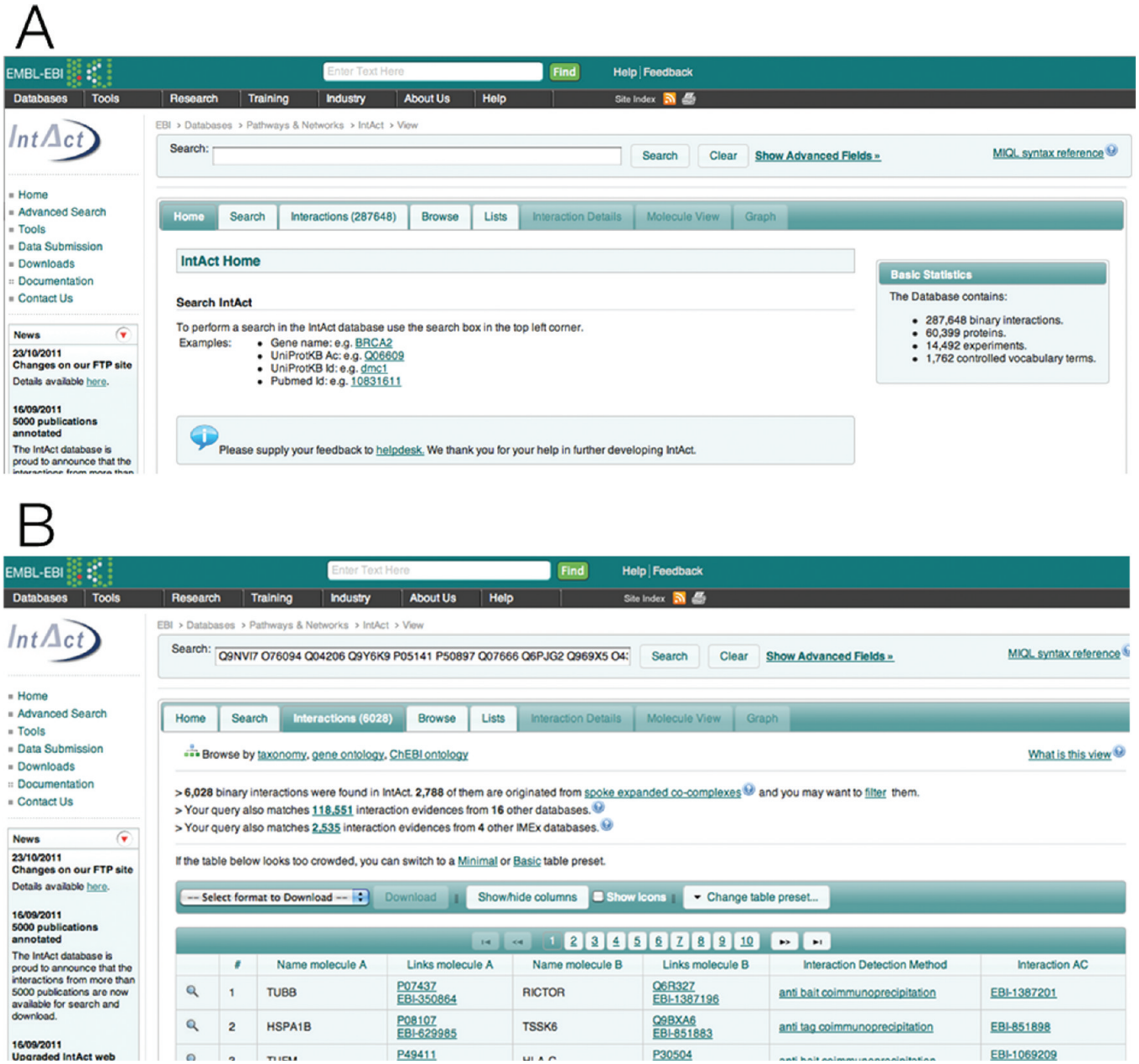


Figure 5. IntAct database home page. (A) IntAct home page: basic and advanced search, along with browsing and listing options are available for navigation through the database. (B) Example search results: Interactions are listed in a table that can be displayed and shown in a number of formats, accessible from the drop-down menus located just above the table headings. The user is also provided with information about the number of interactions found and with the option to filter out spoke-expanded interactions (see text for further explanation).

To retrieve a list of interactions, paste the UniProtKB ACs into the 'Search' dialogue box and click 'Search'. A list of 6028 PPIs was obtained on 4 November 2011 (Figure 5B). Note that the number of interactions returned will change with time, because IntAct is constantly updated.

This list includes spoke-expanded data, which we will filter out in this case. IntAct records all interactions as binary interactions (pairs of proteins) in tabular form. However, some experiments such as AP-MS, protein complexes are pulled-down without any information about how the proteins within the complex interact with each other. IntAct is able to store *n*-ary interactions (interactions among *n* proteins); but when data is exported in a tabular form, only binary interactions can be represented. There are two major ways of transforming *n*-ary into binary data. A spoke expansion records each of the proteins in the experiment as an interactor of the bait. In a matrix expansion, every single possible two-protein combination is recorded as an interaction (accompanying PowerPoint presentation, slide 21). IntAct utilizes the spoke expansion model,[92] but neither of these algorithms is perfect and both may lead to false positives. IntAct gives you the option to remove these interactions from the results it returns. After filtering out the spoke-expanded interactions, we are left with 3240 interactions.

There are several file formats that may be used to export data from IntAct. For example, the PSI MITAB format is a simple representation of the interactions in a plain tabular format that is human-readable[73**] and complies with the requirements of the IMEx consortium for the amount of information given to describe a molecular interaction. As we are going to use the

dx.doi.org/10.1021/pr201211w | *J. Proteome Res.* 2012, 11, 2014−2031

program Cytoscape to represent our network, we choose the XGMML format (the eXtensible Graph Markup and Modeling Language) an XML format, which is used for graph description. After clicking "Download" a new tab is open and we can save the results in a file that we will import into Cytoscape.

Visualizing the PPIN

We will use Cytoscape 2.8.2 to visualize our network. You should also have the clusterMaker 1.9 and BiNGO 2.44 plugins installed. Instructions for installing these are given in the appendix of the accompanying PowerPoint presentation.

The main window in the Cytoscape user interface (Figure 6A) displays the network (all the network manipulations and "working" occurs in this window). The lower-right pane (the Data Panel) contains three tabs that show tabulated information about node, edge and network attributes. The left-hand pane (the Control Panel) is where navigation, visualization, editing and filtering options are displayed. We will limit our explanation to the basic functions required to perform our analysis, but the interested reader is referred to the extensive online documentation to exploit the full range of tools provided by Cytoscape.

In order to draw our protein interaction network, we need to import the XGMML file we obtained from IntAct into Cytoscape. Go to "File" → "Import" → "Network (Multiple File Types)". A dialogue box will then appear in which you can select the XGMML file.

The nodes in the network will be laid out in a grid (Figure 6A). A zoomed-out view of the whole network is always visible in the "Network" tab of the Control Panel (left-hand pane, Figure 6A, marked 1 in red). The main window shows a zoomed-in view of the network. The part of the network visible the main window is marked in blue in the Control Panel (Figure 6A, marked 2 in red). You may navigate around the network by dragging the blue box to the point of interest. Use two-finger drag on your touchpad (Mac OSX) or the middle wheel of your mouse (Windows) to zoom in and out of the network. The blue area will shrink or grow to show you what part of the network you are looking at.

It is important to save your work regularly. Each Cytoscape file contains all the data from a "Session" that may contain several networks, each navigable via the "Network" tab of the Control Panel. Unfortunately, tables generated by plugins are not saved as part of the session file: these have to be saved separately as tables of tab-separated values or comma-separated values. You may also copy and paste the tables into an Excel file.

The default grid layout is not very informative and difficult to read. The first step is therefore to lay the network out again using a force-directed layout. Go to "Layout" → "Cytoscape Layouts" → "Force-Directed Layout" → "(unweighted)".

A force-directed layout models the nodes as if they are electrically charged ("repelling" each other) and the edges as springs (so they pull the nodes together). The first thing that is obvious from the layout is that most of the nodes are connected to each other by interactions, but a small number of orphan interactions are dotted around the edge of the screen. We are going to ignore these orphan interactions for the purposes of this tutorial and focus the analysis on the large central cluster (Figure 6B).

Clustering Analysis using clusterMaker

We are now going to work with the large main cluster only. Select all the nodes in this cluster, while excluding the orphan interactions. Now start clusterMaker. Go to "Plugins" → "Cluster". We are going to use the GLay Community Clustering algorithm due to its superior performance in terms of computing time usage and ease of use. A commonly used alternative is MCL. Choose "Plugins" → "Cluster" → "Community cluster (GLay)" and then follow the instructions on screen to first create and then visualize the clusters.

As seen in Figure 7, 61 clusters are produced and laid out in a new window, arranged in decreasing order of size. When a cluster is selected, a list of participants within that cluster appears in the Data Panel under the "Node attributes" tab. These clusters can represent functional submodules layered behind the network structure. To assign biological meaning within the clusters, we will now perform a GO annotation enrichment analysis.

GO Enrichment Analysis using BiNGO

You need to have the BiNGO plugin installed to perform the analysis. If you do not, follow the instructions given in the appendix of the accompanying PowerPoint presentation. Before performing a BiNGO analysis, remember to update GO reference files to get the most updated version available. The files that BiNGO uses by default are out-of-date and should not be used (www.psb.ugent.be/cbd/papers/BiNGO/Customize.html).

BiNGO is a Cytoscape plugin that annotates proteins (nodes) with gene ontology (GO) terms and then performs an enrichment analysis in order to figure out which terms are over- or underrepresented in the population.

As a starting point, we recommend that the BiNGO analysis is applied to the whole data set to see an overview of all the terms. Subsequent analyses may then focus on individual clusters or subsets of the network, using a view suitable to pick out functional modules.

To select all the nodes in the network, you need to switch back to the windows displaying the whole network, using the "Network" tab in the Control Panel. Use the mouse to select all the nodes in the network by clicking and dragging.

To start BiNGO, go to "Plugins" → "Start BiNGO 2.44". Do this only once: Cytoscape will not stop you from opening multiple copies of the BiNGO setup menu (which will lead to confusion).

The BiNGO setup screen will now appear. Name the cluster you are going to analyze in the text box "Cluster name". Under "Select ontology file" choose the Gene Ontology file you previously downloaded from geneontology.org. Under "Select namespace" select "Biological Process". Under "Select organism/annotation" choose "Homo sapiens". Finally, press the "Start BiNGO" button. You will receive a warning saying, "Some category labels in the annotation file are not defined in the ontology". The warning refers to identifiers that are not properly mapped in the GO reference file by BiNGO. There might often be a discrepancy between the identifiers provided in the interaction network and those found in the GO reference file. In our example, IntAct hosts interactions between different types of molecules, thus there will be a small proportion of interactions in our network that involve nonprotein molecules and will not be mapped to the GO reference file. The number of these interactions is small and will have no GO annotation. Ignore this warning and click OK.

The GO terms found are displayed in two ways (Figure 8A). The first is a table of GO terms found; the second is a directed acyclic network in which nodes are the GO terms found and directed edges link parent terms to child terms.

dx.doi.org/10.1021/pr201211w | J. Proteome Res. 2012, 11, 2014−2031

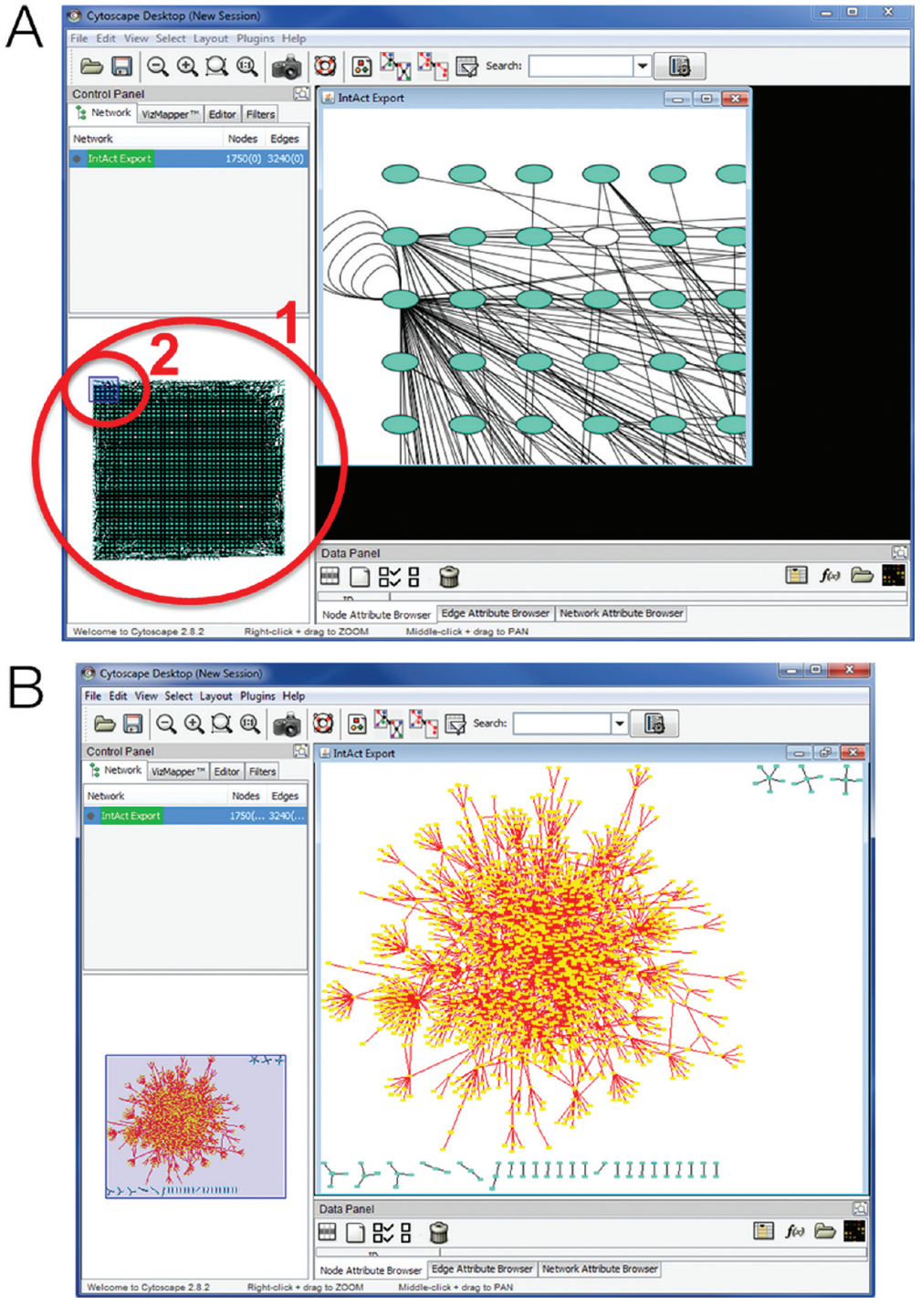


Figure 6. Interaction network imported from IntAct into Cytoscape 2.8.2. (A) Interaction network just imported into Cytoscape. The whole network is depicted in the navigation box in the Control Panel (1). The default layout is for all the nodes to be arranged in a grid. The area magnified in the main window may be navigated with the help of the blue square (2) in the Control Panel that represents the area of the network that is shown in the main window. (B) Same network has been laid out using a force-directed algorithm. There is a single large cluster in the center of the screen with a number of orphan interactions around the periphery of the screen: these are interactions that do not link to any other proteins in the network and consist of only two or three proteins each. The largest cluster in the network has been selected: selected nodes and edges are highlighted in yellow and red, respectively.

The table displays the most over-represented terms sorted in with the smallest *p*-values on top. In this table we see a list of GO terms (with their names and GO-IDs) and the uncorrected *p*-value and corrected *p*-value (the default correction is Benjamini-Hochberg[93]). Apart from that, total frequency values and a list of corresponding proteins (listed under the title "genes") are listed for each term. Since the list is sorted just by *p*-value, many general terms, (less descriptive terms) rise to the top of the table, making it difficult to see the more specific terms that are more useful. If you clicked the "save" option in the BiNGO setup window, then this table is already saved to file. If not, then you will need to copy and paste these results

dx.doi.org/10.1021/pr201211w | *J. Proteome Res.* 2012, 11, 2014–2031

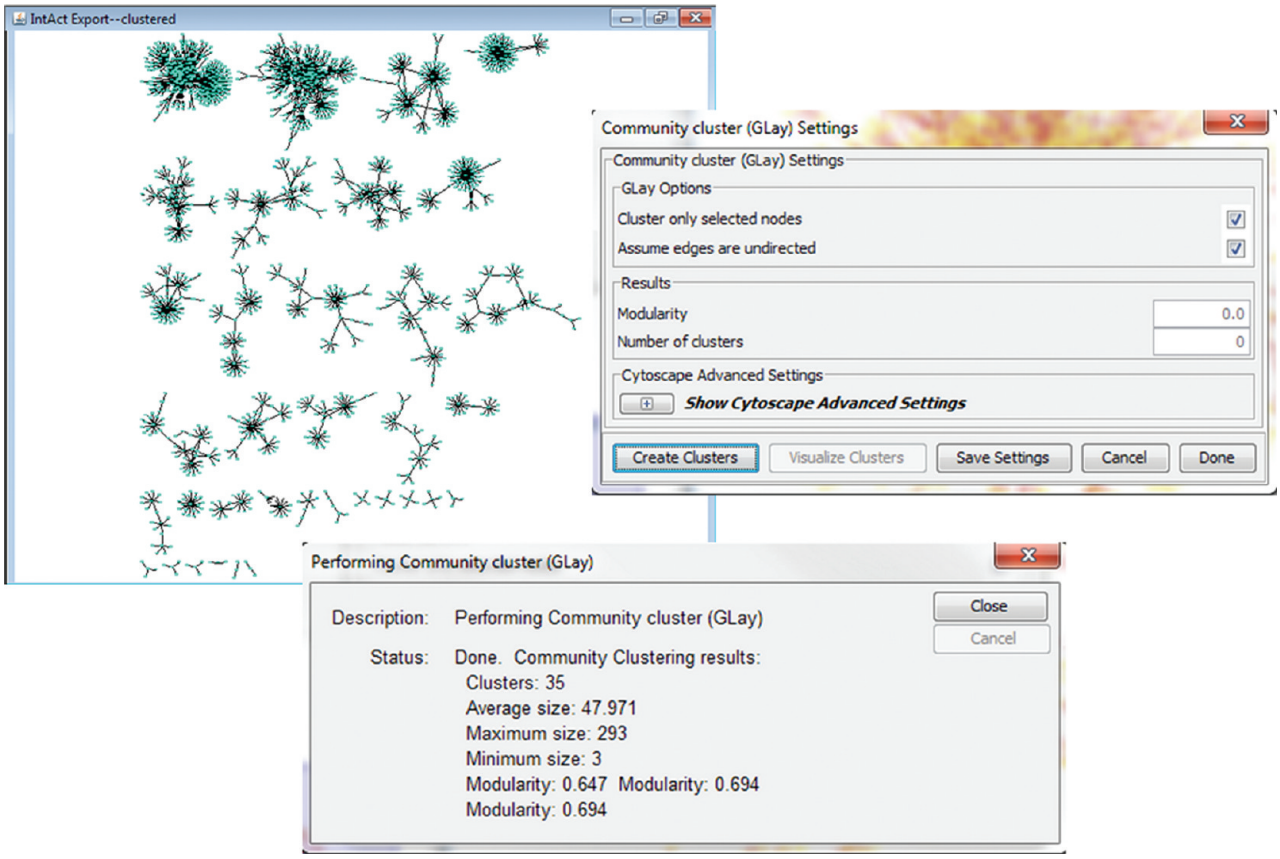


Figure 7. clusterMaker 1.9 plugin for Cytoscape. Community cluster (GLay) window. The settings window (top right) is used first to create the clusters. After the clustering analysis has been performed, an information window (bottom) pops up, reporting details such as the number of clusters created and their size. Go back to the settings window and press the "Visualize Clusters" button to make clusterMaker display the clusters in a new network window (top left).

into an Excel file (or similar). The data in this table is *not* saved as part of a Cytoscape session file and you will lose this data if you do not save it separately.

The other representation of the results is a graphical depiction of the enriched GO terms in the form of a network. Each node is a GO term, and GO terms are linked by directed edges representing parent-to-child relationships. Nodes are colored by *p*-value (orange is $<5 \times 10^{-7}$ and yellow is 0.05) and the size of each node is proportional to the number of proteins annotated with that term (Figure 8A). The default layout is less easy to read, but we may take advantage of one of Cytoscape's tools to provide a user-friendlier representation.

Make sure the graphical representation of the BiNGO results is selected. Choose "Layouts" → "Cytoscape Layouts" → "Hierarchical layout". Gene ontologies are a directed acyclic graph: Cytoscape utilizes this topology to organize the BiNGO results graph so that more specific and informative terms float to the top, while general, less informative terms sink to the bottom (Figure 8B). Navigating through this view provides a more useful impression of what biological processes are present in this PPIN. When you find a term of interest, you may look it up in the table to see what proteins in the network were annotated with that term.

Combining BiNGO and clusterMaker Analysis

We may now use BiNGO to described each of the clusters that were found using Community Cluster algorithm and to locate functional modules within the network. In the example outlined in the accompanying PowerPoint presentation (slide 44) we can see how the proteins interacting in cluster number 4 are involved in metabolic processes related with RNA splicing and processing. BiNGO lists the proteins annotated with the term of interest. We may then highlight these proteins in our cluster or network views to improve our insight on their role within the network.

Once the most relevant functional clusters have been identified, there are additional strategies for highlighting the most relevant information contained in the network. Cytoscape provides a number of visualization features that may be combined to display complex information (e.g., coloring by expression fold-change, or displaying structural information). A comprehensive description of these features is beyond the scope of this tutorial: the interested reader is referred to the online tutorial available on the Cytoscape Web site.

The strategy outlined in this tutorial is but a rough sketch of a "real-life" analysis. Each analysis must be tailored to the research question. We have aimed only to provide the user with an overview of some of the most popular tools used in the analysis of PPINs.

■ CURRENT LIMITATIONS AND USEFUL WORKING LIMITS

Interactome data remains but a glimpse of the complete collection of interactions happening in a living cell. We have to take into account that such a collection is not a static entity, but an ever-changing, highly dynamic landscape of enormous complexity. The nature of interaction data is, as we have already pointed out, inherently artificial, and no technique currently provides us with a real-life vision of what is happening inside a

dx.doi.org/10.1021/pr201211w | *J. Proteome Res.* 2012, 11, 2014−2031

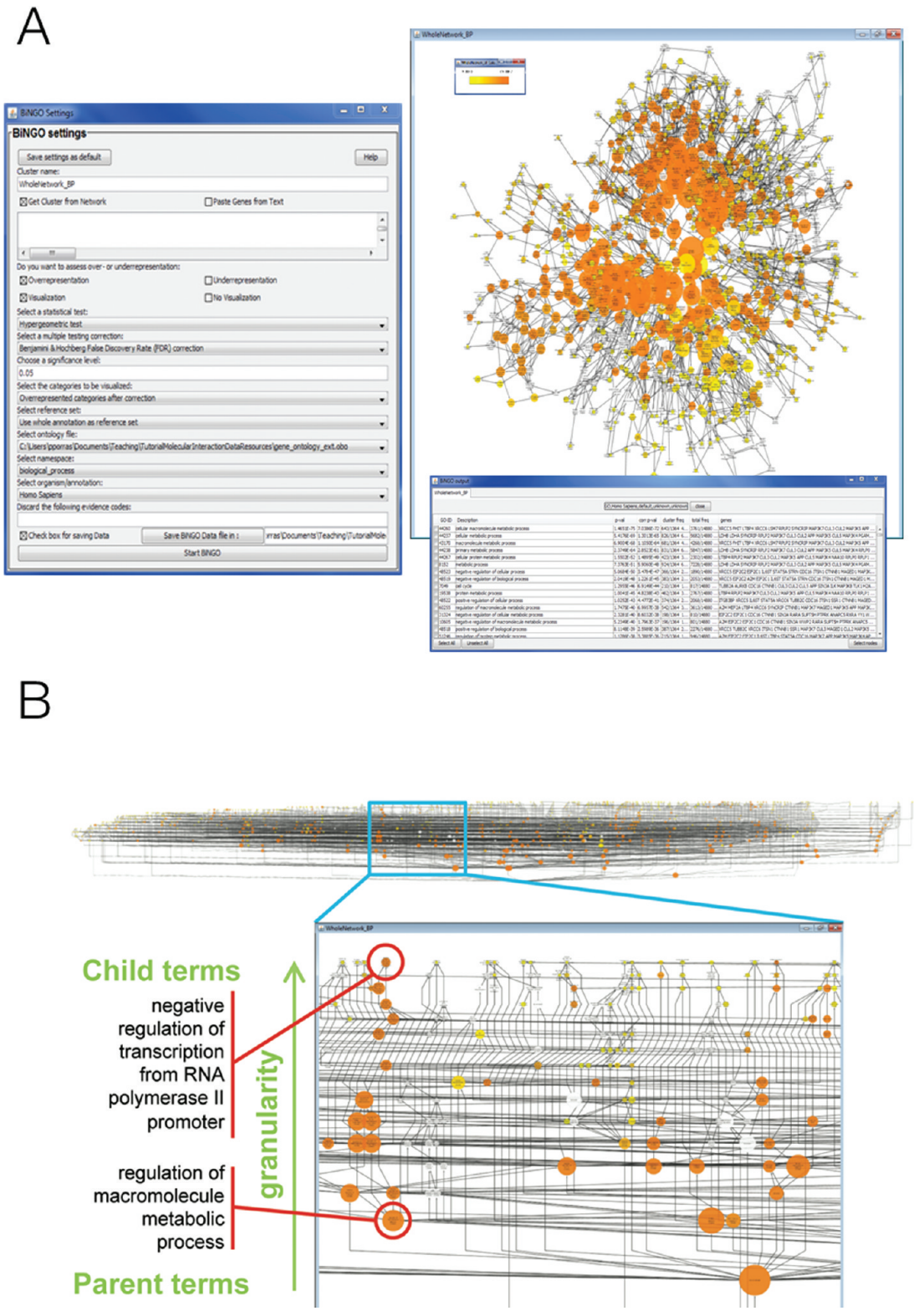


Figure 8. BiNGO 2.44 plugin for Cytoscape. (A) Initial BiNGO setup and results windows. The BiNGO setup window (left) allows the user to set parameters for an enrichment analysis. The user may choose to perform an over- or an under-representation analysis, set different p-value correction options, choose the significance level and decide the species, branch and file that is going to be used to check the ontology. Never use the ontology file that comes installed with BiNGO: instead, make sure you always use the most recent ontology file available from geneontology.org. (B) Graphical representation of the BiNGO analysis is best read using a hierarchical layout. The hierarchical layout places more general (less informative) terms at the bottom, while more specific (more informative) terms float to the top. This makes it easier for the user to identify terms that are most relevant to the goals of the experiment.

cell. A comparison of different interaction data sets will show that even when aimed at characterizing the same biological process, the overlap may be startlingly small.

Curated databases are generally considered to be more reliable than databases holding computationally predicted data or data from text-mining. However, curation is not free from

dx.doi.org/10.1021/pr201211w | J. Proteome Res. 2012, 11, 2014−2031

mistakes or misinterpretation of experimental data, and users need to bear this in mind. The number of publications containing molecular interaction data sets published daily is larger than the rate at which manual curation can keep up. Although the tutorial example has been limited to a curated database, the user may resort to data obtained from predictive or text-mining methods if the curated data is too sparse to be meaningful.

Clustering methods are still an active area of research and it is not possible to make dogmatic recommendations for one technique over any other. ClusterMaker therefore provides the user with a variety of clustering methods (autoSOME, Community cluster, MCL, MCODE, SCPS). Our opinion is that the main value in clustering a network is to increase the granularity of the annotation. We therefore advise that the user try a number of clustering algorithms and then select the one that achieves this.

BiNGO uses a hypergeometric test to calculate *p*-values for its GO terms. In nontechnical language, the hypergeometric test looks at the frequency with which that GO term appears, and the *p*-value consequently falls as the frequency increases. The frequency with which any GO term appears can never be greater than the total number of proteins in the network, so for small networks or clusters (<30 or so), it is possible for BiNGO to find no significant GO terms at all. For smaller networks, the user may find it necessary to increase the significance level from 0.05 to 0.10 (or higher!) to find any terms. For very small networks or lists, it is usually more helpful to examine the proteins one-by-one than to use the methods described here.

FUTURE DEVELOPMENTS

Although most databases focus on protein−protein interactions, the interactome clearly does not consist of proteins alone. Interactions in the living cell will also involve lipids,[94,95] carbohydrates,[96] nucleic acids,[97] and numerous small molecules (e.g., hormones, metabolites or drugs).[90,98] Databases vary widely in their coverage of these interactions. Although some databases, such as IntAct, do store this kind of data, the coverage for molecules other than proteins or gene products is still poor.

Integrating this and other types of data in a standardized fashion in interaction networks remains to be undertaken. Increasing effort is also being devoted to describing of dynamic networks that change as physiological variables change. New methods are being developed to display multiple parameters within the same network, for example, gene expression data[99] or structural information.[100]

A further consideration is description of the 'pathological' interactome (PPIs that only occur in certain disease states), in contrast to the "physiological".[101,102] For example, a gene mutation may result in an abnormal protein, and this abnormality may affect all the interactions made by the protein or just a subset of these interactions; or the protein may be aberrantly expressed in cell compartments or tissues where it is not normally present, with the resultant formation of abnormal protein complexes. Moreover, even in physiological situations, each cell type or situation does indeed have its specific set of molecular interactions. The study of the differences between particular biological contexts in contrast with the static interactome is called differential network biology and it is extensively reviewed by Ideker and Krogan.[103*]

Data standardization remains an important challenge to the exchange of information between and across fields. The development of unified formats is critical to integrate the enormous amount of data that has been and will be performed in the field, and to allow collaboration within the proteomics community. The IMEx consortium is one such effort, but standards require continual updating. The development of new methods and new types of data mean that even after a period of unification, databases will diverge as they develop new way of storing this information. These problems cannot be solved without a coordinated effort on the part of the proteomics community. As molecular interaction databases start to incorporate data from metabolic and genetic interaction networks, new standardization strategies will need to be developed that involve collaboration between researchers working in other fields.

Citation Information

The tools used in this tutorial are free to use. This does not mean that they are free to produce. All these tools represent the work of teams of developers and researchers supported by research grants, and renewal of these grants depends on evidence of use. Please cite the relevant papers if you use any of these tools in your published work.

ASSOCIATED CONTENT

Supporting Information

Supplementary PowerPoint presentation and list of targets file. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +44 (0)1223 494 675. Fax: +44 (0)1223 494 468. E-mail: orchard@ebi.ac.uk.

Author Contributions

‖These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

When presented after a reference callout in the text, * denotes key major reviews and ** denotes key historical/breakthrough articles.

ACKNOWLEDGMENTS

This work has been supported by grants awarded by the European Commision in the context of the PSIMEx consortium (contract number FP7-HEALTH-2007-223411), the APO-SYS consortium (contract number FP7-HEALTH-2007-200767) and by the Wellcome Trust (grant number 086532/Z/08/Z).

REFERENCES

(1) Sanchez, C.; Lachaize, C.; Janody, F.; Bellon, B.; Röder, L.; Euzenat, J.; Rechenmann, F.; Jacq, B. *Nucleic Acids Res.* **1999**, *27*, 89−94.

(2) Vidal, M.; Cusick, M. E.; Barabási, A.-L. *Cell* **2011**, *144*, 986−998.

(3) Vidal, M. *FEBS Lett.* **2009**, *583*, 3891−3894.

(4) Stelzl, U.; Wanker, E. *Curr. Opin. Chem. Biol.* **2006**, *10*, 551−558.

(5) Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J. M. *Nature* **2000**, *403*, 623−627.

(6) Braun, P.; Tasan, M.; Dreze, M.; Barrios-Rodiles, M.; Lemmens, I.; Yu, H.; Sahalie, J. M.; Murray, R. R.; Roncari, L.; de Smet, A.-S.; Venkatesan, K.; Rual, J.-F.; Vandenhaute, J.; Cusick, M. E.; Pawson, T.;

Hill, D. E.; Tavernier, J.; Wrana, J. L.; Roth, F. P.; Vidal, M. *Nat. Methods* **2008**, *6*, 91−97.
(7) Vidal, M.; Braun, P.; Chen, E.; Boeke, J. D.; Harlow, E. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 10321−10326.
(8) Fromont-Racine, M.; Rain, J. C.; Legrain, P. *Nat. Genet.* **1997**, *16*, 277−282.
(9) Finley, R. L. J.; Brent, R. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 12980−12984.
(10) Bartel, P. L.; Roecklein, J. A.; SenGupta, D.; Fields, S. *Nat. Genet.* **1996**, *12*, 72−77.
(11) Lim, J.; Hao, T.; Shaw, C.; Patel, A. J.; Szabo, G.; Rual, J.-F.; Fisk, C. J.; Li, N.; Smolyar, A.; Hill, D. E.; Barabasi, A.-L.; Vidal, M.; Zoghbi, H. Y. *Cell* **2006**, *125*, 801−814.
(12) Colland, F.; Jacq, X.; Trouplin, V.; Mougin, C.; Groizeleau, C.; Hamburger, A.; Meil, A.; Wojcik, J.; Legrain, P.; Gauthier, J.-M. *Genome Res.* **2004**, *14*, 1324−1332.
(13) Suzuki, H.; Fukunishi, Y.; Kagawa, I.; Saito, R.; Oda, H.; Endo, T.; Kondo, S.; Bono, H.; Okazaki, Y.; Hayashizaki, Y. *Genome Res.* **2001**, *11*, 1758−1765.
(14) Albers, M.; Kranz, H.; Kober, I.; Kaiser, C.; Klink, M.; Suckow, J.; Kern, R.; Koegl, M. *Mol. Cell. Proteomics* **2005**, *4*, 205−213.
(15) Goehler, H.; Lalowski, M.; Stelzl, U.; Waelter, S.; Stroedicke, M.; Worm, U.; Droege, A.; Lindenberg, K. S.; Knoblich, M.; Haenig, C.; Herbst, M.; Suopanki, J.; Scherzinger, E.; Abraham, C.; Bauer, B.; Hasenbank, R.; Fritzsche, A.; Ludewig, A. H.; Büssow, K.; Buessow, K.; Coleman, S. H.; Gutekunst, C.-A.; Landwehrmeyer, B. G.; Lehrach, H.; Wanker, E. E. *Mol. Cell* **2004**, *15*, 853−865.
(16) Lehner, B.; Sanderson, C. M. *Genome Res.* **2004**, *14*, 1315−1323.
(17) Butland, G.; Peregrín-Alvarez, J. M.; Li, J.; Yang, W.; Yang, X.; Canadien, V.; Starostine, A.; Richards, D.; Beattie, B.; Krogan, N.; Davey, M.; Parkinson, J.; Greenblatt, J.; Emili, A. *Nature* **2005**, *433*, 531−537.
(18) Arifuzzaman, M.; Maeda, M.; Itoh, A.; Nishikata, K.; Takita, C.; Saito, R.; Ara, T.; Nakahigashi, K.; Huang, H.-C.; Hirai, A.; Tsuzuki, K.; Nakamura, S.; Altaf-Ul-Amin, M.; Oshima, T.; Baba, T.; Yamamoto, N.; Kawamura, T.; Ioka-Nakamichi, T.; Kitagawa, M.; Tomita, M.; Kanaya, S.; Wada, C.; Mori, H. *Genome Res.* **2006**, *16*, 686−691.
(19) Yu, H.; Braun, P.; Yildirim, M. A.; Lemmens, I.; Venkatesan, K.; Sahalie, J.; Hirozane-Kishikawa, T.; Gebreab, F.; Li, N.; Simonis, N.; Hao, T.; Rual, J.-F.; Dricot, A.; Vazquez, A.; Murray, R. R.; Simon, C.; Tardivo, L.; Tam, S.; Svrzikapa, N.; Fan, C.; de Smet, A.-S.; Motyl, A.; Hudson, M. E.; Park, J.; Xin, X.; Cusick, M. E.; Moore, T.; Boone, C.; Snyder, M.; Roth, F. P.; Barabasi, A.-L.; Tavernier, J.; Hill, D. E.; Vidal, M. *Science* **2008**, *322*, 104−110.
(20) Gavin, A.-C.; Bösche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A.-M.; Cruciat, C.-M.; Remor, M.; Höfert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M.-A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G. *Nature* **2002**, *415*, 141−147.
(21) Gavin, A.-C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche, M.; Marzioch, M.; Rau, C.; Jensen, L. J.; Bastuck, S.; Dümpelfeld, B.; Edelmann, A.; Heurtier, M.-A.; Hoffman, V.; Hoefert, C.; Klein, K.; Hudak, M.; Michon, A.-M.; Schelder, M.; Schirle, M.; Remor, M.; Rudi, T.; Hooper, S.; Bauer, A.; Bouwmeester, T.; Casari, G.; Drewes, G.; Neubauer, G.; Rick, J. M.; Kuster, B.; Bork, P.; Russell, R. B.; Superti-Furga, G. *Nature* **2006**, *440*, 631−636.
(22) Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G. D.; Moore, L.; Adams, S.-L.; Millar, A.; Taylor, P.; Bennett, K.; Boutilier, K.; Yang, L.; Wolting, C.; Donaldson, I.; Schandorff, S.; Shewnarane, J.; Vo, M.; Taggart, J.; Goudreault, M.; Muskat, B.; Alfarano, C.; Dewar, D.; Lin, Z.; Michalickova, K.; Willems, A. R.; Sassi, H.; Nielsen, P. A.; Rasmussen, K. J.; Andersen, J. R.; Johansen, L. E.; Hansen, L. H.; Jespersen, H.; Podtelejnikov, A.; Nielsen, E.; Crawford, J.; Poulsen, V.; Sørensen, B. D.; Matthiesen, J.; Hendrickson, R. C.; Gleeson, F.; Pawson, T.; Moran, M. F.; Durocher, D.; Mann, M.; Hogue, C. W. V.; Figeys, D.; Tyers, M. *Nature* **2002**, *415*, 180−183.
(23) Krogan, N. J.; Cagney, G.; Yu, H.; Zhong, G.; Guo, X.; Ignatchenko, A.; Li, J.; Pu, S.; Datta, N.; Tikuisis, A. P.; Punna, T.; Peregrín-Alvarez, J. M.; Shales, M.; Zhang, X.; Davey, M.; Robinson, M. D.; Paccanaro, A.; Bray, J. E.; Sheung, A.; Beattie, B.; Richards, D. P.; Canadien, V.; Lalev, A.; Mena, F.; Wong, P.; Starostine, A.; Canete, M. M.; Vlasblom, J.; Wu, S.; Orsi, C.; Collins, S. R.; Chandran, S.; Haw, R.; Rilstone, J. J.; Gandi, K.; Thompson, N. J.; Musso, G.; St Onge, P.; Ghanny, S.; Lam, M. H. Y.; Butland, G.; Altaf-Ul, A. M.; Kanaya, S.; Shilatifard, A.; O'Shea, E.; Weissman, J. S.; Ingles, C. J.; Hughes, T. R.; Parkinson, J.; Gerstein, M.; Wodak, S. J.; Emili, A.; Greenblatt, J. F. *Nature* **2006**, *440*, 637−643.
(24) Giot, L.; Bader, J. S.; Brouwer, C.; Chaudhuri, A.; Kuang, B.; Li, Y.; Hao, Y. L.; Ooi, C. E.; Godwin, B.; Vitols, E.; Vijayadamodar, G.; Pochart, P.; Machineni, H.; Welsh, M.; Kong, Y.; Zerhusen, B.; Malcolm, R.; Varrone, Z.; Collis, A.; Minto, M.; Burgess, S.; McDaniel, L.; Stimpson, E.; Spriggs, F.; Williams, J.; Neurath, K.; Ioime, N.; Agee, M.; Voss, E.; Furtak, K.; Renzulli, R.; Aanensen, N.; Carrolla, S.; Bickelhaupt, E.; Lazovatsky, Y.; DaSilva, A.; Zhong, J.; Stanyon, C. A.; Finley, R. L.; White, K. P.; Braverman, M.; Jarvie, T.; Gold, S.; Leach, M.; Knight, J.; Shimkets, R. A.; McKenna, M. P.; Chant, J.; Rothberg, J. M. *Science (New York, N.Y.)* **2003**, *302*, 1727−1736.
(25) Walhout, A. J.; Sordella, R.; Lu, X.; Hartley, J. L.; Temple, G. F.; Brasch, M. A.; Thierry-Mieg, N.; Vidal, M. *Science (New York, N.Y.)* **2000**, *287*, 116−122.
(26) Reboul, J.; Vaglio, P.; Rual, J.-F.; Lamesch, P.; Martinez, M.; Armstrong, C. M.; Li, S.; Jacotot, L.; Bertin, N.; Janky, R.; Moore, T.; Hudson, J. R.; Hartley, J. L.; Brasch, M. A.; Vandenhaute, J.; Boulton, S.; Endress, G. A.; Jenna, S.; Chevet, E.; Papasotiropoulos, V.; Tolias, P. P.; Ptacek, J.; Snyder, M.; Huang, R.; Chance, M. R.; Lee, H.; Doucette-Stamm, L.; Hill, D. E.; Vidal, M. *Nat. Genet.* **2003**, *34*, 35−41.
(27) Li, S.; Armstrong, C. M.; Bertin, N.; Ge, H.; Milstein, S.; Boxem, M.; Vidalain, P.-O.; Han, J.-D. J.; Chesneau, A.; Hao, T.; Goldberg, D. S.; Li, N.; Martinez, M.; Rual, J.-F.; Lamesch, P.; Xu, L.; Tewari, M.; Wong, S. L.; Zhang, L. V.; Berriz, G. F.; Jacotot, L.; Vaglio, P.; Reboul, J.; Hirozane-Kishikawa, T.; Li, Q.; Gabel, H. W.; Elewa, A.; Baumgartner, B.; Rose, D. J.; Yu, H.; Bosak, S.; Sequerra, R.; Fraser, A.; Mango, S. E.; Saxton, W. M.; Strome, S.; Van Den Heuvel, S.; Piano, F.; Vandenhaute, J.; Sardet, C.; Gerstein, M.; Doucette-Stamm, L.; Gunsalus, K. C.; Harper, J. W.; Cusick, M. E.; Roth, F. P.; Hill, D. E.; Vidal, M. *Science (New York, N.Y.)* **2004**, *303*, 540−543.
(28) Rual, J.-F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, G. F.; Gibbons, F. D.; Dreze, M.; Ayivi-Guedehoussou, N.; Klitgord, N.; Simon, C.; Boxem, M.; Milstein, S.; Rosenberg, J.; Goldberg, D. S.; Zhang, L. V.; Wong, S. L.; Franklin, G.; Li, S.; Albala, J. S.; Lim, J.; Fraughton, C.; Llamosas, E.; Cevik, S.; Bex, C.; Lamesch, P.; Sikorski, R. S.; Vandenhaute, J.; Zoghbi, H. Y.; Smolyar, A.; Bosak, S.; Sequerra, R.; Doucette-Stamm, L.; Cusick, M. E.; Hill, D. E.; Roth, F. P.; Vidal, M. *Nature* **2005**, *437*, 1173−1178.
(29) Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, F. H.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S. *Cell* **2005**, *122*, 957−968.
(30) Ewing, R. M.; Chu, P.; Elisma, F.; Li, H.; Taylor, P.; Climie, S.; McBroom-Cerajewski, L.; Robinson, M. D.; O'Connor, L.; Li, M.; Taylor, R.; Dharsee, M.; Ho, Y.; Heilbut, A.; Moore, L.; Zhang, S.; Ornatsky, O.; Bukhman, Y. V.; Ethier, M.; Sheng, Y.; Vasilescu, J.; Abu-Farha, M.; Lambert, J.-P.; Duewel, H. S.; Stewart, I. I.; Kuehl, B.; Hogue, K.; Colwill, K.; Gladwish, K.; Muskat, B.; Kinach, R.; Adams, S.-L.; Moran, M. F.; Morin, G. B.; Topaloglou, T.; Figeys, D. *Mol. Syst. Biol.* **2007**, 3.
(31) Venkatesan, K.; Rual, J.-F.; Vazquez, A.; Stelzl, U.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Zenkner, M.; Xin, X.; Goh, K.-I.; Yildirim, M. A.; Simonis, N.; Heinzmann, K.; Gebreab, F.; Sahalie, J. M.; Cevik, S.; Simon, C.; de Smet, A.-S.; Dann, E.; Smolyar, A.; Vinayagam, A.; Yu, H.; Szeto, D.; Borick, H.; Dricot, A.; Klitgord, N.; Murray, R. R.; Lin, C.; Lalowski, M.; Timm, J.; Rau, K.; Boone, C.; Braun, P.; Cusick, M. E.; Roth, F. P.; Hill, D. E.; Tavernier, J.; Wanker, E. E.; Barabasi, A.-L.; Vidal, M. *Nat. Methods* **2009**, *6*, 83−90.

(32) Stumpf, M. P. H; Thorne, T.; de Silva, E.; Stewart, R.; An, H. J.; Lappe, M.; Wiuf, C. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6959−6964.
(33) Wass, M. N.; David, A.; Sternberg, M. J. *Curr. Opin. Struct. Biol.* **2011**, *21*, 382−390.
(34) Hoffmann, R.; Valencia, A. *Nat. Genet.* **2004**, *36*, 664.
(35) Barbosa-Silva, A.; Fontaine, J.-F.; Donnard, E. R.; Stussi, F.; Ortega, J. M.; Andrade-Navarro, M. A. *BMC Bioinform.* **2011**, *12*, 435.
(36) Endy, D.; Brent, R. *Nature* **2001**, *409*, 391−395.
(37) Batagelj, V.; Mrvar, A. *Graph Drawing Software* **2003**, *41*, 871.
(38) Su, G.; Kuchinsky, A.; Morris, J. H.; States, D. J.; Meng, F. *Bioinformatics* **2010**, *26*, 3135−3137.
(39) Bader, G. D.; Hogue, C. W. V. *BMC Bioinform.* **2003**, *4*, 2.
(40) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. *Genome Res.* **2003**, *13*, 2498−2504.
(41) Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P.-L.; Ideker, T. *Bioinformatics (Oxford, England)* **2011**, *27*, 431−432.
(42) Fields, S.; Song, O. *Nature* **1989**, *340*, 245−246.
(43) Suter, B.; Kittanakom, S.; Stagljar, I. *Curr. Opin. Biotechnol.* **2008**, *19*, 316−323.
(44) Rigaut, G.; Shevchenko, A.; Rutz, B.; Wilm, M.; Mann, M.; Séraphin, B. *Nat. Biotechnol.* **1999**, *17*, 1030−1032.
(45) Bauer, A.; Kuster, B. *Eur. J. Biochem.* **2003**, *270*, 570−578.
(46) Pflieger, D.; Gonnet, F.; de la Fuente van Bentem, S.; Hirt, H.; de la Fuente, A. *Mass Spectrom. Rev.* **2011**, *30*, 268−297.
(47) Barrios-Rodiles, M.; Brown, K. R.; Ozdamar, B.; Bose, R.; Liu, Z.; Donovan, R. S.; Shinjo, F.; Liu, Y.; Dembowy, J.; Taylor, I. W.; Luga, V.; Przulj, N.; Robinson, M.; Suzuki, H.; Hayashizaki, Y.; Jurisica, I.; Wrana, J. L. *Science (New York, N.Y.)* **2005**, *307*, 1621−1625.
(48) Eyckerman, S.; Lemmens, I.; Lievens, S.; Van der Heyden, J.; Verhee, A.; Vandekerckhove, J.; Tavernier, J. *Science's STKE: Signal Transduction Knowledge Environment* **2002**, *2002*, pl18.
(49) Ciruela, F. *Current Opinion in Biotechnology* **2008**, *19*, 338−343.
(50) Xia, Z.; Rao, J. *Current Opinion in Biotechnology* **2009**, *20*, 37−44.
(51) Petschnigg, J.; Snider, J.; Stagljar, I. *Current Opinion in Biotechnology* **2011**, *22*, 50−58.
(52) Stubbs, M. T.; Editors-in-Chief: John, B. Taylor; David J. Triggle In *Comprehensive Medicinal Chemistry II;* Elsevier: Oxford, 2007; pp. 449−472.
(53) Pan, Q.; Shai, O.; Lee, L. J.; Frey, B. J.; Blencowe, B. J. *Nature genetics* **2008**, *40*, 1413−1415.
(54) Magrane, M.; Consortium, U. *Database* **2011**, *2011*, bar009.
(55) Pruitt, K. D.; Tatusova, T.; Klimke, W.; Maglott, D. R. *Nucleic Acids Res.* **2009**, *37*, D32−36.
(56) Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. *Proteomics* **2004**, *4*, 1985−1988.
(57) Griss, J.; Martín, M.; O'Donovan, C.; Apweiler, R.; Hermjakob, H.; Vizcaíno, J. A. *Proteomics* **2011**, *11*, 4434−4438.
(58) Turinsky, A. L.; Razick, S.; Turner, B.; Donaldson, I. M.; Wodak, S. J. *J. Biol. Databases Curation* **2010**, *2010*, baq026.
(59) Cusick, M. E.; Yu, H.; Smolyar, A.; Venkatesan, K.; Carvunis, A.-R.; Simonis, N.; Rual, J.-F.; Borick, H.; Braun, P.; Dreze, M.; Vandenhaute, J.; Galli, M.; Yazaki, J.; Hill, D. E.; Ecker, J. R.; Roth, F. P.; Vidal, M. *Nat. Methods* **2009**, *6*, 39−46.
(60) Salwinski, L.; Licata, L.; Winter, A.; Thorneycroft, D.; Khadake, J.; Ceol, A.; Aryamontri, A. C.; Oughtred, R.; Livstone, M.; Boucher, L.; Botstein, D.; Dolinski, K.; Berardini, T.; Huala, E.; Tyers, M.; Eisenberg, D.; Cesareni, G.; Hermjakob, H. *Nat. Methods* **2009**, *6*, 860−861.
(61) De Las Rivas, J.; Fontanillo, C. *PLoS Comput. Biol.* **2010**, *6*, e1000807.
(62) Hernandez-Toro, J.; Prieto, C.; De las Rivas, J. *Bioinformatics (Oxford, England)* **2007**, *23*, 2495−2497.
(63) Wu, J.; Vallenius, T.; Ovaska, K.; Westermarck, J.; Makela, T. P.; Hautaniemi, S. *Nat. Methods* **2009**, *6*, 75−77.
(64) Tarcea, V. G.; Weymouth, T.; Ade, A.; Bookvich, A.; Gao, J.; Mahavisno, V.; Wright, Z.; Chapman, A.; Jayapandian, M.; Ozgur, A.; Tian, Y.; Cavalcoli, J.; Mirel, B.; Patel, J.; Radev, D.; Athey, B.; States, D.; Jagadish, H. V. *Nucleic Acids Res.* **2009**, *37*, D642−D646.
(65) McDowall, M. D.; Scott, M. S.; Barton, G. J. *Nucleic Acids Res.* **2009**, *37*, D651−D656.
(66) Goll, J.; Rajagopala, S. V.; Shiau, S. C.; Wu, H.; Lamb, B. T.; Uetz, P. *Bioinformatics (Oxford, England)* **2008**, *24*, 1743−1744.
(67) Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; Jensen, L. J.; von Mering, C. *Nucleic Acids Res.* **2011**, *39*, D561−568.
(68) Brown, K. R.; Jurisica, I. *Bioinformatics (Oxford, England)* **2005**, *21*, 2076−2082.
(69) Chaurasia, G.; Iqbal, Y.; Hänig, C.; Herzel, H.; Wanker, E. E.; Futschik, M. E. *Nucleic Acids Res.* **2007**, *35*, D590−594.
(70) Orchard, S.; Kerrien, S.; Jones, P.; Ceol, A.; Chatr-Aryamontri, A.; Salwinski, L.; Nerothin, J.; Hermjakob, H. *Proteomics* **2007**, *7*, 28−34.
(71) Orchard, S.; Kerrien, S.; Abbani, S.; Aranda, B.; Bhate, J.; Bidwell, S.; Bridge, A.; Briganti, L.; Brinkman, F. S. L.; Cesareni, G.; Chatr-aryamontri, A.; Chautard, E.; Chen, C.; Dumousseau, M.; Eisenberg, D.; Goll, J.; Hancock, R. E. W.; Hannick, L. I.; Jurisica, I.; Khadake, J.; Lynn, D. J.; Mahadevan, U.; Perfetto, L.; Raghunath, A.; Ricard-Blum, S.; Roechert, B.; Salwinski, L.; Stümpflen, V.; Tyers, M.; Uetz, P.; Xenarios, I.; Hermjakob, H. *Nature Methods* **2012**, in press.
(72) Orchard, S.; Salwinski, L.; Kerrien, S.; Montecchi-Palazzi, L.; Oesterheld, M.; Stümpflen, V.; Ceol, A.; Chatr-aryamontri, A.; Armstrong, J.; Woollard, P.; Salama, J. J.; Moore, S.; Wojcik, J.; Bader, G. D.; Vidal, M.; Cusick, M. E.; Gerstein, M.; Gavin, A.-C.; Superti-Furga, G.; Greenblatt, J.; Bader, J.; Uetz, P.; Tyers, M.; Legrain, P.; Fields, S.; Mulder, N.; Gilson, M.; Niepmann, M.; Burgoon, L.; Rivas, J. D. L.; Prieto, C.; Perreau, V. M.; Hogue, C.; Mewes, H.-W.; Apweiler, R.; Xenarios, I.; Eisenberg, D.; Cesareni, G.; Hermjakob, H. *Nat. Biotechnol.* **2007**, *25*, 894−898.
(73) Kerrien, S.; Orchard, S.; Montecchi-Palazzi, L.; Aranda, B.; Quinn, A. F.; Vinod, N.; Bader, G. D.; Xenarios, I.; Wojcik, J.; Sherman, D.; Tyers, M.; Salama, J. J.; Moore, S.; Ceol, A.; Chatr-Aryamontri, A.; Oesterheld, M.; Stümpflen, V.; Salwinski, L.; Nerothin, J.; Cerami, E.; Cusick, M. E.; Vidal, M.; Gilson, M.; Armstrong, J.; Woollard, P.; Hogue, C.; Eisenberg, D.; Cesareni, G.; Apweiler, R.; Hermjakob, H. *BMC Biol.* **2007**, *5*, 44.
(74) Gehlenborg, N.; O'Donoghue, S. I.; Baliga, N. S.; Goesmann, A.; Hibbs, M. A.; Kitano, H.; Kohlbacher, O.; Neuweger, H.; Schneider, R.; Tenenbaum, D.; Gavin, A.-C. *Nat. Methods* **2010**, *7*, S56−68.
(75) Cline, M. S.; Smoot, M.; Cerami, E.; Kuchinsky, A.; Landys, N.; Workman, C.; Christmas, R.; Avila-Campilo, I.; Creech, M.; Gross, B.; Hanspers, K.; Isserlin, R.; Kelley, R.; Killcoyne, S.; Lotia, S.; Maere, S.; Morris, J.; Ono, K.; Pavlovic, V.; Pico, A. R.; Vailaya, A.; Wang, P.-L.; Adler, A.; Conklin, B. R.; Hood, L.; Kuiper, M.; Sander, C.; Schmulevich, I.; Schwikowski, B.; Warner, G. J.; Ideker, T.; Bader, G. D. *Nat. Protocols* **2007**, *2*, 2366−2382.
(76) Brohée, S.; van Helden, J. *BMC Bioinformatics* **2006**, *7*, 488.
(77) Newman, M. E. J.; Girvan, M. *Phys. Rev., E* **2004**, *69*, 026113.
(78) Wang, J.; Li, M.; Deng, Y.; Pan, Y. *BMC Genomics* **2010**, *11* (Suppl3), S10.
(79) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. *Nat. Genet.* **2000**, *25*, 25−29.
(80) Croft, D.; O'Kelly, G.; Wu, G.; Haw, R.; Gillespie, M.; Matthews, L.; Caudy, M.; Garapati, P.; Gopinath, G.; Jassal, B.; Jupe, S.; Kalatskaya, I.; Mahajan, S.; May, B.; Ndegwa, N.; Schmidt, E.; Shamovsky, V.; Yung, C.; Birney, E.; Hermjakob, H.; D'Eustachio, P.; Stein, L. *Nucleic Acids Res.* **2011**, *39*, D691−697.
(81) Haw, R.; Hermjakob, H.; D'Eustachio, P.; Stein, L. *Proteomics* **2011**, *11*, 3598−3613.
(82) Kanehisa, M.; Goto, S. *Nucleic Acids Res.* **2000**, *28*, 27−30.
(83) Pico, A. R.; Kelder, T.; van Iersel, M. P.; Hanspers, K.; Conklin, B. R.; Evelo, C. *PLoS Biol.* **2008**, *6*, e184.

(84) Soler-López, M.; Zanzoni, A.; Lluís, R.; Stelzl, U.; Aloy, P. *Genome Res.* **2011**, *21*, 364−376.
(85) Sowa, M. E.; Bennett, E. J.; Gygi, S. P.; Harper, J. W. *Cell* **2009**, *138*, 389−403.
(86) Côté, R. G.; Jones, P.; Martens, L.; Kerrien, S.; Reisinger, F.; Lin, Q.; Leinonen, R.; Apweiler, R.; Hermjakob, H. *BMC Bioinform.* **2007**, *8*, 401.
(87) Morris, J. H.; Apeltsin, L.; Newman, A. M.; Baumbach, J.; Wittkop, T.; Su, G.; Bader, G. D.; Ferrin, T. E. *BMC Bioinformatics*, **2011**, *12*, 436.
(88) Maere, S.; Heymans, K.; Kuiper, M. *Bioinformatics (Oxford, England)* **2005**, *21*, 3448−3449.
(89) Vizcaíno, J. A.; Côté, R.; Reisinger, F.; Foster, J. M.; Mueller, M.; Rameseder, J.; Hermjakob, H.; Martens, L. *Proteomics* **2009**, *9*, 4276−4283.
(90) Bantscheff, M.; Hopf, C.; Savitski, M. M.; Dittmann, A.; Grandi, P.; Michon, A.-M.; Schlegl, J.; Abraham, Y.; Becher, I.; Bergamini, G.; Boesche, M.; Delling, M.; Dümpelfeld, B.; Eberhard, D.; Huthmacher, C.; Mathieson, T.; Poeckel, D.; Reader, V.; Strunk, K.; Sweetman, G.; Kruse, U.; Neubauer, G.; Ramsden, N. G.; Drewes, G. *Nat. Biotechnol.* **2011**, *29*, 255−265.
(91) Aranda, B.; Achuthan, P.; Alam-Faruque, Y.; Armean, I.; Bridge, A.; Derow, C.; Feuermann, M.; Ghanbarian, A. T.; Kerrien, S.; Khadake, J.; Kerssemakers, J.; Leroy, C.; Menden, M.; Michaut, M.; Montecchi-Palazzi, L.; Neuhauser, S. N.; Orchard, S.; Perreau, V.; Roechert, B.; van Eijk, K.; Hermjakob, H. *Nucleic Acids Res.* **2010**, *38*, D525−31.
(92) Hakes, L.; Robertson, D. L.; Oliver, S. G.; Lovell, S. C. *Comp. Funct. Genomics* **2007**, 49356.
(93) Benjamini, Y. *Ann. Stat.* **2001**, *29*, 1165−1188.
(94) Lu, N.; Shen, Q.; Mahoney, T. R.; Liu, X.; Zhou, Z. *Mol. Biol. Cell* **2011**, *22*, 354−374.
(95) Sasaki, T.; Demura, M.; Kato, N.; Mukai, Y. *Biochemistry* **2011**, *50*, 2283−2290.
(96) Park, S.; Shin, I. *Angew. Chem., Int. Ed.* **2002**, *41*, 3180−3182.
(97) Pagano, J. M.; Clingman, C. C.; Ryder, S. P. *RNA* **2011**, *17*, 14−20.
(98) Lazaridis, I.; Charalampopoulos, I.; Alexaki, V.-I.; Avlonitis, N.; Pediaditakis, I.; Efstathopoulos, P.; Calogeropoulou, T.; Castanas, E.; Gravanis, A. *PLoS Biol.* **2011**, *9*, e1001051.
(99) Zheng, S.; Tansey, W. P.; Hiebert, S. W.; Zhao, Z. *BMC Med. Genomics* **2011**, *4*, 62.
(100) Stein, A.; Mosca, R.; Aloy, P. *Curr. Opin. Struct. Biol.* **2011**, *21*, 200−208.
(101) Dreze, M.; Charloteaux, B.; Milstein, S.; Vidalain, P.-O.; Yildirim, M. A.; Zhong, Q.; Svrzikapa, N.; Romero, V.; Laloux, G.; Brasseur, R.; Vandenhaute, J.; Boxem, M.; Cusick, M. E.; Hill, D. E.; Vidal, M. *Nat. Methods* **2009**, *6*, 843−849.
(102) Zhong, Q.; Simonis, N.; Li, Q.-R.; Charloteaux, B.; Heuze, F.; Klitgord, N.; Tam, S.; Yu, H.; Venkatesan, K.; Mou, D.; Swearingen, V.; Yildirim, M. A.; Yan, H.; Dricot, A.; Szeto, D.; Lin, C.; Hao, T.; Fan, C.; Milstein, S.; Dupuy, D.; Brasseur, R.; Hill, D. E.; Cusick, M. E.; Vidal, M. *Mol. Syst. Biol.* **2009**, *5*, 321.
(103) Ideker, T.; Krogan, N. J. *Mol. Syst. Biol.* **2012**, *8*, 565.
(104) Kleiger, G.; Saha, A.; Lewis, S.; Kuhlman, B.; Deshaies, R. J. *Cell* **2009**, *139*, 957−968.
(105) Ceol, A.; Chatr Aryamontri, A.; Licata, L.; Peluso, D.; Briganti, L.; Perfetto, L.; Castagnoli, L.; Cesareni, G. *Nucleic Acids Res.* **2010**, *38*, D532−539.
(106) Salwinski, L. *Nucleic Acids Res.* **2004**, *32*, D449−451.
(107) Chautard, E.; Fatoux-Ardore, M.; Ballut, L.; Thierry-Mieg, N.; Ricard-Blum, S. *Nucleic Acids Res.* **2011**, *39*, D235−240.
(108) Lynn, D. J.; Chan, C.; Naseer, M.; Yau, M.; Lo, R.; Sribnaia, A.; Ring, G.; Que, J.; Wee, K.; Winsor, G. L.; Laird, M. R.; Breuer, K.; Foroushani, A. K.; Brinkman, F. S. L.; Hancock, R. E. W. *BMC Syst. Biol.* **2010**, *4*, 117.
(109) Pagel, P.; Kovac, S.; Oesterheld, M.; Brauner, B.; Dunger-Kaltenbach, I.; Frishman, G.; Montrone, C.; Mark, P.; Stümpflen, V.; Mewes, H.-W.; Ruepp, A.; Frishman, D. *Bioinformatics (Oxford, England)* **2005**, *21*, 832−834.
(110) Güldener, U.; Münsterkötter, M.; Oesterheld, M.; Pagel, P.; Ruepp, A.; Mewes, H.-W.; Stümpflen, V. *Nucleic Acids Res.* **2006**, *34*, D436−441.
(111) Alfarano, C.; Andrade, C. E.; Anthony, K.; Bahroos, N.; Bajec, M.; Bantoft, K.; Betel, D.; Bobechko, B.; Boutilier, K.; Burgess, E.; Buzadzija, K.; Cavero, R.; D'Abreo, C.; Donaldson, I.; Dorairajoo, D.; Dumontier, M. J.; Dumontier, M. R.; Earles, V.; Farrall, R.; Feldman, H.; Garderman, E.; Gong, Y.; Gonzaga, R.; Grytsan, V.; Gryz, E.; Gu, V.; Haldorsen, E.; Halupa, A.; Haw, R.; Hrvojic, A.; Hurrell, L.; Isserlin, R.; Jack, F.; Juma, F.; Khan, A.; Kon, T.; Konopinsky, S.; Le, V.; Lee, E.; Ling, S.; Magidin, M.; Moniakis, J.; Montojo, J.; Moore, S.; Muskat, B.; Ng, I.; Paraiso, J. P.; Parker, B.; Pintilie, G.; Pirone, R.; Salama, J. J.; Sgro, S.; Shan, T.; Shu, Y.; Siew, J.; Skinner, D.; Snyder, K.; Stasiuk, R.; Strumpf, D.; Tuekam, B.; Tao, S.; Wang, Z.; White, M.; Willis, R.; Wolting, C.; Wong, S.; Wrong, A.; Xin, C.; Yao, R.; Yates, B.; Zhang, S.; Zheng, K.; Pawson, T.; Ouellette, B. F. F.; Hogue, C. W. V. *Nucleic Acids Res.* **2005**, *33*, D418−424.
(112) Stark, C.; Breitkreutz, B.-J.; Chatr-Aryamontri, A.; Boucher, L.; Oughtred, R.; Livstone, M. S.; Nixon, J.; Van Auken, K.; Wang, X.; Shi, X.; Reguly, T.; Rust, J. M.; Winter, A.; Dolinski, K.; Tyers, M. *Nucleic Acids Res.* **2011**, *39*, D698−704.
(113) Goel, R.; Muthusamy, B.; Pandey, A.; Prasad, T. S. K. *Mol. Biotechnol.* **2011**, *48*, 87−95.

NOTE ADDED AFTER ASAP PUBLICATION

This article was published ASAP on March 2, 2012. A new Supporting Information file containing IPI data sets in PDF and RTF formats has been added. The correct version was published on March 16, 2012.