Popular Computational Methods to Assess Multiprotein Complexes Derived From Label-Free Affinity Purification and Mass Spectrometry (AP-MS) Experiments*\(\S\)

Irina M. Armean‡, Kathryn S. Lilley‡¶, and Matthew W. B. Trotter§

Advances in sensitivity, resolution, mass accuracy, and throughput have considerably increased the number of protein identifications made via mass spectrometry. Despite these advances, state-of-the-art experimental methods for the study of protein-protein interactions yield more candidate interactions than may be expected biologically owing to biases and limitations in the experimental methodology. In silico methods, which distinguish between true and false interactions, have been developed and applied successfully to reduce the number of false positive results yielded by physical interaction assays. Such methods may be grouped according to: (1) the type of data used: methods based on experiment-specific measurements (e.g., spectral counts or identification scores) versus methods that extract knowledge encoded in external annotations (e.g., public interaction and functional categorisation databases); (2) the type of algorithm applied: the statistical description and estimation of physical protein properties versus predictive supervised machine learning or text-mining algorithms; (3) the type of protein relation evaluated: direct (binary) interaction of two proteins in a cocomplex versus probability of any functional relationship between two proteins (e.g., co-occurrence in a pathway, sub cellular compartment); and (4) initial motivation: elucidation of experimental data by evaluation versus prediction of novel protein-protein interaction, to be experimentally validated a posteriori. This work reviews several popular computational scoring methods and software platforms for protein-protein interactions evaluation according to their methodology, comparative strengths and weaknesses, data representation, accessibility, and availability. The scoring methods and platforms described include: CompPASS, SAINT, Decontaminator, MINT, IntAct, STRING, and FunCoup. References to related work are provided throughout in order to provide a concise but thorough introduction to a rapidly growing interdisciplinary field of investigation. Molecular & Cellular Proteomics 12: 10.1074/mcp.R112.019554, 1–13, 2013.

Proteins are extremely diverse in both structure and function. Despite this fact, most perform their function via direct interactions with other proteins. Many exist as part of multiprotein complexes ranging in size from two interacting partners to more than 81 in the case of the large ribosomal subunit (1, 2), some with multiple copies of the same protein forming large complexes with concerted functional activity (3). To fully understand cellular mechanisms at the molecular level, it is important to determine which proteins interact and how these interactions change in response to a highly dynamic cellular environment. Determination of multiprotein complex composition is not without its challenges. Protein complexes span a wide abundance range, and may be membrane bound or associated with different subcellular compartments. Characterization of multiprotein complexes is further confounded by varying affinities of proteins to their interacting partners, transient binding of partners, and differential interaction according to protein isoforms. Accordingly, experimental conditions must facilitate solubilization of complexes without destroying their contributing interactions. Contaminating proteins, co-solubilized during purification, must also be identified and discarded from genuine members of a multiprotein complex.

Most protein-protein interactions (PPI)\(^1\) recorded to date by the public databases collected in IRefIndex (4) have not been

\(^1\) The abbreviations used are: PPI, protein-protein interaction; AD, activating domain; AIN, autophagy interaction network; AP, affinity purification and mass spectrometric analysis; BD, DNA binding domain; BIND, the Biomolecular Interaction Network Database; BioCreative, Critical Assessment of Information Extraction systems in Biology; BioGrid, Biological General Repository for Interaction Databases; BLAST, basic local alignment search tool; COG, Clusters of Orthologous Groups; CompPASS, Comparative Proteomics Analysis Software Suite; DIP, Database of Interacting Proteins; DOM, domain associations; DUB, deubiquitinating enzymes; EEL, enhancer element locator; FP, false positive; FPR, false positive rate; FRET, fluorescence resonance energy transfer; FunCoup, Networks of functional coupling; GFP, green fluorescent protein; GIN, genetic interactions; GO, Gene Ontology; HCIP, high confidence candidate interacting proteins; HEK, Human Embryonic Kidney 293 cells; HPRD, Human Protein Reference Database; IMEX, International Molecular Exchange Consortium; INferAD, interaction network for ER-associated degradation; InnateDB, a knowledge resource for innate immunity interac-
detected by isolation of complexes from their biological source, but instead originate from studies using the yeast two-hybrid (Y2H) system (supplemental Table S1). Despite a growing trend in deposition of large protein interaction data sets from isolation studies and components identified by proteomics approaches, as described later in this publication (5, 6); Y2H studies have yielded a rich source of protein interaction data (supplemental Tables S1, S2) over the past three decades. Y2H determines physical interaction between two proteins (binary interactions) by means of transcriptional activity (7). The method employs yeast transcription factors, such as Gal4, with activating (AD) and DNA binding (BD) domains. The AD and BD domains can be fused respectively to a variety of prospective interacting partners, commonly referred to as baits and preys. If bait and prey interact, the two domains are brought together to form a functional transcription factor that indicates protein interaction via expression of a reporter gene. Developments of the approach include attempts to improve specificity, for example via use of cytokine receptor activation for greater physiological relevance (8). Regardless of the specific approach taken, libraries of baits and preys can be genetically engineered to test numerous combinations in large-scale studies. Despite the advantages of obtaining large data sets (9–11), the false positive rate (FPR) associated with this approach has been reported in the past to be greater than 50% in some instances (12–15). Recent studies have reduced error rates, in particular via the combination of additional methodologies such as LUMIER (16), MAPPIT (17), protein complementation assay (PCA) (18), and NAPPA (19), as shown by Vidal and coworkers (20). False positives (FP) arise from differences in genetic background, differences in quantitative measures of interaction, use of nonphysiological conditions, or extensive testing of protein partners that do not normally exist in the same cellular compartment (12, 15). Finally, although Y2H largely tests direct interactions between bait and prey, there may also be occasions in which interaction is mediated by a third yeast protein and thus bait and prey interact indirectly. For further reviews focused on the minimization of FP associated with direct binary protein interaction data (primarily Y2H), the reader is directed to the works of Vidal, Uetz, and Vazquez (21–23).

Protein microarrays also enable the identification of (binary) protein-protein interactions systematically on a genome scale. Bait proteins are individually expressed in yeast strains and subsequently purified. The purified proteins are arrayed on a slide and subsequently incubated with prey proteins labeled in some manner, such as tagging with a fluorescent marker, to monitor their interaction. Protein arrays have also been used to investigate interactions between proteins and small-molecules, kinase-substrate interactions, or antibody-mediated protein profiling. Observed limitations of this method concern aberrant protein folding and nonphysiologically relevant post-translational modification (PTM) status of the yeast-derived bait (24–26). In addition, there exists a plethora of biophysical experimental methods with which to investigate PPIs, including structural studies of protein complexes using x-ray crystallography and NMR, surface plasmon resonance (SPR), fluorescence resonance energy transfer (FRET), and isothermal titration calorimetry (ITC). Many of these approaches are reviewed in (27, 28).

Affinity purification coupled to mass spectrometry (AP-MS) represents a more recent approach that involves tagging bait proteins with an epitope before purification of both bait and interacting partners (prey) using antibody affinity. Proteins of the retrieved complex are digested into peptides, which are then subject to liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) for identification. Commonly used tags are GFP, 6xHis, Myc, StrepII, and FLAG. Tandem affinity purification (TAP) (29, 30) is a popular variant, wherein baits are engineered to contain two or more tags that permit isolation of protein complexes via sequential two-step purification in native conditions before identification. Sequential purification reduces contaminating proteins, but often results in poor yields. Interactomes by parallel affinity capture (iPAC) (31) aims to overcome sequential losses in yield by parallel purification of protein complexes interacting with multiply tagged baits. Eluted complexes resulting from different affinity matrices are analyzed separately by LC-MS/MS. Contaminants are identified by processing a tagless negative control, and matrix-specific contamination is subtracted in silico.

AP-MS methods have facilitated the generation of large-scale protein complex studies in native conditions in Saccharomyces cerevisiae (5, 32, 33), C. elegans (34), Escherichia coli (35), Drosophila melanogaster (36), and Homo sapiens (37),...
and have also been used to determine dynamic changes in complex components upon perturbation (38–43). When compared with other published data sets, such as MIPS (44) and HPRD (73), IntAct (74), MINT (75), DIP (76), BIND (77), BioGRID (78), MPact (79), and InnateDB (80), collect interaction data via direct submission, expert curation, or computational text-mining of protein interaction data from publications under curator supervision. In addition to the limitations associated with automated text-mining (81, 82), one report suggests that any two databases fully agree on only 42% of interaction data and 62% of proteins curated from the same publication (83). It is demonstrated, however, that such discrepancies are due predominantly to different curation methods, e.g., focus on specific interactions of interest, or mapping mixed-species interactions into a single model organism (84–86). Accordingly, the International Molecular Exchange Consortium (IMEx) has developed a set of curation rules to facilitate both consistency and quality, and now provides a nonredundant set of protein interaction data annotated to a uniform standard (87). Also in support of curation and data comparability, the Proteomics Standards Initiative (PSI) has developed a controlled vocabulary and file format for molecular interactions (PSI-MI) (88), which may be evaluated by any of a selected list of methods (PSISCORE) via the PSI common query interface (PSICQUIC) (89) for a proof of PSICQUIC use see supplementary Table S3 and Table S4 and supplementary Fig. S1). IRefWeb (90) and IRefIndex (4) are systems allowing the merging and interrogation of protein interactions from more than 10 public databases.

In response to unacceptably high FPR in empirical data sets, computational evaluation methods have emerged over the last decade that may be used to assist quality assessment of protein-protein interactions. Such in silico evaluation methods explore different information sources, ranging from experimental output to public knowledge-bases, with different algorithmic approaches, including statistical over-representation, probability distribution modeling of varying complexity, supervised machine learning, and text-mining. Some methods act directly upon AP-MS experimental output to filter false positive interactions, whereas others have the wider aim...
of predicting protein interactions de novo, but may equally be applied as filters on experimental output. Accordingly, the broad spectrum of in silico filtering methods may be characterized by their input data (1), modeling and decision methods (2), type of protein relation (3), and motivation (4). These four distinctions can be used to separate techniques along broad lines and with the specific context of identifying biologically relevant protein interactions in AP-MS protein association data. Here, we review the computational evaluation of protein interactions via several contemporary methods, which mark paradigms from platform-based univariate assessment (PPI prediction based on one variable or parameter), through weighted combinations of multiple scores derived from related annotation and wider experimental sources, toward currently evolving methods that make use of adaptively-weighted, predictive ensembles, derived from all three information sources (platform, annotation and wider data; Fig. 1).

Platform-Based Methods—A plethora of experimental platform-based scoring methods for AP-MS data exemplify the first of the three paradigms introduced above. Platform-based assessments are specifically motivated by FP removal from data derived via AP-MS-based studies (distinction 4, defined earlier), and filter direct interactions (3) via computational assessment of platform output (1). They differ in the statistical methods employed for PPI appraisal (2), especially in the use or otherwise of a subset of “true” and “false” interactions to guide decision thresholds placed on their output.

The Comparative Proteomics Analysis Software Suite, CompPASS (91) was developed to identify high confidence candidate interacting proteins (HCIPs) in AP-MS experiments via mass spectrometry spectral counts (91), and has been applied to human deubiquitinating enzymes (DUB) (91), the human autophagy interaction network (AIN) (92), and the human interaction network for ER-associated degradation (INERAD) (93). CompPASS employs univariate descriptive statistics to assess observed interactions according to functions of the number of baits evaluated and the number of both purifications and replicate runs in which the prey protein is detected. The score is assessed by comparison against randomized data, does not employ control experiments, and is designed for nonreciprocal purifications, giving lower scores to proteins found in multiple experiments. The suite is available online and, if queried with a protein not present in the evaluated experiments, displays results for similar proteins via protein name synonyms. Dependence of the algorithm upon spectral counts may cause it to be adversely affected by inefficiencies in protein digestion, peptide ionization, and performance within an MS/MS experiment. These factors have particular influence on the quantitation of low abundance proteins (94–97), highlighting a potential weakness of experiment-specific methods.

The Significance Analysis of INTeractome (SAINT) system (98) was first introduced to filter kinase-phosphatase interaction data (KPI; 1844 interactions between 887 protein partners) (99) and has been applied subsequently to smaller data sets (98). Like CompPASS, SAINT evaluates the probability of interaction between proteins as a function of spectral counts. The latest version assumes separate Poisson spectral count distributions for bait and prey: one to represent true interactions (test experiments) and another for false interactions (negative control purification experiments). In the absence of a control data set (known false interactions), preys are divided into high- and low-frequency groups according to a user-specified threshold. The final score reflects the probability of observed spectral count belonging to the true interaction distribution, compared with that of false interactions. For optimal evaluation, the authors suggest 15 to 20 different bait purifications (supplementary Data (99)). In comparison with CompPASS, SAINT has displayed increased overlap with published data (98). It relies, however, upon the use of robust controls, good overlap of spectral measurements over successive LC-MS/MS runs, and the assumption that preys in control samples are characterized by lower quantitation values than those in experimental samples. In one published study, for example, peptide lists from two technical replicates were observed to overlap by 35–60% largely because of the stochastic nature of LC-MS/MS selection of peptide ions for fragmentation in successive runs (97).

It is noteworthy that, although most applications of CompPASS and SAINT have involved spectral counting, both approaches can be modified to take other types of quantitation data. A recent development, SAINT-MS1 (100), models MS1 intensity data to filter AP-MS PPI data in similar manner. For an in-depth review of the SAINT see Nesvizhskii (101). Alternatives to both methods are provided in the form of in-house analytical pipelines, some recent examples of which are outlined below.

Guruharsha and coworkers (102) evaluate the confidence of ~3500 D. melanogaster affinity purifications to extract 556 protein complexes in several stages. In the first instance, spectral counts are modeled using the hypergeometric probability distribution as in Hart et al. (103). Using a randomized set for FDR computation, a 0.05% threshold reduces the set to 10,969 binary interactions among 1297 proteins before identification of complexes via a Markov clustering algorithm (104).
Computational Assessment of Protein Interaction Data

DAVID Functional Annotation Tools (105), KEGG (106), and modEncode RNA-Seq data (107) provide additional functional analyses. The hypergeometric distribution score shows improved performance when compared with both Hart et al. (103) and the socio-affinity index (SAI) (32), two methods based on frequency of observation of specific bait-prey combinations, and also when compared with CompPASS and SAINT, methods based on spectral counts. The analysis pipeline delivers scores expressing confidence in protein identification, protein complex co-membership, and protein co-expression.

A further recent study, of membrane-protein complexes in S. cerevisae, uses the “purification enrichment” (PE) score (108) based on observation of two pairs of proteins (bait and prey) in the same and in separate purifications, followed by application of Markov clustering to a combined set of published and newly detected complexes to deliver over 500 putative protein complexes (5). Similarly, Havugimana et al. (6) recently applied Pearson correlation analysis to evaluate spectral counts in combination with the analysis of published protein networks to integrate new experimental data and build a core set of human protein complexes.

Decontaminator, removes contaminants from AP-MS experiments by modeling Mascot protein identification scores (109), as demonstrated by application to a set of 14 “induced” and 14 “noninduced” TAP-tagged RPAP3 complexes in HEK 293 cells (purification protocol (110)). Between five and 516 preys, with a mean of 135, were detected per bait in the induced samples, whereas between 206 and 626 preys, with a mean of 316 proteins, were detected in the noninduced samples. Noise models are constructed on MASCOT scores of noninduced samples, which feature higher mean and range of prey detection frequencies, and their induced counterparts. The resulting posterior probabilities of true PPIs are used subsequently to compute significance of each potential prey interaction. Decontaminator has shown improved performance in comparison to five other methods: two variations of SAINT, and three descriptive statistics, on the evaluation of a combined BioGRID (78) and HPRD (73) data set (109) and may be applied even when few control experiments are available. It should be noted, however, that strong conclusions are elusive when comparing algorithms designed for use with large unconnected networks, such as SAINT, with those designed for application to smaller data collections, such as Decontaminator. The provision of an alternative to conventional subtraction of preys identified in a control sample from those of an experimental sample is welcomed; but the method remains dependent upon both biological sample (i.e. protein purification) and the analytical system employed (i.e. mass spectrometric set-up and protein detection levels). In addition, the universal assumption that FPs are defined by lower Mascot scores than those of true interactors may not always generalize sufficiently in the presence of inconsistencies in peptide detection and mass-spectra-peptide-protein matching between runs (111, 112).

Finally, two recent methods for platform-based filtering of false positive interactions warrant introduction. Quantitative BAC Interactomics (QUBIC) uses either SILAC or “label-free” quantitation in MaxQuant (113) to differentiate between false positives, background binding proteins and bait interaction partners (114), and Mass spectrometry interaction statistics (MiST) combines protein abundance (peak intensity), reproducibility, and specificity (observance of a prey uniquely for one bait) by principal component analysis (115) in an approach that proved competitive against both CompPASS and SAINT when applied to AP-MS data from an experiment describing host-pathogen interactions (116).

Databases and Annotation Confidence Scoring Systems—It is clear that the experiment or platform-specific direct assessment of binding assay output limits the amount of data available from which to derive interaction statistics of sufficient power to discriminate fully between true and false interactions.

Several computational methods aim to provide more generic tools that allow scoring and inference of interactions beyond those putatively identified by single AP-MS experiments. The following recent, well-known, PPI prediction systems all derive a series of single scores from public data sources (distinction i, Introduction) and combine them to assess putative interaction (4) of varying degrees of association (3). The methods described progress from fixed to adaptive weighting when combining PPI scores, and from data description to use of “high-confidence” data subsets to inform decision thresholds placed on them (2).

The Molecular Interaction (MINT) repository evaluates stored protein interactions according to experimental information quality (117). The evaluation score ranges between 0 and 1 to describe the strength of association between two proteins, and is a weighted combination of: experiment size, experiment type, published orthologue interactions, and number of supporting publications, via an exponential function with initial slope provided by an empirically set exponent. Experiments with less than 50 interactions receive an experiment size weight of 1, larger experiments receive 0.5. Experiment type is weighted 1 for experiments supporting a direct interaction, or 0.5 for affinity purifications. Ortholog scores extracted from InParanoid (118), a tool to extract orthologs from multiple organisms using a BLAST-based algorithm, are multiplied to yield a single ortholog weight ranging between 0 and 1. Interactions with more than five supporting publications obtain a weight greater than or equal to 1. The latest version of MINT introduces a “community recognition” weight (75) via a combination of experiment type and ratio between citations and reported interactions in each paper. The improved score rewards focused interaction studies (e.g. Y2H) in highly cited publications. Ease of implementation and functional clarity enable application of this approach independent of experimental method, and is particularly useful in the fast filtering of large data sets. The method is not sufficiently
sensitive for detailed investigation of a smaller set of proteins, however, and the fixed a priori thresholds on publication count, experiment size, and experiment type may require revision to prevent large-scale experiments (e.g. AP-MS) being disadvantaged within this scoring system.

The Interaction database (IntAct) (74) employs a simple system to evaluate constituent experimentally observed PPIs. All IntAct experiments are clustered (manuscript in preparation (74)) to produce a nonredundant protein subset (i.e., a set containing only one entry for each protein) and the confidence of each protein-protein pair is scored via the sum of fixed weights relating to detection method, interaction type, and number of database publications in which the interaction has been observed. For example, weights for the detection method are derived by expert curators from the PSI-MI controlled vocabulary, collapsed to top-level terms in the vocabulary hierarchy, and range from 0.2 (colocalization) to 3 (biochemical). The combined score is normalized to scale between 0 and 1. The algorithm is designed to work on any PSI-formatted file annotated to MIMIx standards (119), and future directions include the incorporation of further properties from PSI-MI controlled vocabularies. A precursor of such methods, the Database of Interacting Proteins (DIP) (76), used three scoring systems based on expression profiles (12), paralogs (12), and domain-domain interactions (120) respectively to evaluate binary physical PPI in experimentally derived collections.

The Search Tool for Retrieval of Interacting Genes (STRING) (121) employs multiple scoring algorithms and annotation sources to retrieve and evaluate possible protein pairs and reject pairs deemed not to interact according to lack of data-driven evidence. Eight distinct scores are incorporated to assess interaction partners for every protein stored in the accompanying database: neighborhood, (gene) fusion, (phyloge netic) co-occurrence, homology, coexpression, experiments, databases, and textmining. Some are novel to STRING itself, whereas others already exist as stand-alone scoring algorithms (122, 123).

The Neighborhood score (124) infers functional linkage based on conserved gene clusters in prokaryotes (125, 126), gene fusion (124) assumes a functional relation between genes evolved from a single gene via fission (127–129) and phylogenetic co-occurrence assesses functional linkage among gene families via similarity in evolutionary history (124, 126, 130). Despite potential weaknesses because of assumed application of prokaryotic properties to eukaryotes (128, 131–139), and when used alone (140), predictive strength is derived from their combination alongside wider information obtained via data-mining.

For example, the Homology score (141) employs external ortholog assignments stored in Clusters of Orthologous Groups (COGs), with interaction confidence score between two proteins from different COGs assumed to be either partially or wholly (depending on mode of use) transferable to any pair of proteins belonging to the same two COGs. Co-expression (141) combines correlation of gene expression across multiple transcriptional profiles, retrieved via ArrayProspector (142), with the context of known interaction pathways (KEGG) to assess putative PPIs for consistency, accompanied by familiar advantages and disadvantages of using microarray data to infer interactions (143, 144) and transcription to approximate resulting gene or protein activity. The Experiments score (141) uses protein interactions imported from databases—some described above—including MINT (75), BIND (77), DIP (76), BioGrid (78), IntAct (74), and PDB (145). The imported experiments are evaluated based on the size of the experiment: small-scale experiments get a fixed empirical score, whereas protein complexes obtained from large-scale studies are evaluated as the log ratio of the number of studies supporting the interaction and the number of studies supporting the contrary (141). Database evidence (126, 141) uses external published pathway information from databases including KEGG (106), Reactome (146), and BioCarta2, and derives from the frequency with which each protein pair appears in the same pathway. Finally, the Textmining score (126, 147) searches abstracts from SGD (148), OMIM (149), The Interactive Fly, and PubMed for published regulatory gene interactions, parsing for co-occurring gene names and interaction information using Natural Language Processing tools (123). Ongoing improvements are supported by community initiatives such as the Critical Assessment of Information Extraction systems in Biology (BioCreAtivE)3.

Each individual subscore described above is benchmarked against a common TP reference set, KEGG (106) pathway membership, and represents on its own the confidence of a true interaction. A naïve Bayesian combination of these subscores under the assumption of independence (141) is employed subsequently to produce a single confidence score scaled to range between 0 and 1 (141). STRING thereby combines multiple annotation sources evaluated via a range of computational methods into one composite interaction score.

In relation to AP-MS experimental data, putative interactions may be filtered via intersection with a list of proteins suggested by STRING to interact with the same bait protein, taking composite scores and scoring methods into consideration upon interpretation of the result. A study in Mycoplasma pneumoniae bacteria used STRING in conjunction with an experimental scoring scheme resulting in some lower experimental scores being elevated and as such used this approach to validate lower scoring empirical interactions using orthogonal evidence (45). A potential weakness of the STRING framework is that it is gene-orientated and, therefore, does not acknowledge specific protein isoforms or post-translational modifications. Accordingly, a measure of care

---

2 http://www.biocarta.com.
3 http://biocreative.sourceforge.net.
should be taken during interpretation, especially when applied to filter the output of AP-MS experiments. In mitigation, the composite interaction score reflects individual subscores in a transparent fashion, which greatly assists interpretation, and the combination of subscores across varied approaches and biological information lends robustness beyond that of each individual constituent.

In the above sections, the scoring systems described for potential application to filter PPIs suggested by AP-MS experiments have progressed from descriptive statistics on individual data sets to predictive statistics based on wider experimental data, moving toward conjoint analysis of multiple variables or parameters and the derivation of decision thresholds from wider annotation data.

**Composite Methods**—Recent techniques to evaluate true physiological interactions tend toward the adaptively weighted combination of multiple PPI assessments "learned" from a variety of information sources. A good example is provided by the Functional Coupling database—FunCoup (150), which constructs global functional coupling networks for the most-researched eukaryotes based on proteomic and genomic data. The input sources are listed and given brief descriptions below, with references to related works:

- mRNA co-expression (MEX): data from species-specific expression studies;
- phylogenetic profile similarity (PHP);
- protein-protein interactions (PPIs): obtained from IntAct (see previous section);
- subcellular colocalization (SCL): obtained from GO (151);
- protein co-expression (PEX): data obtained from Human Protein Atlas (152);
- shared transcription factor binding (TFB): from enhancer element locator tool (EEL) (153), MPromDb (154), Interolog/Regulog Database (155);
- co-miRNA regulation (MIR): from miRBase (156);
- genetic interactions (GIN): for *S. cerevisiae* (157); and
- domain associations (DOM): obtained via UniDomInt (158)

Co-expression (MEX) is evaluated from normalized microarray data by Pearson and Spearman rank correlations. Phylogenetic signatures (PHP) based on the major clades (*Fungi, Plantae, Animalia*) are employed to calculate log-likelihood ratios pertaining to shared phylogeny. Protein-protein interaction data (PPI) are evaluated based on number of papers, experiment type and experiment size. Data linked to one paper, one experiment and multiple preys are penalized, whereas data with multiple papers and multiple experiments are favored. GO terms relating to SCL are weighted based on their mutual information, an information theory quantifier (159) of the codependence of each GO term in the evaluated GO term pair, and number of proteins assigned to each term. The PEX score evaluates Human Protein Atlas expression and subcellular localization data using the same mutual information quantification principle as for SCL. TFB and MIR compute scores directly proportional to the ratio of binding sites shared by two proteins over all binding sites associated with the pair. The UniDomInt descriptor incorporates nine domain interaction databases (158) and is incorporated directly. Likewise, GIN scores are incorporated from genetic interaction profiles suggested by Costanzo et al. (157).

**FunCoup** delivers a discrete evaluation of physical interaction, protein complex membership, and metabolic or signaling link. To map combined, un-normalized metric scores to these four classes, FunCoup employs four corresponding training sets built from data retrieved from IntAct, HPRD, BIND, KEGG, and UniProt. The four sets correspond to: **physical interaction**—proteins reported in one PPI experiment and belonging to the same KEGG pathway, or reported in multiple PPI experiments; **protein complex**—members of the same UniProt complex; **metabolic link**—members of either a small KEGG metabolic pathway, or partners in a specified minimum number of (organism specific) metabolic pathways; **signaling link**—similar to metabolic link, but for signaling pathways. Corresponding negative training sets and the pitfalls associated with them (160) are avoided by use of Bayesian methods to implicitly define a background evidence probability (with background defined as sets of protein pairs less likely to interact when compared the ‘gold standard’ sets of true interacting partners). An interaction model is created from each training set and outputs are combined using a naive Bayesian network to deliver a single confidence value that ranges between 0 and 1 (161).

FunCoup scores have displayed Spearman rank correlation of 0.19 when compared with STRING, with relatively low positive correlation attributed to difference in methodologies and input data (161). A positive correlation is expected, as both systems use external annotation to infer PPIs, but it should be noted that Spearman rank correlation is sensitive to changes in rank, rather than individual score, which gives small score changes (e.g. from 0.88 to 0.89) a larger influence on correlation than merited by biological context.

Van Haagen et al. (162) combine text mining (*Peregrine* (163)), Gene Ontology (GO) over-representation analysis, microarray data (*COXPRESdb* (164)), tissue specific gene expression data (*TiGER* (165), and domain-domain interaction (*DOMINE*) (166) information into a single PPI prediction system and demonstrate their approach by inferring potential interaction partners for dysferlin and huntingtin (162). MEDLINE abstracts are processed via text-mining with a dictionary of gene name synonyms, spelling variation of concepts, and protein-specific context profiles (128). Microarray data (*COXPRESdb* (129)) are evaluated using Pearson correlation coefficients to quantify the co-expression of each gene pair. The DDI score appraises binary PPIs via number of interacting domains according to DOMINE. Information content of the lowest common GO term ancestor (LCA) is employed as a similarity measure between two proteins. Fisher’s method is used to combine the individual database scores.
A reported increase in performance over STRING (127) may be caused by the latter method’s bias towards reporting all potentially functionally linked proteins, plus differences in annotation sources and text-mining procedures. The systems differ also in their treatment of microarray data, as STRING uses ArrayProspector whereas van Haagen uses COXPRESdb and TIGER to assess co-expression across tissue-specific expression data. A potential improvement might be obtained by taking into consideration the different evidence codes that indicate source of information for each GO term assignment.

Further online resources that combine data types and eukaryotic gene networks include GeneMANIA (167), which covers A. thaliana, C. elegans, D. melanogaster, M. musculus, H. sapiens and S. cerevisiae, and bioPIXIE, which builds S. cerevisiae networks based on gene expression, genetic and physical interactions, localization annotation, and literature curation (168). Alternatively, Kiel and coworkers describe a new conceptual analysis pipeline that combines experimental proteomic data, literature mining, computational analyses, and structural information to generate a multiscale signal transduction network based on tissue-specific gene expression and domain-domain interaction data for the study of rhodopsin and its interactions (169).

The systems outlined in this section provide examples of contemporary PPI inference methods to which the filtering of false-positive (and identification of false negative) interactions from AP-MS experiments lends a dual purpose. This dual purpose also frames the present state of the field, because it appears that successful computational approaches to the postreduction of FPR in AP-MS experiments may require a balance to be struck between the specificity of direct analysis on experimental output and the strong phenotypic relationships inferred from abundant annotation and large databases of generic curated PPIs.

Summary and Future Prospects—The approaches described in this review may all be employed to assist the interpretation of empirical AP-MS-derived protein interaction data, by filtering false-positive interactions in particular. The methods presented differ in motivation and consideration of what constitutes interaction, ranging from acute assessment of direct interactions from single experimental output to generic PPI inference from proteome-wide annotation (see Table I). The approaches also differ in their approach to evaluation and decision-making, from single descriptive statistics to distinctions between exemplar PPIs and noninteracting counterparts learned across multiple descriptors. Each approach has strengths and weaknesses and arguably there exists considerable scope to refine current methodologies.

In the immediate term, increasing generation and expansion of published interactomes is likely to encourage the evaluation of multiple data sets to reveal system-wide trends. Regardless of the specific approach taken, weighted combinations of multiple PPI scores, and the decision thresholds placed upon them, provoke new considerations regarding the selection and representation of training data that exemplifies the distinction sought between proteins that interact and those that do not. In this context, avenues of potential improvement include: further development of combined scoring systems to provide multisource data representations for PPI inference via supervised machine learning or modeling techniques based on multiple sources of information; the co-analysis of readily available annotation data to augment a limited quantity of more specific experimental data obtained via platform or experiment-specific methods (data fusion); and the application of training data design strategies from other disciplines, such as high-throughput molecular screening (170), to guide the production of more appropriate exemplar data collections from which to discriminate between true and false PPIs.

The present trend of combining multiple information sources in pursuit of protein interaction filters that are (a) specific to experimental output, and also (b) reliable and comparable across multiple experiments and platforms, comprises a genuinely interdisciplinary endeavor. The required combination of expertise includes computer scientists and bioinformaticians, to design and build multisource computational filters, working alongside developers of the experimental methods employed and the expert knowledge of curators, biologists and experimentalists.

In conclusion, the importance of determining sets of interacting proteins and how these interactions change in response to a highly dynamic cellular environment cannot be underestimated. The field has impact on many fundamental areas of cell biology research and beyond, including the selection of drug targets and the elucidation of therapeutic mechanisms of action, which play an increasingly important role in human health. Continuing platform development may alleviate the unacceptably high FPRs that presently impede the unfiltered use of common empirical methods, but sources of noise and unwanted variance are inherent to the underpinning technologies and are likely to perpetuate a need for computational postanalysis of experimental output. Accord-

---

**Table I**

| Systems          | Production stage | Postproduction stage |
|------------------|------------------|----------------------|
| CompPASS        | X                |                      |
| SAINT            | X                |                      |
| Guruharsha et al. | X              | X                    |
| Decontaminator   | X                |                      |
| MINT             | X                |                      |
| IntAct           | X                |                      |
| STRING           | X                |                      |
| FunCoup          | X                |                      |
| van Haagen et al. | X              |                      |
ingly, the development and implementation of sophisticated computational tools to assist in removal of erroneous interactions is, and will continue to be, the linchpin of such technologies.

Acknowledgments—We would like to thank Sandra Orchard (EMBL-EBI) and Daniel Nightingale (Cambridge Centre for Proteomics) for useful comments on the manuscript.

IMA is funded by BBSRC grant BB/F017464/1. MWBT and KSL are members of PRIME-XS a 7th Framework Programme of the European Union (262067 - PRIME-XS).

This article contains supplemental Fig. S1 and Tables S1 to S4.

REFERENCES

1. Lee, S. W., Berger, S. J., Martinovic, S., Pasa-Tolic, L., Anderson, G. A., Shen, Y., Zhao, R., and Smith, R. D. (2002) Direct mass spectrometric analysis of intact proteins of the yeast large ribosomal subunit using capillary LC/FTICR. Prot. Natl. Acad. Sci. U. S. A. 99, 5942–5947.
2. Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. (2009) Up-to-date catalogue of yeast protein complexes. Nucleic Acids Res. 37, 825–831.
3. Alberts, B. (1998) The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists. Cell 92, 291–294.
4. Razick, S., Magkيرانis, G., and Donaldson, I. M. (2008) iRefindex: a consolidated protein interaction database with provenance. BMC Bioinformatics 9, 405.
5. Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B. D. M., Yu, H., Sahalie, J. M., Murray, R. R., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Pochart, P., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Ezin, P., Alonso, M. A., Davis, N., Chait, B. T., Rost, B., and Roth, B. (2005) Comparative analysis of protein-protein interactions in the secretory pathway of living cells. Proc. Natl. Acad. Sci. U.S.A. 102, 6350–6355.
6. Barrios-Rodiles, M., Liu, Z., Ozdamar, B., Bose, R., Liu, Z., Donovan, R. S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I. W., Luga, V., Pruzil, N., Robinson, M., Suzuki, H., Hayashizaki, Y., Jurisica, I., and Wrana, J. L. (2005) High-throughput mapping of a dynamic signaling network in mammalian cells. Science 307, 1621–1625.
7. Eyckerman, S., Verhee, A., der Heyden, J. V., Lemmens, I., Ostade, X. V., Vandekerckhove, J., and Tavernier, J. (2001) Design and application of cytokine-receptor-based affinity-trap interaction map. Nat. Cell Biol. 3, 1114–1119.
8. Nyfeler, B., Michnick, S. W., and Hauri, H. P. (2005) Capturing protein interactions in the secretory pathway of living cells. Proc. Natl. Acad. Sci. U.S.A. 102, 6350–6355.
9. Ramachandran, N., Raphael, J. V., Hainsworth, E., Demirkan, G., Fuentes, M. G., Rolfs, A., Hu, Y., and LaBaer, J. (2008) Next-generation high-density self-assembling functional protein arrays. Nat. Methods 5, 535–538.
10. Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J. M., Murray, R. R., Conover, D., de Smet, A. S., Venkatases, K., Rual, J. F., Vandenhaute, J., Cuskic, M. E., Pawson, T., Hill, D. E., Tavernier, J., Wrana, J. L., Roth, F. P., and Vidal, M. (2009) An experimentally derived confidence score for binary protein-protein interactions. Nat. Methods 6, 91–97.
11. Venkatases, K., Rual, J. F., Vazquez, A., Steib, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K. I., Yildrim, M. A., Simonis, N., Heinzmann, K., Gebreb, F., Sahalie, J. M., Cevik, S., Sahalie, J. M., Murray, R. R., and Vidal, M. (2009) An empirical framework for binary interactome mapping. Nat. Methods 6, 535–538.
12. Chen, Y.-C., Rajagopala, S. V., Stellberger, T., and Uetz, P. (2010) Exhaustive benchmarking of the yeast two-hybrid system. Nat. Methods 7, 667–668.
13. Vazquez, A., Rual, J. F., and Venkatases, K. (2011) Quality control methodology for high-throughput protein-protein interaction screening. Methods Mol. Biol. 781, 279–294.
14. MacBeath, G., and Schreiber, S. L. (2000) Printing proteins as microarrays for high-throughput observations. Science 290, 1723–1725.
15. Rajagopala, S. V., Hughes, K. T., and Uetz, P. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol. Cell. Proteomics 1, 349–356.
16. Barrios-Rodiles, M., Liu, Z., Ozdamar, B., Bose, R., Liu, Z., Donovan, R. S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I. W., Luga, V., Pruzil, N., Robinson, M., Suzuki, H., Hayashizaki, Y., Jurisica, I., and Wrana, J. L. (2005) High-throughput mapping of a dynamic signaling network in mammalian cells. Science 307, 1621–1625.
17. Eyckerman, S., Verhee, A., der Heyden, J. V., Lemmens, I., Ostade, X. V., Vandekerckhove, J., and Tavernier, J. (2001) Design and application of cytokine-receptor-based affinity-trap interaction map. Nat. Cell Biol. 3, 1114–1119.
18. Nyfeler, B., Michnick, S. W., and Hauri, H. P. (2005) Capturing protein interactions in the secretory pathway of living cells. Proc. Natl. Acad. Sci. U.S.A. 102, 6350–6355.
19. Ramachandran, N., Raphael, J. V., Hainsworth, E., Demirkan, G., Fuentes, M. G., Rolfs, A., Hu, Y., and LaBaer, J. (2008) Next-generation high-density self-assembling functional protein arrays. Nat. Methods 5, 535–538.
20. Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J. M., Murray, R. R., Conover, D., de Smet, A. S., Venkatases, K., Rual, J. F., Vandenhaute, J., Cuskic, M. E., Pawson, T., Hill, D. E., Tavernier, J., Wrana, J. L., Roth, F. P., and Vidal, M. (2009) An experimentally derived confidence score for binary protein-protein interactions. Nat. Methods 6, 91–97.
21. Venkatases, K., Rual, J. F., Vazquez, A., Steib, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K. I., Yildrim, M. A., Simonis, N., Heinzmann, K., Gebreb, F., Sahalie, J. M., Cevik, S., Sahalie, J. M., Murray, R. R., and Vidal, M. (2009) An empirical framework for binary interactome mapping. Nat. Methods 6, 535–538.
22. Chen, Y.-C., Rajagopala, S. V., Stellberger, T., and Uetz, P. (2010) Exhaustive benchmarking of the yeast two-hybrid system. Nat. Methods 7, 667–668.
23. Vazquez, A., Rual, J. F., and Venkatases, K. (2011) Quality control methodology for high-throughput protein-protein interaction screening. Methods Mol. Biol. 781, 279–294.
24. MacBeath, G., and Schreiber, S. L. (2000) Printing proteins as microarrays for high-throughput function determination. Science 289, 1760–1763.
25. Kung, L. A., and Snyder, M. (2006) Proteome chips for whole-organism assays. Nat. Rev. Mol. Cell Biol. 7, 617–622.
26. Tessler, L. A., and Mitra, R. D. (2011) Sensitive single-molecule protein quantification and protein complex detection in a microarray format. Proteomics 11, 4731–4735.
27. Schramaker, B. A., and Panchenko, A. R. (2007) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. Proteomics 7, 1–50.
28. Collins, M. O., and Choudhary, J. S. (2008) Mapping multiprotein complexes by affinity purification and mass spectrometry. Curr. Opin. Biotechnol. 19, 324–330.
29. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999) A generic protein purification method for the functional characterization and proteome exploration. Nat. Biotechnol. 17, 1030–1032.
30. Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouvetet, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. (2001) The tandem affinity purification (TAP) method: a general procedure of complex protein purification. Methods 24, 218–229.

Molecular & Cellular Proteomics 12.1
31. Rees, J. S., Lowe, N., Armean, I. M., Roote, J., Johnson, G., Drummond, E., Spriggs, H., Ryder, E., Russell, S., St Johnston, D., and Lilley, K. S. (2011) In vivo analysis of proteomes and interactomes using Parallel Affinity Capture (iPAC) coupled to mass spectrometry. *Mol. Cell. Proteomics* **10**, M110.002386

32. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Bahler, J., Schueler, M., Friebe, M., Rau, C., Huhse, B., Uetz, P., et al. (2004) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **429**, 145–150

33. Kratchmarova, I., Blagoev, B., Haack-Sorensen, M., Kassem, M., and Mann, M. (2005) Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. *Science* **308**, 1472–1477

34. Mewes, H. W., Ruepp, A., Theis, F., Rattie, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K. F. X., Stümpflen, V., and Antonov, A. (2011) MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.* **39**, D220–224

35. A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpfelde, B., Edelmann, A., Heurteur, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hutvak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Dewes, G., Neubauer, G., Wick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636

36. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Bahler, J., Schueler, M., Friebe, M., Rau, C., Huhse, B., Uetz, P., et al. (2004) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **429**, 145–150

37. Mann, M. (2003) A proteomics strategy to elucidate functional protein-protein interactions by mass spectrometry. *Nat. Biotechnol.* **21**, 2132–2143

38. Mann, M. (2005) Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. *Science* **308**, 1472–1477

39. Mewes, H. W., Ruepp, A., Theis, F., Rattie, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K. F. X., Stümpflen, V., and Antonov, A. (2011) MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.* **39**, D220–224

40. Kühner, S., van Noort, V., Betts, M. J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., Casteño-Diez, D., Chen, W. H., Devos, D., Güell, M., Norambuena, T., Rake, I., Rybin, V., Schmidt, A., Yus, E., Aebersold, R., Herrmann, R., Böttcher, B., Franøgakis, A. S., Russell, R. B., Serrano, L., Bork, P., and Gavin, A. C. (2009) Proteome organization in a genome-reduced bacterium. *Science* **326**, 1235–1240

41. Gavin, A. C., Maeda, K., and Kühner, S. (2011) Recent advances in charting protein-protein interaction: mass spectrometry-based approaches. *Curr. Opin. Biotechnol.* **22**, 42–49

42. Glatter, T., Wepf, A., Aebersold, R., and Gstaiger, M. (2009) An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Mol. Syst. Biol.* **5**, 237

43. Chang, I. F. (2006) Mass spectrometry-based proteomic analysis of the proteomics. *Mol. Cell. Proteomics* **1**, 376–386

44. Perseus, A., Andersson, P., Draghici, S., von Mering, C., and Stumpf, M. P. (2015) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Science* **347**, 976–989

45. Boutilier, K., Ross, M., Podtelejnikov, A. V., Orsi, C., Taylor, R., Taylor, P., and Aebersold, R. (2003) The study of macromolecular complexes by affinity purification coupled to mass spectrometry. *Nature* **429**, 397–403

46. Goyon, Y., Canete, M. M., Vlasblom, J., Wu, S., J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., and Selbach, M. (2006) Interaction network and proteomic arrays. *Mol. Cell. Proteomics* **5**, 1870–1883

47. Schmidt, A., Aloy, P., Grandi, P., Krause, R., Boesche, M., Bahler, J., Schueler, M., Friebe, M., Rau, C., Huhse, B., Uetz, P., et al. (2004) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **429**, 145–150

48. Mann, M. (2003) A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat. Biotechnol.* **21**, 315–318

49. Rual, J. F., Venkatesan, K., Hao, Y., Zhang, Y., Dricot, A., Zeevenberg, M., J. E., Sheung, A., Beattie, B., Krogan, N., and Wodak, S. J. (2004) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637–643

50. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386
validated MS/MS test datasets. Anal. Chim. Acta 534, 11–20
65. Baigley, B. M., Laudeman, T., Yang, L., Song, T., and Lee, C. S. (2007) Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. Mol. Cell. Proteomics 6, 1599–1608
66. Nesvizhskii, A. I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat. Methods 4, 759–777
67. Reiter, L., Claassen, M., Schrimp, S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O., and Aebersold, R. (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol. Cell. Proteomics 8, 2405–2417
68. Yu, W., Taylor, J. A., Davis, M. T., Bonilla, L. E., Lee, K. A., Auger, P. L., Farnsworth, C. C., Welcher, A. A., and Patterson, S. D. (2010) Maximizing the utility of peptide identification in large-scale proteomic experiments by harnessing multiple search engines. Proteomics 10, 1172–1189
69. Stengel, F., Aebersold, R., and Robinson, C. V. (2012) Joining forces: integrating proteomics and cross-linking with the mass spectrometry of intact complexes. Mol. Cell. Proteomics 11, R111.014027
Soehn, C. H., Lee, J. E., Sweredoski, M. J., Graham, R. L. J., Bader, G. D., Betel, D., and Hogue, C. W. V. (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 31, D540–544
71. Bader, G. D., Cary, M. P., and Sander, C. (2006) Pathguide: a pathway resource list. Nucleic Acids Res. 34, D504–506
72. Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, P., Periasamy, M., Mohan, S., Ramabadran, S., Chae, D. H., Nusinov, B. A., Mohmood, R., Ramachandra, Y. L., Krishna, V., Yau, M., Acab, M., Tulpan, D., Whiteside, M. D., Chikatamarla, S., Cesareni, G., and Hermjakob, H. (2007) Literature-curated protein interaction datasets. Nat. Methods 4, 934–935
73. Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod, N., Bader, D. G., Xenarios, I., Wojcik, J., Sherman, D., Salama, J. J., Moore, S., Ceol, A., Chuat-Aryamontri, A., Oesterheld, M., Stumpfens, V., Salwinski, L., Nerotin, J., Cerami, E., Cusick, M. E., Vidum, V., Gilson, M., Armstrong, J., Woolard, P., Hogue, C., Eisenberg, D., Cesareni, G., Apweiler, R., and Hermjakob, H. (2007) Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biol. 5, 44
74. Kerrien, S., Orchard, S., Blankenburg, H., Aranda, B., Brickman, F. S., Ceol, A., Chuat, E., Dana, J. M., De Las Rivas, J., Dujardins, N., De Las Rivas, J., Dumousseau, M., Galeota, E., Goll, J., Hancock, R. E., Hancock, R., Hani, A., Iserlin, R., Jimenez, R. C., Kerssemakers, J., Khadake, J., Lynn, D. J., Michaud, M., O’Kelly, O., Ono, K., Pedraza, A., Prieto, C., Ricard-Blum, S., Rual, J.-F., Rives, J., Sadler, M., Schirru, P., Stu¨ mpfen, T., Tyrrell, G., Uetz, P., Xenarios, I., and Hermjakob, H. (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat. Methods 9, 528–529
75. Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., Morrison, K., Donaldson, I. M., and Wodak, S. J. (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database 2010, baq023
76. Sowa, M. E., Bennett, E. J., Gygi, S. P., and Harper, J. W. (2009) Defining the human deubiquitinating enzyme interaction landscape. Cell 138, 389–403
77. Behrends, C., Sowa, M. E., Gygi, S. P., and Harper, J. W. (2010) Network organization of the human autophagy system. Nature 466, 68–76
78. Christianson, J. C., Oltzm, J. A., Shaler, T. A., Sowa, M. E., Bennett, E. J., Richter, C. M., Tyler, R. E., Greenblatt, E. J., Harper, J. W., and Kopito, R. R. (2012) Defining human ERAD networks through an integrative mapping strategy. Nat. Cell Biol. 14, 93–105
79. Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Rosing, K. A., and Ahn, N. G. (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol. Cell Proteomics 4, 1487–1502
80. Autschbech, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007) Quantitative mass spectrometry in proteomics: a critical review. Anal. Bioanal. Chem. 389, 1017–1031
81. Barton, S. J., and Whittaker, J. C. (2009) Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. Mass Spectrom. Rev. 28, 177–187
proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol. 2, RESEARCH0034.

132. Enright, A. J., Kunin, V., and Ouzounis, C. A. (2003) Protein families and TRIBES in genome sequence space. Nucleic Acids Res. 31, 4632–4638

133. Lee, J. M., and Sonnhammer, E. L. L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 13, 857–882

134. Hurst, L. D., Pal, C., and Lercher, M. J. (2004) The evolutionary dynamics of neighboring genes in eukaryotic genomes. Genomics 91, 243–248

135. Blumenthal, T. (2004) Operons in eukaryotes. Brief Funct. Genomic Proteomic 3, 199–211

136. Hurst, L. D., Pal, C., and Lercher, M. J. (2004) The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. 5, 299–310

137. Michalak, P. (2008) Coexpression, coregulation, and cofunctionality of prokaryotic gene fusion events. Nucleic Acids Res. 32, D273–276

138. Singh, S., and Wall, D. P. (2008) Testing the accuracy of eukaryotic phylogenetic profiles for prediction of biological function. Evol Bioinform Online 4, 217–223

139. Huynen, M., Snel, B., Lathe, W., 3rd, and Bork, P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. 10, 1204–1210

140. von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffret, N., Huynen, M. A., and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 33, D333–437

141. von Mering, C., Jensen, L. J., Lagarde, J., von Mering, C., and Bork, P. (2004) ArrayProspector: a web resource of functional associations inferred from microarray expression data. Nucleic Acids Res. 32, W445–448

142. Donoho, D. L. (2000) High-Dimensional Data Analysis: the curses and blessings of dimensionality. Technical Review.

143. Donoho, D. L. (2000) High-Dimensional Data Analysis: the curses and blessings of dimensionality. Technical Review.

144. Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249–255

145. Sherman, R. A., Croft, D., Yung, C. K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L. D. (2011) The Reactome BioMart. Database 2011, bta031

146. Sherman, R. A., Croft, D., Yung, C. K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L. D. (2011) The Reactome BioMart. Database 2011, bta031

147. Sherman, R. A., Croft, D., Yung, C. K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L. D. (2011) The Reactome BioMart. Database 2011, bta031

148. Sherman, R. A., Croft, D., Yung, C. K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L. D. (2011) The Reactome BioMart. Database 2011, bta031

149. Sherman, R. A., Croft, D., Yung, C. K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L. D. (2011) The Reactome BioMart. Database 2011, bta031

150. Sherman, R. A., Croft, D., Yung, C. K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L. D. (2011) The Reactome BioMart. Database 2011, bta031

151. Sherman, R. A., Croft, D., Yung, C. K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L. D. (2011) The Reactome BioMart. Database 2011, bta031

152. Sherman, R. A., Croft, D., Yung, C. K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L. D. (2011) The Reactome BioMart. Database 2011, bta031

153. Sherman, R. A., Croft, D., Yung, C. K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L. D. (2011) The Reactome BioMart. Database 2011, bta031

154. Sherman, R. A., Croft, D., Yung, C. K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L. D. (2011) The Reactome BioMart. Database 2011, bta031

155. Sherman, R. A., Croft, D., Yung, C. K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L. D. (2011) The Reactome BioMart. Database 2011, bta031

156. Kozomara, A., and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 39, D152–157

157. Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S., Prinz, J. St Onge, R. P., VanderSluis, B., Maknevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahri, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z. Y., Liang, W., Marbach, M., Pav, J., San Luis, B. J., Shriver, I., Tong, A. H., van Dyk, N., Wallace, I. M., Whitney, J. A., Weriauch, M. T., Zong, G., Zhu, H., Houny, W. A., Brudno, M., Bagibiezadeh, S., Bopp, B. P., Pál, C., Roth, F. P., Glaevie, G., Nislov, C., Taryanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A. C., Morris, O. D., Kim, P. M., Kaiser, C. A., Myers, C. L., Andrews, B. J., and Boone, C. (2010) The genetic landscape of a cell. Science 327, 425–431

158. Björkholm, P., and Sonnhammer, E. L. L. (2009) Comparative analysis and unification of domain-domain interaction networks. Bioinformatics 25, D202–D205

159. Cover, T. M., and Thomas, J. A. (2006) Elements of Information Theory, 2nd Ed., Wiley-Interscience.

160. Ben-Hur, A., and Noble, W. S. (2006) Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinf 7, 82

161. Alexeyenko, A., and Sonnhammer, E. L. L. (2009) Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome Biol. 10, 1107–1116

162. van Hagen, H. H., ‘t Hoen, P. A., de Morréée, A., van Roon-Mom, W. M., Peters, D. J., Roos, M., Monis, B., van Ommen, G. J., and Schuermie, M. J. (2011) In silico discovery and experimental validation of new protein-protein interactions. Proteomics 11, 843–853

163. Jelier, R., Schuemie, M. J., Roes, P. J., van Mulligen, E. M., and Kors, J. A. (2008) Literature-based concept profiles for gene annotation: the issue of weighting. Int. J. Med. Inform. 77, 354–362

164. Yellaboina, S., and Kinosita, K. (2011) COXPRESdb: a database to compare gene coexpression in seven model animals. Nucleic Acids Res. 39, D1016–1022

165. Liu, X., Xu, Y., Zack, D. J., Zhu, H., and Qian, J. (2008) TiGER: a database for tissue-specific gene expression and regulation. BMC Bioinformatics 9, 271

166. Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Mattlant, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., and Morris, Q. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 38, W214–220

167. Myers, C. L., Dolinski, K., and Troyanskaya, O. G. (2005) Discovery of biological networks from diverse functional genomic data. Genome Biol. 6, R114

168. Kiel, C., Vogt, A., Campagna, A., Chatr-aryamontri, A., Swiatek-de Lange, M., Beer, M., Bolz, S., Zack, D., Kinf, N., Cesareni, G., Serrano, L., and Ueffing, M. (2011) Structural and functional protein network analyses predict novel signaling functions for rhodopsin. Mol. Syst. Biol. 7, 523

169. Kennard, R. W., and Stone, L. A. (1969) Computer Aided Design of Experiments. McGraw-Hill, New York.