Protein complex analysis : from raw protein lists to protein interaction networks

Protein complex analysis: from raw protein lists to protein interaction networks

Pieter Meysman[1,2,*,§], Kevin Titeca[3,4,*], Sven Eyckerman[3,4], Jan Tavernier[3,4], Bart Goethals[1], Lennart Martens[3,4], Dirk Valkenborg[5,6,7], Kris Laukens[1,2]

Affiliations

[1] Advanced Database Research and Modelling (ADReM), Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium
[2] Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp / Antwerp University Hospital, Edegem, Belgium
[3]Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium
[4]Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium
[5]Flemish Institute for Technological Research (VITO), Mol, Belgium
[6]I-BioStat, Hasselt University, Hasselt, Belgium
[7]CFP-CeProMa, University of Antwerp, Antwerp, Belgium

* These authors contributed equally to this work.

[§]Corresponding author

Email: pieter.meysman@uantwerpen.be

Telephone: +32 3 265 3407

Address: Advanced Database Research and Modelling (ADReM), Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, Antwerp, Belgium

Running title: **Protein complex analysis**

Keywords: Bioinformatics, co-complex purification, protein-protein interaction networks

Abstract

The elucidation of molecular interaction networks is one of the pivotal challenges in the study of biology. Affinity purification - mass spectrometry and other co-complex methods have become widely employed experimental techniques to identify protein complexes. These techniques typically suffer from a high number of false negatives and false positive contaminants due to technical shortcomings and purification biases. To support a diverse range of experimental designs and approaches, a large number of computational methods have been proposed to filter, infer and validate protein interaction networks from experimental pull-down MS data. Nevertheless, this expansion of available methods complicates the selection of the most optimal ones to support systems biology-driven knowledge extraction. In this review, we give an overview of the most commonly used computational methods to process and interpret co-complex results, and we discuss the issues and unsolved problems that still exist within the field.

Table of Contents

I. Introduction
II. Reliability of interactions from co-complex experiments
III. The source data of a co-complex experiment
IV. Strategies for filtering false positive interactions
    A. Non-parametric filtering
    B. Parametric filtering
    C. Cluster-based filtering
V. Use of external protein interaction data for contaminant removal
    A. Reference interaction data sets
    B. Integration of reference data into the filtering step
VI. Prioritization of candidate partners
    A. Based on functional data
    B. Based on orthology information
VII. Strategies for prediction of false negative interactions
VIII. From processed interactions to biological insight
    A. Analysis of specific protein complexes or interactions
    B. Analysis of protein complex stoichiometry
    C. Analysis of protein complex structures
    D. Analysis of the entire network
    E. Visualisation of protein interaction networks
IX. Conclusions and future prospects
    A. Different analytical methods for different experimental setups
    B. Rise of standard workflows and quality control
X. Acknowledgements
XI. References

I. Introduction

The study of protein interactions remains highly challenging due to the diversity in the way proteins interact, the subcellular context of the interactions and the possible involvement of post-translational modifications. Current methods to detect protein-protein interactions (PPIs) can be roughly divided in binary and co-complex approaches. The yeast-two hybrid system (Rolland et al., 2014), MAmmalian Protein-Protein Interaction Trap (MAPPIT) (Eyckerman et al., 2001) and protein complementation assays (Tarassov et al., 2008; Morell et al., 2009) are examples of methods that perform a binary test between an expressed candidate bait and prey pair, typically coupled to a reporter activity read-out (Suter et al., 2008; Lievens et al., 2009). Positive data from these assays suggest that the tested proteins interact in a direct way, although endogenous proteins within the test system (e.g. MAPPIT is performed in human cells) often confound this interpretation. Affinity purification-mass spectrometry (AP-MS), immunoprecipitation-mass spectrometry (IP-MS) and tandem affinity purification-mass spectrometry (TAP-MS) are the best known examples of co-complex methods, wherein a bait of interest is typically expressed in the cells of interest, followed by purification and subsequent MS-based identification of the complex constituents (Gingras et al., 2007). Several recent reviews are available that detail the technical operation and differences of the co-complex MS methods (Collins & Choudhary, 2008; Gavin et al., 2011; Dunham et al., 2012; Oeffinger, 2012; Walzthoeni et al., 2013).

It is clear that the interactome changes with the cellular context and is critically dependent on the activation state of the cells. Current global binary approaches do not readily capture these aspects but rather give a rough estimate of the total number of interactions that a protein could be involved in (Rolland et al., 2014; Grossmann et al., 2015). However, the resulting interactome maps are limited by technical aspects of the underlying technology, both in terms of false positives (false interactors detected with the method) and in particular false negatives (true interactors not detected by the method) (Braun et al., 2009). Binary technologies need forced expression of both bait and prey, and this often in a limited set of possible cell types, while co-complex technologies generally do not force prey expression and are more flexible in the cell types that can be examined. Nevertheless, most systematic screens were mainly performed in typical cellular model systems, largely driven by practical considerations (Couzens et al., 2013; Huttlin et al., 2015).

No single technology currently addresses all the different aspects of PPIs, raising not only the need for complementary binary assays to define the total interactome (Braun et al., 2009), but also the need for different co-complex methods to identify the cell-specific complexes in a comprehensive way (Couzens et al., 2013). Note that many proteins are a member of different complexes, often depending on external conditions such as extracellular stimuli to activate pathways.

The proper framework to pre-process and filter the data originating from these experiments is therefore a critical aspect of the analysis. However, even if the interactions found by these technologies would perfectly represent the ground truth, simply constructing a network of protein interactions would not be the end point of the study. Indeed, the true potential of collecting these

interactions is to actually use them to learn more about the signalling and interaction pathways in living organisms and to gain new biological insights. This requires methods beyond those that merely pre-process the data. To this end, techniques must be used that find interesting answers to relevant research questions about these interaction networks and represent them in a way that is readily interpreted.

In this review, we give an outline of how to go from the raw data collected from a co-complex MS experiment to new biological knowledge (Figure 1). The goal is to explore the most commonly used procedures and highlight some key emergent concepts that are expected to have a large impact on the field, without aiming to exhaustively cover all available options.

II. Reliability of interactions from co-complex experiments

The interpretation of the output of co-complex technologies is generally hampered by both false negatives and false positives. One of the causes of the high number of false negatives (which are often weak or transient interactions) is the homogenization or lysis, and the use of (harsh) washing steps required in many of the co-complex protocols. Therefore, novel techniques are getting developed that are 'lysis-free' or stabilize complexes, like BioID (Roux et al., 2012) and Virotrap (Eyckerman et al., submitted). Some recent techniques focus on co-elution, instead of specific purification (Havugimana et al., 2012; Kristensen et al., 2012). The undersampling by mass spectrometry instruments is another cause of the high number of false negatives. New data independent approaches (DIA) aim to reduce this problem, like Affinity Purification combined with Sequential Window Acquisition of all THeoretical spectra (AP-SWATH) (Gillet et al., 2012; Collins et al., 2013), but part of the solution can also come from spectral matching between runs (e.g. as implemented in MaxQuant/Perseus) (Cox & Mann, 2008). On the other hand, many of the false positives in the output of co-complex technologies are due to highly abundant proteins. These proteins are for example contaminants that are introduced by the technical handling of the samples (e.g. keratins), that stick to the purification matrix or the (often overexpressed and hence sometimes misfolded) bait (e.g. chaperones), or that are generally very highly expressed (e.g. ribosomes and cytoskeletal proteins).

Some of the causes of these false positives can be partly tackled by using "lysis-free" methods (Roux et al., 2012), by limiting overexpression (Couzens et al., 2013), or by purifying endogenous complexes (Malovannaya et al., 2011; Gibson et al., 2013). Nevertheless, there is a constantly increasing need for filter techniques to differentiate true from false positives, and this need will grow even more with the continuous increase in sensitivity of mass spectrometry instruments that can currently generate protein lists containing hundreds of proteins per co-complex MS experiment (Selbach & Mann, 2006; Mellacheruvu et al., 2013).

An important consideration is the inclusion of background or unrelated experiments in the design of the experiment. It is clear that the choice and extent of background experiments affect the filtering profoundly with varying sensitivities coupled to the different filter approaches. The actual sample composition (only purification matrix, matrix with unspecific antibody, unrelated baits) is an important factor in successful filtering and should be carefully considered before starting a purification experiment.

III. The source data of a co-complex experiment

After pull-down with a tagged bait or equivalent, the samples are subjected to mass spectrometry analysis to identify proteins that have remained in a complex with the bait. The primary data source for further analysis will arise from identification software such as SEQUEST, Mascot or MaxQuant (Eng et al., 1994; Perkins et al., 1999; Cox & Mann, 2008). The different layers in the identification process result in a range of features for further analysis, such as protein identities, identification scores, peptide counts (and scores), spectral counts and intensities. The most basic feature, 'protein presence' (the presence of a certain protein identifier), is simply a binary variable corresponding to whether a protein was detected as a prey or not in a given sample, and is a commonly used feature in a large number of processing methods. More recent methods make use of features related to the abundance and confidence of each protein, typically the spectral count or intensity of the prey protein (or a derived feature). The 'spectral count' of a protein is defined as the total number of peptide-to-spectrum matches assigned to a specific protein in the project of interest. The 'protein intensity' corresponds to the measured intensity of the spectra assigned to the protein at the MS1 stage. Both have been shown to be relatively good proxies for the abundance of the protein in the sample (Old et al., 2005; Gingras et al., 2007; Asara et al., 2008) and are therefore often used. Other commonly used features that also reflect the confidence of the protein observed in the sample include the 'peptide count' of a protein (the *unique* peptide-to-spectrum matches), and the identification scores.

IV. Strategies for filtering false positive interactions

Several steps can be undertaken to remove the false positive interactions found in a co-complex MS experiment. Table 1 gives an overview of several pipelines and the corresponding features. The available options often depend on the type and scale of the performed experiment. For example, simple approaches include retaining only the preys that occur in all biological repeats with the same bait (Glatter et al., 2009), or the interactions where a reciprocal match is found with each interaction partner in turn as bait and prey (Butland et al., 2005). However, both approaches are often avoided because they require significant investment to repeat or perform new experiments and are not always feasible due to technical limitations. Furthermore, they still retain a large fraction of false positives at the expense of losing many true positive interactions due to their intrinsic simplicity.

The most common approach to eliminate the false positives in a co-complex MS experiment involves quantifying the confidence for each measured interaction and comparing this to a specific cut-off to remove preys that are likely contaminants or indirect binders, *i.e.* a filter for false positives. In general, when determining this cut-off, there is a trade-off between the false positive and false negative rate. Stricter cut-offs will result in less contaminant proteins, but will return less reported interactions. Mild cut-offs will give plenty of results, but may have a high fraction of false interactions. The choice where to end up in this spectrum is highly dependent on the goal of the constructed interaction network, which is addressed further below. Typically

the confidence value assigned to an interaction is based on a combination of the promiscuity of the proteins, i.e. how often do the involved proteins occur in other purifications; the reproducibility of the interaction, i.e. does this interaction consistently occur in repeats; and/or the protein concentration of the interacting partners, usually determined by their spectral counts or protein intensity measurement.

The choice of the filtering approach is dependent on several factors, such as the size of the experiment and the connectivity of bait proteins, but each filtering approach makes several assumptions about the true interaction network composition. As the topology of the protein complex methods is unknown, a critical assumption that these filtering approaches must make fits in two possible models: the 'spoke' model or the 'matrix' model. In the 'spoke' model, interactions are assumed between the bait and each identified prey, but not between the preys, while in the 'matrix' model, the identified preys are assumed to also interact with each other. Comprehensive reviews about the inner workings of these filtering approaches are available (Armean et al., 2012; Nesvizhskii, 2012). Here we will provide a brief overview by dividing the methods in three categories based on their intended use and required input parameters. The first category includes filtering approaches based on straightforward calculations that are mostly independent from the application setting. They typically involve count data or intensity values. The second category corresponds to parametric approaches where one or more parameters must be tuned to fit the data set. The confidence values for each potential interaction can then be modelled. The third category consists of methods that attempt to identify clusters of co-occurring proteins instead of quantifying each individual interaction. In the next sections we will further describe each of these categories.

A. Non-parametric filtering

Non-parametric filtering approaches are defined as those that do not require any parameter tuning for the given data set and often consist of only a few straightforward calculations. These approaches are often expert-driven, with the most basic versions typically more based on what the expected experimental outcome should be like, rather than fitting a statistical model. As such, these methods are sometimes called 'empirical methods'. The cut-offs that are used in the filters of the more basic versions are often estimated from synthetic data or golden standard data sets, although some variants estimate the cut-offs directly from the underlying data. Furthermore, several of the basic approaches are embedded in more complex methods, e.g. SAI (Gavin et al., 2006), often as a first step to remove the most obvious of false positives.

Some of the most straightforward non-parametric filters remove potential interactors that are found at frequencies above those observed in the negative control samples (Arifuzzaman et al., 2006; Ewing et al., 2007). Alternative straightforward approaches eliminate proteins that co-elute with a large fraction of the tested baits, or directly remove all ribosomal proteins, a pool of common contaminants (Ho et al., 2002). These are all basic filtering operations that already have a large effect on the false positive ratio of the final network as they will remove most clearly promiscuous preys. Nevertheless, these approaches are far from stringent and may often also arbitrarily remove many true interactors.

Because of the clear shortcomings of the most basic non-parametric filtering approaches, several filters scale the input data to correct for biases e.g. introduced by bait or prey abundance differences. These scaled data are thereafter used to estimate confidences and are compared to expert-based cut-offs or basic statistical distributions that are inherent to the underlying data and hence do not need parameter input. Methods of this type include the CompPASS Z- and D-scores (Sowa et al., 2009) and SFINX (Titeca et al., submitted). Many of the non-parametric approaches are embedded in extensive pipelines that make use of several filters. One such 'hybrid' example is the typical usage of the Socio-affinity index, which logarithmically converts the basic spectral count data, while trying to correct bait and prey biases in order to yield a mutual co-occurrence matrix that is then used as input for a cluster-based filter (Gavin et al., 2006; Kühner et al., 2009). Another example is the PP-NSAF method (Posterior Probabilities based on Normalized Spectral Abundance Factors), which requires a complex series of many different filtering techniques (Sardiu et al., 2008).

While the principle of Occam's razor - to select the solution with the least assumptions - underlies all filter techniques, it is absolutely critical for the non-parametric methods. These techniques have to get their power from their user-friendliness and algorithmic straightforwardness, from their tight link with the experimental experts, and from avoiding external data integration and parameter optimization.

B. Parametric filtering

The parametric filtering approaches are defined as those that require the explicit setting or estimation of (several) parameters for each new data set. These parameters are often calculated based on the negative control samples. Typically a background distribution is calculated against which each interaction is tested. In this case the confidence value tested against the cut-off will have an explicit meaning, such as the probability of a given interaction being a true positive based on the trained null model. The cut-off of these approaches is therefore based on statistical significance rather than tuned towards an expected result. The used models for which the parameters must be tuned, range from simple statistical tests to complex probabilistic models.

Basic statistical filters are often built upon the assumption that the false positive interactions follow a well-characterized distribution, such as the normal distribution (Hubner & Mann, 2011; Malovannaya et al., 2011) or the hypergeometric distribution (Hart et al., 2007). However, such approaches often require the imputation of missing values in the data matrix. In addition, probabilistic approaches built on a Bayesian principle are frequently used to estimate the probability of each interaction given the prior distribution of the evidence (Collins et al., 2007; Choi et al., 2010, 2012; Lavallée-Adam et al., 2011; Skarra et al., 2011). The advantage of a Bayesian approach is its versatility and its possibilities for the integration of evidence sources, even those from external data sources as discussed later. Furthermore, the Bayesian approach should also be inherently more stable under shortage of underlying data, as it allows approximations of the true values. The most prominent disadvantage is that they require explicit modelling or assumptions regarding the distributions of many of these features, and the need for the incorporation of external data and sometimes extensive parameter optimization might also introduce significant complications, as will be discussed in later paragraphs.

C. Cluster-based filtering

A final broad category of often-used filtering approaches clusters proteins together that share many potential interactors. This conceptual difference is what distinguishes them from the previous two categories. Instead of regarding the problem as a data matrix that needs to be filtered in a pairwise fashion and assigning each interaction pair a confidence value, they attempt to identify those sets of proteins that co-elute as a group and are therefore likely part of the same complex. Derived confidence values or cut-offs are then evaluated at the group level and not at the level of individual interactions. Two subcategories are defined, those that find groupings in the data matrix and those that find groupings in a graph representation where each node is a protein and each edge is a possible interaction. The former are often based on unsupervised classification methods, such as hierarchical clustering or biclustering (Gavin et al., 2006; Sardiu et al., 2008; Choi et al., 2010; Xie et al., 2011). The latter of these approaches consists of the identification of 'locally dense regions' in the interaction network (Enright et al., 2002; Bader & Hogue, 2003; Newman & Girvan, 2004). In the most extreme case these will be cliques, i.e. sets of proteins that all share reciprocal interactions (Zhang et al., 2008). The networks to be clustered can include weighted edges, where each potential interaction is given a confidence score based on a previous filtering step or based on orthogonal evidence, which can be used in probabilistic approaches (Asthana et al., 2004).

The data output of the co-complex techniques AP-MS, tandem affinity purification coupled to mass spectrometry (TAP-MS), immunoprecipitation coupled to mass spectrometry (IP-MS), Virotrap and BioID show many similarities. Hence, several of the described algorithms from section IV will be able to analyse data from multiple of these sources. Nevertheless, both the co-elution approaches and AP-SWATH generate fundamentally different forms of data output. The co-elution approaches are only focussed on very large interaction networks and often employ customized data analysis pipelines that first search for correlations between chromatographic elution profiles before performing other steps. Some (Havugimana et al., 2012) heavily use external data for the filtering, while others do not (Kristensen et al., 2012). For AP-SWATH data, the filtering originally happened by a customized non-parametric pipeline (Lambert et al., 2013) and later by one of the existing Bayesian parametric approaches (Tsou et al., 2015). Whether the other described filter approaches can also handle this type of data is still unknown.

V. Use of external protein interaction data for contaminant removal

Nowadays, the available data are not limited to the data generated by your own experiment. Various online databases include a myriad of protein interaction networks, interaction predictions and functional annotation of proteins. Such data are not only valuable for prioritizing the most promising interaction candidates, but are also useful for the primary removal of potential false positives.

A. Reference interaction data sets

Different collections of characterized protein interactions exist for a variety of species. These include a small number of high quality protein complexes to entire interaction networks collected from different sources. However, a good golden standard data set must be of high

quality, as any false positive interaction might skew the results. Nowadays, several large databases have been constructed containing a large number of protein interactions, of which table 2 lists the most frequently used. We distinguish three types of databases: databases with only curated interactions from experimental data, databases which include predictions, and meta-databases that integrate the others. The databases that contain only curated interactions from experimental data, such as IntAct and MIPS (Pagel et al., 2005; Kerrien et al., 2012), are expected to contain the lowest number of false positives and are therefore commonly used for validation of co-complex MS experimental data. The databases that include interaction predictions, e.g. STRING, contain interaction networks for a large number of organisms based on text mining, curated knowledge and predictions based on functional features (Franceschini et al., 2013). Such databases are expected to contain a larger fraction of false positives but are useful for lesser-studied organisms and very valuable as an additional filter. While in the databases that integrate several other databases, such as iRefIndex and DroID (Razick et al., 2008; Yu et al., 2008), the number of possible false positive interactions depends mostly on the databases that they include and the manner in which they are integrated.

It is common to not only compare with known true interactions, but also with known negative interactions, such as those collected in the Negatome database or the CRAPome. The Negatome includes confirmed negative interactions manually curated from studies in the literature and indirect binders from protein complexes as stored in PDB (Kouranov et al., 2006; Smialowski et al., 2009). This collection thus contains a set of several thousands of protein pairs that are known not to be direct binders. The CRAPome on the other hand is a collection of negative control samples from different co-complex MS experiments (Mellacheruvu et al., 2013). As the eluted proteins in a negative control should be bait-independent, this data set can be used as an additional control for any similar co-purification experiment.

B. Integration of reference data into the filtering step

We distinguish two possible approaches to use the interaction data from public databases in the filtering framework. In the first approach, the data are directly integrated into the filtering approach, in order to boost true positive interaction prediction accuracy from co-complex MS data. For example, the parametric filter SAINTexpress incorporates the protein interaction data from the iRefIndex database among other features in a probabilistic model (Razick et al., 2008; Teo et al., 2014), so that preys that are known interaction partners have a boosted probability of eluting with the same bait. Using similar approaches, other methods have also integrated prediction data from prediction databases, such as STRING (Franceschini et al., 2013), so that protein pairs that have been predicted to interact will be assigned a higher confidence value.

In the second approach, the known protein interaction data are indirectly used to train supervised filtering models that predict the confidence of an interaction. These supervised methods are trained using a wide spectrum of features. The possible features include those generated or derived from the co-complex MS experiment itself, such as protein concentration measures, identification scores, or final network topologies, but also additional external features, such as protein functional data, orthology information or co-expression information, as discussed in the section on prioritization. These features are then integrated into a supervised framework, such as a random forest or a logistic regression model, to predict the confidence for

all potential interaction partners (Cloutier et al., 2009; Havugimana et al., 2012; Huttlin et al., 2015). The advantage of such an approach is that the contribution of each feature to the final model is trained on the dataset and requires little to no expert oversight. However, it does require a good training data set that is not only representative for the co-complex MS experiment that it will be applied on but is also free from any experimental or curation biases. Good training sets are nevertheless hard to find.

VI. Annotation and prioritization of candidate partners

A common task after the interactions have been filtered for false positives is to annotate and prioritize the remaining interactions for further (experimental) validation. In general, orthogonal evidence from a variety of sources is used to select the most interesting and the most relevant interactions. To this end, the confidence scores from the filtering methods can be used to identify the interactions with the highest quality. However, these scores might be biased towards certain interactions as many filtering approaches nowadays explicitly use external data sources or are trained on a validated data set. Furthermore, many filtering methods do not have sufficient resolution to distinguish between a good and an excellent interactor, resulting in a pool of hundreds or thousands of partners that need further testing.

A. Based on functional data

Most co-complex experimental techniques do not reveal anything about the functional characteristics of the interaction, such as where in the cell it takes place or its consequence for the cellular processes. Integration of external functional information about the involved proteins can be used to prioritize interactions based on a specific localization or pathway of interest. Several sources of protein functional annotation are available which are used in this context, such as Pfam protein domains, Gene Ontology terms or KEGG pathways (Ashburner et al., 2000; Kanehisa et al., 2012; Punta et al., 2012). Interacting proteins are known to share similar annotation with regards to their composition, their biological function and their cellular location. Therefore, statistics of assigning protein complexes to a specific function are often based on identifying an overrepresentation of overlapping terms, using methods such as a Jaccard index or a Hypergeometric test. However, there are several caveats with such an approach. First, any annotation of proteins is likely to be incomplete and sometimes incorrect as our knowledge of protein function is still greatly lacking. Secondly, any statistical test must take great care to account for the proper background set as not all proteins are identified during a co-complex experiment, nor do they all have annotation. Finally, many annotations have a strict hierarchy, which must be accounted for in any test as each term is not independent from another.

Several other functional characteristics are also typically used for prioritization. Interacting proteins also have a higher chance to be co-regulated and thus to be co-transcribed under similar conditions (Ge et al., 2001; Deane et al., 2002; Jansen et al., 2002). Hence, a common procedure is to check if a putative interaction pair is also co-expressed in a gene expression compendium under the experimental conditions of interest. Many online expression databases that collect gene expression information now allow such targeted queries, such as in Colombos, GEO or ArrayExpress (Barrett et al., 2013; Rustici et al., 2013; Meysman et al., 2014). The actual check for co-expression can involve straightforward correlation analysis to more

advanced biclustering approaches (Eren et al., 2012; Naulaerts et al., 2015). In addition, co-regulation can be ascertained using transcription factor target annotation from databases such as TRANSFAC (Wingender et al., 1996), with similar overrepresentation tests as those described in the previous paragraph.

B. Based on orthology information

In the past, co-complex MS experiments were only performed on well-studied model organisms, such as yeast or human. However, in recent years, many studies have been performed on other organisms (Fernández et al., 2009; Van Leene et al., 2010; Jäger et al., 2011; Płociński et al., 2014). As only limited functional data might be available for such a species, researchers commonly map the proteins from the studied species to a model organism to identify so-called 'orthologs' and use its functional data. Orthology mapping is typically done based on protein sequence similarity using one of the many methods that are available, such as OrthoMCL or Inparanoid (Li et al., 2003; Ostlund et al., 2010). However, nowadays many databases exists that contain this information; such as KEGG orthology, Ensembl-compara, EggNOG, OMA orthology and COG databases (Kanehisa et al., 2012; Flicek et al., 2014; Powell et al., 2014; Altenhoff et al., 2015; Galperin et al., 2015). Detailed reviews are available that describe the advantages and uses of the commonly used orthology mapping methods and databases (Kristensen et al., 2011; Altenhoff & Dessimoz, 2012; Gabaldón & Koonin, 2013).

The same is possible with the interactions themselves by identifying the 'interologs' in another species. This can be useful to involve golden standard datasets or functional information at the interaction level, such as phosphorylation targets. The actual mapping of the interaction can be done using the straightforward reasoning that any orthologs with sufficient sequence homology of two interaction partners are likely to also interact (Matthews et al., 2001; Butland et al., 2005). However, more sophisticated approaches that correct for the conservation of protein interaction domains and function also exist (Michaut et al., 2008). The main caveat with these approaches for mapping information across species is that this only works for proteins for which homologs are available.

VII. Strategies for prediction of false negative interactions

In parallel with the annotation and prioritization of interaction partners, researchers are often interested in the prediction of possible false negatives to compensate for technical and biological limitations. The combination of any PPI technique with complementary experimental techniques, like other PPI techniques or even genetic interaction techniques, helps to identify false negatives, and the analysis of expression profiles can point to cell specificities or stimulus dependencies. Nevertheless, *in silico* techniques have become an increasingly powerful and useful alternative to the often labour intensive and costly orthogonal experimental techniques. Broadly accessible databases, like those described in table 2, are rich in relevant information and enable the researcher to compare the obtained results. Furthermore, other databases contain extensive protein expression profiles, like the Human Protein Atlas (Uhlen et al., 2010) or RNA-Seq and gene expression microarray profiles (Su et al., 2004; Hruz et al., 2008; Wang et al., 2008; Krupp et al., 2012), which for example help to find cell or tissue specificities.

Furthermore, many algorithms exist for the prediction of PPIs, but all inherently pose the risk of overinterpreting the experimental data. Hence, this type of predictions is omitted from many co-complex MS analyses. These algorithms typically belong to one of four categories, based on the information that is used in the prediction: genetic location, protein structure, network topology or a heterogeneous combination. The algorithms using genetic location features take into account close localisation of genes in the genome (Tamames et al., 1997; Overbeek et al., 1999), evolutionary gene-fusion events (Enright et al., 1999; Marcotte et al., 1999), or phylogenic conservation of gene order and location (Dandekar, 1998; Gaasterland & Ragan, 1998; Pellegrini et al., 1999; Goh et al., 2000; Huynen, 2000; Pazos & Valencia, 2001; Juan et al., 2008), but most of these are only useful for prokaryotes. The algorithms using the protein structure take into account the primary protein structure with the detection of short relevant sequences for protein interaction (Bock & Gough, 2001; Sprinzak & Margalit, 2001; Gomez et al., 2003; Martin et al., 2005; Pitre et al., 2006; Najafabadi & Salavati, 2008; Guo et al., 2010; Yu et al., 2010), or the secondary or tertiary structure (Edwards et al., 2002; Smith & Sternberg, 2002; Aloy & Russell, 2003; Hue et al., 2010; Singh et al., 2011; Wass et al., 2011). These are generally more functionally relevant but limited by our knowledge about the structure of the involved proteins, although the Protein Data Bank (PDB) (Kouranov et al., 2006) has been growing steadily and homologous structures can also serve well in these predictions (Zhang et al., 2012). The algorithms based on network topology take into account features that separate PPI-networks from random ones, like preferential attachment (Barabási, 1999), and can prove powerful for larger networks (Goldberg & Roth, 2003; Wuchty, 2006; Yu et al., 2006; Chen et al., 2008). The last group of algorithms combine many different features within classifier machine learning techniques, like support vector machines or random forests, trained on large high-quality positive and negative training sets (Jansen et al., 2003; Ben-Hur & Noble, 2005; Chen & Liu, 2005; Rhodes et al., 2005; Guo et al., 2008; Xia et al., 2010; Lin & Chen, 2013; Kotlyar et al., 2015).

VIII. From processed interactions to biological insight

Once a high quality protein interaction network has been constructed, the final step is to extract relevant biological knowledge from the identified protein interactions. This can occur at different scales, namely from studies concerning only a single protein complex to those that encompass the entire network.

A. Analysis of specific protein complexes or interactions

Targeted functional analysis of a single or a small set of proteins often implies tedious literature surveys. Some solutions exist to support such searches. Various public repositories, such as NCBI and UniProt, allow look-up of single proteins and offer a variety of curated information. Such databases are often used as the first step into a more in-depth survey. Further functional information, such as the annotated protein domains, gives insight into the nature of the interaction. For example, interaction with proteins containing kinase domains might suggest phosphorylation, which could be further validated using phosphorylation prediction tools (Dang et al., 2008; Xue et al., 2008; Fermin et al., 2015). In addition, pathway analysis software tools, like KEGG Atlas, Ingenuity Pathway Analysis, MetaCore or BioCyc Omics Viewer, are often

used to aid in the further functional characterization of found interactions and protein complexes (Caspi et al., 2012; Kanehisa et al., 2012). Such tools allow users to place their protein complexes of interest within the greater scope of a specific pathway or functional category. An extensive review of approaches for the functional interpretation of proteome data is described in (Laukens et al., 2014).

B. Analysis of protein complex stoichiometry

The composition of a protein complex plays a large role in its function and its inner workings. As source data such as protein intensities are considered a proxy for protein abundance, a commonly used tactic is to infer the stoichiometry underlying the protein complexes based on this information (Smits et al., 2013; van Nuland et al., 2013). In such a case these values must be corrected for the size and observability of the protein, upon which the ratio between interacting proteins represents the stoichiometry. It has been shown that such a relatively straightforward procedure is sufficiently suitable to study the protein content in a complex (Fabre et al., 2014).

C. Analysis of protein complex structures

Once a potential interaction between proteins has been identified, the three-dimensional molecular structure of the complex can be revealed through a more detailed analysis (Janin & Séraphin, 2003). Both experimental and computational procedures exist to characterize the protein complex structure (Russell et al., 2004; Melquiond et al., 2012). Experimentally, one can express the postulated complex subunits together and perform cryo-electron-microscopy, nuclear magnetic resonance (NMR) or X-ray crystallography of the subsequent complex (Nogales et al., 1998; Fieulaine et al., 2002; Terrak et al., 2004). The difficulty of experimentally characterizing a protein structure has lead to the development of computational protein-protein docking methods to model the complex structure of interacting proteins, such as HADDOCK, SwarmDock, ClusPro and ZDOCK (Comeau et al., 2003; van Dijk et al., 2006; Torchala et al., 2013; Pierce et al., 2014). Protein-protein docking is substantially more complex than the type of prediction introduced in section VII, as it not only involves predicting whether proteins interact but also how. A wide range of docking methods are available and several reviews have provided an overview of the field (Moreira et al., 2010; Kozakov et al., 2013; Huang, 2014). In brief, the starting point for protein-protein docking typically requires known crystallography structures for the individual subunits of the complex. These methods will attempt to identify the interface alignment for the involved proteins and any conformational changes that may occur during complex formation. Some tools are able to also start from reliable structural models based on close homologs. However, the quality of the starting structures has a great impact on the reliability of the eventual docking prediction. An independent evaluation of the performance of various docking prediction methods is frequently assessed during the CAPRI (Critical Assessment of PRedicted Interactions) challenges organized by EMBL-EBI. There are several prediction rounds organised per year, where various research groups are tasked with using their methods to predict a set of protein complex structures that have been recently solved but are still unpublished. The performance of the methods are then evaluated based on the similarity of their prediction to the solved molecular structure (Fernández-Recio & Sternberg, 2010; Lensink & Wodak, 2013).

D. Analysis of the entire network

As mentioned previously, the collection of a large number of protein interactions can be conceptualized as an undirected graph where each node is a protein and each edge represents a potential physical interaction. These interaction networks have certain typical topological features. For example, they are often considered scale-free, i.e. the degree distribution obeys a power-law, so that there is a small number of hub proteins with many interactions and a large number of proteins with few interactions (Barabási, 1999). Various topological features, such as centrality, average shortest path and network diameter, can be investigated for any protein interaction network to give an idea of the density or connectedness of the proteins. Different data mining methods can be applied at the network level to extract novel and potentially interesting patterns (Naulaerts et al., 2015). For example, a typical problem at the topological level is to identify the subgraphs within the network. These subgraphs are collections of nodes and edges in a pattern or motif that appears several times within the network, and may thus have a functional purpose. Different approaches exist to find such frequent subgraphs (Ghazizadeh & Chawathe, 2002; Przulj et al., 2004; Jiang et al., 2012). A distinction is made between induced subgraphs and partial subgraphs. Induced subgraphs, or 'graphlets', require any matches to nodes in the network to also feature the same interconnecting edges without any additional ones. The distribution of the graphlets in the protein interaction network gives valuable information into the overall topology and the type of protein interactions that are present, which can then be coupled to the biological functions of the protein (Milenković & Przulj, 2008; Davis et al., 2015). Partial subgraphs or 'network motifs' allow additional edges to be present between the matching nodes. The enrichment of network motifs in the network is typically determined by comparing to random background networks, generated with an appropriate degree model (Yeger-Lotem et al., 2004; Ciriello & Guerra, 2008). However, in most cases, the final protein interaction network is not fully connected, and any analysis tool must be able to deal with this fact or the network must be split into smaller connected subnetworks.

E. Visualisation of protein interaction networks

The most commonly used visual representation of a protein interaction network is a graph as introduced in the previous paragraph. However, this visualization becomes unreadable for dense networks involving many baits and preys. Therefore, other variants such as bait-bait connectivity maps, which only visualize the shared interactors between baits, are also frequently used (Ewing et al., 2007). The shape, colour and size of the nodes and edges can be linked to specific characteristics of the proteins and interactions, such as their function or their certainty. The represented networks can be large, i.e. spanning an entire characterized interactome, or small, focusing on only one or a few protein complexes. The layout of the network, i.e. the positioning of the nodes, is typically based on grouping dense interactions clusters or on grouping proteins with similar biological functions. Many standalone tools are available for network visualisation (Goldovsky et al., 2005; Brown et al., 2009; Hu et al., 2009; Smoot et al., 2011); table 3 gives an overview of the most commonly used ones. Each tool has its advantages and the choice of the tool will typically depend on two factors. The first factor is the size and density of the network. Many tools do not support full protein network representation or have difficulties to generate clear overviews with the available layouts. The second factor is the goal of the network visualisation. Many tools offer a great range of options and extensions to

allow users to tune the look of their graph or perform topological analyses such as those discussed in the previous section. Several comprehensive comparative reviews into the technicalities of these visualisation tools are available (Suderman & Hallett, 2007; Pavlopoulos et al., 2008; Gehlenborg et al., 2010).

An upcoming trend is to share protein interaction networks from co-complex MS experiments online, i.e. in 'the cloud', in conjunction with the publication so that readers can explore these for themselves (Havugimana et al., 2012). Different libraries are now available that allow construction of such a web-based network to be placed on a webserver, such as the Cytoscape.js (http://cytoscape.github.io/cytoscape.js/) or D3.js (http://d3js.org) JavaScript libraries. These allow design of dynamic and interactive networks of relatively large size that will run on the commonly used browsers. Furthermore, several standalone visualisation tools support the export of network visualization to web-compatible formats.

IX. Conclusions and future prospects

In this review, we discussed the major computational challenges and solutions to process and interpret co-complex experimental results. Essential steps include identification of the proteins, filtering of the false positives (using both internal and external data), validating the found interactions, making new biological-relevant discoveries and visualising these results.

A. Different analytical methods for different experimental setups

The choice of the filtering strategy for primary false positive removal is an essential but challenging one, which is not facilitated by the myriad of available options. One of the main elements that determines the choice of the technology is the size of the dataset. Cluster-based filters and some specific non-parametric filters are typically most relevant for the analysis of complete interactomes and dense networks, but often underperform on smaller subnetworks. At the other side of the spectrum, the parametric filters, especially those based on Bayesian frameworks, have a strong performance on the smaller networks with incomplete data (Pu et al., 2015).

However, one main disadvantage of many of these methods is that they are specifically tuned towards one type of experimental setup, or even one specific experiment, resulting in poor performance on other data (Choi et al., 2012; Nesvizhskii, 2012; Pu et al., 2015). This means that it remains hard to know which method will perform best on your data, as each method has been optimized for their own test case. There is no guarantee that you will achieve similar results as those that have been reported, even if the experimental setup seems similar. There is a great need for unbiased comparisons of different approaches on a variety of good golden standard datasets, but the quality of golden standard datasets also urgently needs improvement (Pu et al., 2015). These golden standard datasets are good for the raw comparative estimation of accuracies, but detailed close-call comparisons should be handled prudently. Furthermore, the tendency towards dedicated approaches tuned to specific data sets means that every new technological platform for co-complex studies would need to be accompanied by the creation of new optimized filtering strategies for removing false positives, which is far from efficient. As we can expect more variations on the traditional co-complex MS experiment to occur with mass

spectrometry becoming cheaper and more accurate, there will be a need for generic preprocessing frameworks that either combine various methods or are sufficiently robust by introducing the least amount of biases and assumptions. In this case, the non-parametric filters have the potential for the broadest applicability, as long as they focus on algorithmic simplicity and experimental expertise, and stay away from extensive parameter optimisation and integration of external data sources. Approaches that combine several known filters with downstream analysis and visualisation functions could also be an intermediary filtering solution (Krumsiek et al., 2008), and may yield an even more accurate approach upon correct combination, just like the algorithms in the peptide identification field have been advanced in a similar way (Vaudel et al., 2015).

B. Rise of standard workflows and quality control

A recent evolution in the field of co-complex MS and MS in general is a focus on the quality control and reproducibility of experiments. There is a growing need to address the reliability of the protein interactions, both for clinical and fundamental research purposes. A recent study revealed that using a standardized workflow a reproducibility of 81% is possible for a single AP-MS experiment performed in different labs (Varjosalo et al., 2013). Hence, any co-complex MS technique must be robust at all levels, from sample preparation to network inference, and each step has an essential impact on the final experimental result. First, the choice of sample preparation and purification technologies is expected to result in different outcomes due to the intrinsic biases of each method. Even when the same standard protocol is followed, experimental variation is a factor that needs to be characterised and must be accounted for. Second, the analytical instruments show variation in performance, both at the level of chromatography, ionization, MS1 and MS2 and ion selection for MS2. Several papers have discussed the need for quality control. Rudnick and coworkers have proposed 46 performance metrics to monitor the consistency of instrument performance (Rudnick et al., 2010). Recently qcML was introduced, a standard format for the exchange of quality control data (Walzer et al., 2014). In addition, several software tools and frameworks have been released to handle QC and instrument performance metrics (Sturm et al., 2008; Pichler et al., 2012; Bittremieux et al., 2014, 2015). It is anticipated that these developments will drive the adoption of stringent quality control in MS in general, and more specifically also lead to further improved consistency across multiple co-complex MS experiments once the community picks them up. Finally, the method used to infer networks from experimental co-complex MS data impacts the outcome to a substantial extent. A recent comparative analysis of six scoring methods showed not only poor overlap, but also a highly variable performance dependent on the dataset, which was both attributed to sensitivity to noise in the experimental data and biases arising from the Golden Standard datasets used for tuning and training network inference methods (Pu et al., 2015). This leaves important challenges on the road to reproducible and robust co-complex MS based interactomics.

X. Acknowledgements

This work was supported by the Research Foundation-Flanders (FWO) project “Evolving graphs” (G.0903.13N); and the agency for Innovation by Science and Technology (IWT) SBO project “InSPECtor” (120025) and a personal grant to KT.

XI. References

Aloy P, Russell RB. 2003. InterPreTS: protein Interaction Prediction through Tertiary Structure. *Bioinformatics* **19**:161–162.

Altenhoff AM, Dessimoz C. 2012. Inferring orthology and paralogy. *Methods in molecular biology (Clifton, N.J.)* **855**:259–79.

Altenhoff AM, Škunca N, Glover N, Train C-M, Sueki A, Piližota I, Gori K, Tomiczek B, Müller S, Redestig H, Gonnet GH, Dessimoz C. 2015. The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic acids research* **43**:D240–9.

Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, Saito R, Ara T, Nakahigashi K, Huang H, 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. 2006. Large-scale identification of protein – protein interaction of Escherichia coli K-12. *Genome Research* **16**:686–691.

Armean IM, Lilley KS, Trotter MWB. 2012. Popular computational methods to assess multi-protein complexes derived from label-free affinity purification and mass spectrometry (AP-MS) experiments. *Molecular & Cellular Proteomics* **12**:1–13.

Asara JM, Christofk HR, Freimark LM, Cantley LC. 2008. A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen. *Proteomics* **8**:994–999.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**:25–9.

Asthana S, King OD, Gibbons FD, Roth FP. 2004. Predicting protein complex membership using probabilistic network reliability. *Genome Research* **14**:1170–1175.

Bader GG, Hogue CC. 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**:2.

Barabási A. 1999. Emergence of Scaling in Random Networks. *Science* **286**:509–512.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. 2013. NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research* **41**:D991–5.

Ben-Hur A, Noble WS. 2005. Kernel methods for predicting protein-protein interactions. *Bioinformatics (Oxford, England)* **21 Suppl 1**:i38–46.

Bittremieux W, Kelchtermans P, Valkenborg D, Martens L, Laukens K. 2014. jqcML: an open-source java API for mass spectrometry quality control data in the qcML format. *Journal of proteome research* **13**:3484–7.

Bittremieux W, Willems H, Kelchtermans P, Martens L, Laukens K, Valkenborg D. 2015. iMonDB: Mass spectrometry quality control through instrument monitoring. *Journal of proteome research*.

Bock JR, Gough DA. 2001. Predicting protein-protein interactions from primary structure. *Bioinformatics* **17**:455–460.

Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, de Smet A-S, Venkatesan K, Rual J-F, Vandenhaute J, Cusick ME, Pawson T, Hill DE, Tavernier J, Wrana JL, Roth FP, Vidal M. 2009. An experimentally derived confidence score for binary protein-protein interactions. *Nature methods* **6**:91–7.

Brown KR, Otasek D, Ali M, McGuffin MJ, Xie W, Devani B, Toch IL van, Jurisica I. 2009. NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics (Oxford, England)* **25**:3327–9.

Butland G, Peregrín-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, Davey M, Parkinson J, Greenblatt J, Emili A, Imaizumi-Anraku H, Webb J, Takeda N, Kojima T, Charpentier M, Oldroyd G, Perry J, Soll J, Miwa H, Kru J, Umehara Y, Kouchi H, Research C, Murakami Y, Mulder L, Vickers K, Pike J, Downie JA, Wang T, Sato S, Asamizu E, Tabata S, Yoshikawa M, Murooka Y, Wu G-J, Kawaguchi M, Kawasaki S, Parniske M, Hayashi M. 2005. Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433**:527–31.

Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD. 2012. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome

databases. *Nucleic acids research* **40**:D742–53.

Chen P-Y, Deane CM, Reinert G. 2008. Predicting and validating protein interactions using network structure. *PLoS computational biology* **4**:e1000118.

Chen X-W, Liu M. 2005. Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics (Oxford, England)* **21**:4394–400.

Choi H, Glatter T, Gstaiger M, Nesvizhskii AI. 2012. SAINT-MS1: Protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *Journal of Proteome Research* **11**:2619–2624.

Choi H, Kim S, Gingras A-C, Nesvizhskii AI. 2010. Analysis of protein complexes through model-based biclustering of label-free quantitative AP-MS data. *Molecular systems biology* **6**:385.

Ciriello G, Guerra C. 2008. A review on models and algorithms for motif discovery in protein-protein interaction networks. *Briefings in functional genomics & proteomics* **7**:147–56.

Cloutier P, Al-Khoury R, Lavallée-Adam M, Faubert D, Jiang H, Poitras C, Bouchard A, Forget D, Blanchette M, Coulombe B. 2009. High-resolution mapping of the protein interaction network for the human transcription machinery and affinity purification of RNA polymerase II-associated complexes. *Methods* **48**:381–386.

Collins MO, Choudhary JS. 2008. Mapping multiprotein complexes by affinity purification and mass spectrometry. *Current opinion in biotechnology* **19**:324–30.

Collins BC, Gillet LC, Rosenberger G, Röst HL, Vichalkovski A, Gstaiger M, Aebersold R. 2013. Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nature methods* **10**:1246–53.

Collins SR, Kemmeren P, Zhao X-C, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS, Krogan NJ. 2007. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Molecular & cellular proteomics : MCP* **6**:439–450.

Comeau SR, Gatchell DW, Vajda S, Camacho CJ. 2003. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* **20**:45–50.

Couzens AL, Knight JDR, Kean MJ, Teo G, Weiss A, Dunham WH, Lin Z-Y, Bagshaw RD, Sicheri F, Pawson T, Wrana JL, Choi H, Gingras A-C. 2013. Protein Interaction Network of the Mammalian Hippo Pathway Reveals Mechanisms of Kinase-Phosphatase Interactions. *Science Signaling* **6**:rs15–rs15.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26**:1367–1372.

Dandekar T. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences* **23**:324–328.

Dang TH, Van Leemput K, Verschoren A, Laukens K. 2008. Prediction of kinase-specific phosphorylation sites using conditional random fields. *Bioinformatics (Oxford, England)* **24**:2857–64.

Davis D, Yaveroğlu ÖN, Malod-Dognin N, Stojmirovic A, Pržulj N. 2015. Topology-function conservation in protein-protein interaction networks. *Bioinformatics (Oxford, England)* **31**:1632–9.

Deane CM, Salwiński Ł, Xenarios I, Eisenberg D. 2002. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular & cellular proteomics : MCP* **1**:349–356.

van Dijk M, van Dijk ADJ, Hsu V, Boelens R, Bonvin AMJJ. 2006. Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic acids research* **34**:3317–25.

Dunham WH, Mullin M, Gingras AC. 2012. Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *Proteomics* **12**:1576–1590.

Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M. 2002. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends in Genetics* **18**:529–536.

Eng JK, McCormack AL, Yates JR. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5**:976–89.

Enright a J, Van Dongen S, Ouzounis C a. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* **30**:1575–1584.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**:86–90.

Eren K, Deveci M, Küçüktunç O, Catalyürek U V. 2012. A comparative analysis of biclustering algorithms for gene expression data. *Briefings in bioinformatics* **14**:279–292.

Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, 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 HS, Stewart II, Kuehl B, Hogue K, Colwill K, Gladwish K, Muskat B, Kinach R, Adams S-L, Moran MF, Morin GB, Topaloglou T, Figeys D. 2007. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular systems biology* **3**:89.

Eyckerman S, Verhee A, der Heyden J V, Lemmens I, Ostade X V, Vandekerckhove J, Tavernier J. 2001. Design and application of a cytokine-receptor-based interaction trap. *Nature cell biology* **3**:1114–9.

Fabre B, Lambour T, Bouyssié D, Menneteau T, Monsarrat B, Burlet-Schiltz O, Bousquet-Dubouch M-P. 2014. Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry. *EuPA Open Proteomics* **4**:82–86.

Fermin D, Avtonomov D, Choi H, Nesvizhskii AI. 2015. LuciPHOr2: site localization of generic post-translational modifications from tandem mass spectrometry data. *Bioinformatics (Oxford, England)* **31**:1141–1143.

Fernández E, Collins MO, Uren RT, Kopanitsa M V, Komiyama NH, Croning MDR, Zografos L, Armstrong JD, Choudhary JS, Grant SGN. 2009. Targeted tandem affinity purification of PSD-95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. *Molecular systems biology* **5**:269.

Fernández-Recio J, Sternberg MJE. 2010. The 4th meeting on the Critical Assessment of Predicted Interaction (CAPRI) held at the Mare Nostrum, Barcelona. *Proteins: Structure, Function, and Bioinformatics* **78**:3065–3066.

Fieulaine S, Morera S, Poncet S, Mijakovic I, Galinier A, Janin J, Deutscher J, Nessler S. 2002. X-ray structure of a bifunctional protein kinase in complex with its protein substrate HPr. *Proceedings of the National Academy of Sciences of the United States of America* **99**:13437–41.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kähäri AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ruffier M, Sheppard D, Taylor K, Thormann A, Trevanion SJ, Vullo A, Wilder SP, Wilson M, Zadissa A, Aken BL, Birney E, Cunningham F, Harrow J, Herrero J, Hubbard TJP, Kinsella R, Muffato M, Parker A, Spudich G, Yates A, Zerbino DR, Searle SMJ. 2014. Ensembl 2014. *Nucleic acids research* **42**:D749–55.

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez

P, Bork P, von Mering C, Jensen LJ. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research* **41**:D808–15.

Gaasterland T, Ragan MA. 1998. Microbial Genescapes: Phyletic and Functional Patterns of ORF Distribution among Prokaryotes. *Microbial & Comparative Genomics* **3**:199–217.

Gabaldón T, Koonin E V. 2013. Functional and evolutionary implications of gene orthology. *Nature reviews. Genetics* **14**:360–6.

Galperin MY, Makarova KS, Wolf YI, Koonin E V. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic acids research* **43**:D261–9.

Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, 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 JM, Kuster B, Bork P, Russell RB, Superti-Furga G. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**:631–636.

Gavin A-C, Maeda K, Kühner S. 2011. Recent advances in charting protein-protein interaction: mass spectrometry-based approaches. *Current opinion in biotechnology* **22**:42–9.

Ge H, Liu Z, Church GM, Vidal M. 2001. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nature genetics* **29**:482–486.

Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs M a, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, Gavin A-C. 2010. Visualization of omics data for systems biology. *Nature methods* **7**:S56–S68.

Ghazizadeh S, Chawathe SS. 2002. SEuS: Structure Extraction Using Summaries. *Proceedings of the 5th International Conference on Discovery Science*:71–85.

Gibson TJ, Seiler M, Veitia RA. 2013. The transience of transient overexpression. *Nature methods* **10**:715–21.

Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, Aebersold R. 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & cellular proteomics : MCP* **11**:O111.016717.

Gingras A-C, Gstaiger M, Raught B, Aebersold R. 2007. Analysis of protein complexes using mass spectrometry. *Nature reviews. Molecular cell biology* **8**:645–654.

Glatter T, Wepf A, Aebersold R, Gstaiger M. 2009. An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Molecular systems biology* **5**:237.

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with their interaction partners. *Journal of molecular biology* **299**:283–93.

Goldberg DS, Roth FP. 2003. Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences of the United States of America* **100**:4372–6.

Goldovsky L, Cases I, Enright AJ, Ouzounis CA. 2005. BioLayout(Java): versatile network visualisation of structural and functional relationships. *Applied bioinformatics* **4**:71–4.

Gomez SM, Noble WS, Rzhetsky A. 2003. Learning to predict protein-protein interactions from protein sequences. *Bioinformatics* **19**:1875–1881.

Grossmann A, Benlasfer N, Birth P, Hegele A, Wachsmuth F, Apelt L, Stelzl U. 2015. Phospho-tyrosine dependent protein-protein interaction network. *Molecular systems biology* **11**:794.

Guo Y, Li M, Pu X, Li G, Guang X, Xiong W, Li J. 2010. PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment. *BMC research notes* **3**:145.

Guo Y, Yu L, Wen Z, Li M. 2008. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic acids research* **36**:3025–30.

Guruharsha KG, Rual J-F, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee DY, Cenaj O, McKillip E, Shah S, Stapleton M, Wan KH, Yu C, Parsa B, Carlson JW, Chen X, Kapadia B, VijayRaghavan K, Gygi SP, Celniker SE, Obar RA, Artavanis-Tsakonas S. 2011. A Protein Complex Network of Drosophila melanogaster. *Cell* **147**:690–703.

Hart GT, Lee I, Marcotte ER. 2007. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC bioinformatics* **8**:236.

Havugimana P, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S, Babu M, Craig S a., Hu P, Wan C, Vlasblom J, Dar V-NUN, Bezginov A, Clark GW, Wu GC, Wodak SJ, Tillier ERM, Paccanaro A, Marcotte

EM, Emili A. 2012. A census of human soluble protein complexes. *Cell* **150**:1068–1081.

Ho Y, Gruhler a, Heilbut a, Bader G, Moore L, Adams S, 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, Sassi H, Nielsen P a, Rasmussen K, Andersen J, Podtelejnikov a, Nielsen E, Crawford J, Poulsen V, Sorensen B, Matthiesen J, Hendrickson R, Gleeson F, Pawson T, Moran M, Durocher D, Mann M, Hogue C, Figeys D, Tyers M. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**:180–183.

Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, Widmayer P, Gruissem W, Zimmermann P. 2008. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Advances in bioinformatics* **2008**:420747.

Hu Z, Hung J-H, Wang Y, Chang Y-C, Huang C-L, Huyck M, DeLisi C. 2009. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic acids research* **37**:W115–21.

Huang S-Y. 2014. Search strategies and evaluation in protein-protein docking: principles, advances and challenges. *Drug discovery today* **19**:1081–96.

Hubner NC, Mann M. 2011. Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC). *Methods* **53**:453–459.

Hue M, Riffle M, Vert J-P, Noble WS. 2010. Large-scale prediction of protein-protein interactions from structures. *BMC bioinformatics* **11**:144.

Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S, Zarraga G, Colby G, Baltier K, Dong R, Guarani V, Vaites LP, Ordureau A, Rad R, Erickson BK, Wühr M, Chick J, Zhai B, Kolippakkam D, Mintseris J, Obar RA, Harris T, Artavanis-Tsakonas S, Sowa ME, De Camilli P, Paulo JA, Harper JW, Gygi SP. 2015. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**:425–440.

Huynen M. 2000. Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. *Genome Research* **10**:1204–1210.

Jäger S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, Shales M, Mercenne G, Pache L, Li K, Hernandez H, Jang GM, Roth SL, Akiva E, Marlett J, Stephens M, D'Orso I, Fernandes J, Fahey M, Mahon C, O'Donoghue AJ, Todorovic A, Morris JH, Maltby D a., Alber T, Cagney G, Bushman FD, Young J a., Chanda SK, Sundquist WI, Kortemme T, Hernandez RD, Craik CS, Burlingame A,

Sali A, Frankel AD, Krogan NJ. 2011. Global landscape of HIV–human protein complexes. *Nature* **481**:5–10.

Janin J, Séraphin B. 2003. Genome-wide studies of protein–protein interaction. *Current Opinion in Structural Biology* **13**:383–388.

Jansen R, Greenbaum D, Gerstein M. 2002. Relating Whole-Genome Expression Data with Protein-Protein Interactions. *Genome Research* **12**:37–46.

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science (New York, N.Y.)* **302**:449–53.

Jiang C, Coenen F, Zito M. 2012. A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review* **28**:75–105.

Juan D, Pazos F, Valencia A. 2008. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences of the United States of America* **105**:934–9.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **40**:D109–14.

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H. 2012. The IntAct molecular interaction database in 2012. *Nucleic acids research* **40**:D841–6.

Kotlyar M, Pastrello C, Pivetta F, Lo Sardo A, Cumbaa C, Li H, Naranian T, Niu Y, Ding Z, Vafaee F, Broackes-Carter F, Petschnigg J, Mills GB, Jurisicova A, Stagljar I, Maestro R, Jurisica I. 2015. In silico prediction of physical protein interactions and characterization of interactome orphans. *Nature methods* **12**:79–84.

Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM. 2006. The RCSB PDB information portal for structural genomics. *Nucleic acids research* **34**:D302–5.

Kozakov D, Beglov D, Bohnuud T, Mottarella SE, Xia B, Hall DR, Vajda S. 2013. How good is automated protein docking? *Proteins* **81**:2159–66.

Kristensen AR, Gsponer J, Foster LJ. 2012. A high-throughput approach for measuring temporal changes in the interactome. *Nature methods* **9**:907–9.

Kristensen DM, Wolf YI, Mushegian AR, Koonin E V. 2011. Computational methods for Gene Orthology inference. *Briefings in Bioinformatics* **12**:379–391.

Krumsiek J, Friedel CC, Zimmer R. 2008. ProCope--protein complex prediction and evaluation. *Bioinformatics* **24**:2115–2116.

Krupp M, Marquardt JU, Sahin U, Galle PR, Castle J, Teufel A. 2012. RNA-Seq Atlas--a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics (Oxford, England)* **28**:1184–5.

Kühner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, Castaño-Diez D, Chen W-H, Devos D, Güell M, Norambuena T, Racke I, Rybin V, Schmidt A, Yus E, Aebersold R, Herrmann R, Böttcher B, Frangakis AS, Russell RB, Serrano L, Bork P, Gavin A-C. 2009. Proteome organization in a genome-reduced bacterium. *Science (New York, N.Y.)* **326**:1235–40.

Lambert J-P, Ivosev G, Couzens AL, Larsen B, Taipale M, Lin Z-Y, Zhong Q, Lindquist S, Vidal M, Aebersold R, Pawson T, Bonner R, Tate S, Gingras A-C. 2013. Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nature methods* **10**:1239–45.

Laukens K, Naulaerts S, Berghe W Vanden. 2014. Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis. *Proteomics* **15**:981–96.

Lavallée-Adam M, Cloutier P, Coulombe B, Blanchette M. 2011. Modeling contaminants in AP-MS/MS experiments. *Journal of Proteome Research* **10**:886–895.

Van Leene J, Hollunder J, Eeckhout D, Persiau G, Van De Slijke E, Stals H, Van Isterdael G, Verkest A, Neirynck S, Buffel Y, De Bodt S, Maere S, Laukens K, Pharazyn A, Ferreira PCG, Eloy N, Renne C, Meyer C, Faure J-D, Steinbrenner J, Beynon J, Larkin JC, Van de Peer Y, Hilson P, Kuiper M, De Veylder L, Van Onckelen H, Inzé D, Witters E, De Jaeger G. 2010. Targeted interactomics reveals a complex core cell cycle machinery in Arabidopsis thaliana. *Molecular systems biology* **6**:397.

Lensink MF, Wodak SJ. 2013. Docking, scoring, and affinity prediction in CAPRI. *Proteins* **81**:2082–95.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13**:2178–89.

Lievens S, Lemmens I, Tavernier J. 2009. Mammalian two-hybrids come of age. *Trends in biochemical sciences* **34**:579–88.

Lin X, Chen X. 2013. Heterogeneous data integration by tree-augmented naïve Bayes for protein-protein interactions prediction. *Proteomics* **13**:261–8.

Malovannaya A, Lanz RB, Jung SY, Bulynko Y, Le NT, Chan DW, Ding C, Shi Y, Yucer N, Krenciute G, Kim BJ, Li C, Chen R, Li W, Wang Y, O'Malley BW, Qin J. 2011. Analysis of the human endogenous coregulator complexome. *Cell* **145**:787–799.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**:83–6.

Martin S, Roe D, Faulon J-L. 2005. Predicting protein-protein interactions using signature products. *Bioinformatics (Oxford, England)* **21**:218–26.

Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M. 2001. Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or “Interologs.” *Genome Research* **11**:2120–2126.

Mellacheruvu D, Wright Z, Couzens AL, Lambert J-P, St-Denis NA, Li T, Miteva Y V, Hauri S, Sardiu ME, Low TY, Halim VA, Bagshaw RD, Hubner NC, Al-Hakim A, Bouchard A, Faubert D, Fermin D, Dunham WH, Goudreault M, Lin Z-Y, Badillo BG, Pawson T, Durocher D, Coulombe B, Aebersold R, Superti-Furga G, Colinge J, Heck AJR, Choi H, Gstaiger M, Mohammed S, Cristea IM, Bennett KL, Washburn MP, Raught B, Ewing RM, Gingras A-C, Nesvizhskii AI. 2013. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nature methods* **10**:730–6.

Melquiond ASJ, Karaca E, Kastritis PL, Bonvin AMJJ. 2012. Next challenges in protein-protein docking: from proteome to interactome and beyond. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2**:642–651.

Meysman P, Sonego P, Bianco L, Fu Q, Ledezma-Tejeida D, Gama-Castro S, Liebens V, Michiels J, Laukens K, Marchal K, Collado-Vides J, Engelen K, Qiang F. 2014. COLOMBOS v2.0: An ever expanding collection of bacterial expression compendia. *Nucleic Acids Res.* **42**:D649–53.

Michaut M, Kerrien S, Montecchi-Palazzi L, Chauvat F, Cassier-Chauvat C, Aude JC, Legrain P, Hermjakob H. 2008. InteroPORC: Automated inference of highly conserved protein interaction networks. *Bioinformatics* **24**:1625–1631.

Milenković T, Przulj N. 2008. Uncovering biological network function via graphlet degree signatures. *Cancer informatics* **6**:257–73.

Moreira IS, Fernandes PA, Ramos MJ. 2010. Protein-protein docking dealing with the

unknown. *Journal of computational chemistry* **31**:317–42.

Morell M, Ventura S, Avilés FX. 2009. Protein complementation assays: approaches for the in vivo analysis of protein interactions. *FEBS letters* **583**:1684–91.

Najafabadi HS, Salavati R. 2008. Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome biology* **9**:R87.

Naulaerts S, Meysman P, Bittremieux W, Vu TN, Vanden Berghe W, Goethals B, Laukens K. 2015. A primer to frequent itemset mining for bioinformatics. *Briefings in bioinformatics* **16**:216–31.

Nesvizhskii AI. 2012. Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. *Proteomics* **12**:1639–1655.

Newman M, Girvan M. 2004. Finding and evaluating community structure in networks. *Physical Review E* **69**:026113.

Nogales E, Wolf SG, Downing KH. 1998. Structure of the alpha beta tubulin dimer by electron crystallography. *Nature* **391**:199–203.

van Nuland R, Smits AH, Pallaki P, Jansen PWTC, Vermeulen M, Timmers HTM. 2013. Quantitative dissection and stoichiometry determination of the human SET1/MLL histone methyltransferase complexes. *Molecular and cellular biology* **33**:2067–77.

Oeffinger M. 2012. Two steps forward--one step back: advances in affinity purification mass spectrometry of macromolecular complexes. *Proteomics* **12**:1591–608.

Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing K a, Ahn NG. 2005. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Molecular & cellular proteomics : MCP* **4**:1487–1502.

Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic acids research* **38**:D196–203.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. Use of contiguity on the chromosome to predict functional coupling. *In silico biology* **1**:93–108.

Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes HW, Ruepp A, Frishman D. 2005. The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**:832–834.

Pavlopoulos G, Wegener A, Schneider R. 2008. A survey of visualization tools for

biological network analysis. *BioData Mining* **1**:12.

Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering Design and Selection* **14**:609–614.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences* **96**:4285–4288.

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**:3551–67.

Pichler P, Mazanek M, Dusberger F, Weilnböck L, Huber CG, Stingl C, Luider TM, Straube WL, Köcher T, Mechtler K. 2012. SIMPATIQCO: a server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on Orbitrap instruments. *Journal of proteome research* **11**:5540–7.

Pierce BG, Wiehe K, Hwang H, Kim B-H, Vreven T, Weng Z. 2014. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics (Oxford, England)* **30**:1771–3.

Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, Gebbia M, Greenblatt J, Jessulat M, Krogan N, Luo X, Golshani A. 2006. PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC bioinformatics* **7**:365.

Płociński P, Laubitz D, Cysewski D, Stoduś K, Kowalska K, Dziembowski A. 2014. Identification of protein partners in mycobacteria using a single-step affinity purification method. *PLoS ONE* **9**.

Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T, Rattei T, Creevey C, Kuhn M, Jensen LJ, von Mering C, Bork P. 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic acids research* **42**:D231–9.

Przulj N, Corneil DG, Jurisica I. 2004. Modeling interactome: scale-free or geometric? *Bioinformatics (Oxford, England)* **20**:3508–15.

Pu S, Vlasblom J, Turinsky A, Marcon E, Phanse S, Trimble SS, Olsen J, Greenblatt J, Emili A, Wodak SJ. 2015. Extracting high confidence protein interactions from affinity purification data: At the crossroads. *Journal of Proteomics*.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A,

Finn RD. 2012. The Pfam protein families database. *Nucleic acids research* **40**:D290–301.

Razick S, Magklaras G, Donaldson IM. 2008. iRefIndex: a consolidated protein interaction database with provenance. *BMC bioinformatics* **9**:405.

Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM. 2005. Probabilistic model of the human protein-protein interaction network. *Nature biotechnology* **23**:951–9.

Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE, Braun P, Brehme M, Broly MP, Carvunis A-R, Convery-Zupan D, Corominas R, Coulombe-Huntington J, Dann E, Dreze M, Dricot A, Fan C, Franzosa E, Gebreab F, Gutierrez BJ, Hardy MF, Jin M, Kang S, Kiros R, Lin GN, Luck K, MacWilliams A, Menche J, Murray RR, Palagi A, Poulin MM, Rambout X, Rasla J, Reichert P, Romero V, Ruyssinck E, Sahalie JM, Scholz A, Shah AA, Sharma A, Shen Y, Spirohn K, Tam S, Tejeda AO, Trigg SA, Twizere J-C, Vega K, Walsh J, Cusick ME, Xia Y, Barabási A-L, Iakoucheva LM, Aloy P, De Las Rivas J, Tavernier J, Calderwood MA, Hill DE, Hao T, Roth FP, Vidal M. 2014. A Proteome-Scale Map of the Human Interactome Network. *Cell* **159**:1212–1226.

Roux KJ, Kim DI, Raida M, Burke B. 2012. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *The Journal of cell biology* **196**:801–10.

Rudnick PA, Clauser KR, Kilpatrick LE, Tchekhovskoi D V, Neta P, Blonder N, Billheimer DD, Blackman RK, Bunk DM, Cardasis HL, Ham A-JL, Jaffe JD, Kinsinger CR, Mesri M, Neubert TA, Schilling B, Tabb DL, Tegeler TJ, Vega-Montoto L, Variyath AM, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Carr SA, Fisher SJ, Gibson BW, Paulovich AG, Regnier FE, Rodriguez H, Spiegelman C, Tempst P, Liebler DC, Stein SE. 2010. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Molecular & cellular proteomics : MCP* **9**:225–41.

Russell RB, Alber F, Aloy P, Davis FP, Korkin D, Pichaud M, Topf M, Sali A. 2004. A structural perspective on protein-protein interactions. *Current opinion in structural biology* **14**:313–24.

Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M, Kurbatova N, Malone J, Mani R, Mupo A, Pedro Pereira R, Pilicheva E, Rung J, Sharma A, Tang YA, Ternent T, Tikhonov A, Welter D, Williams E, Brazma A, Parkinson H, Sarkans U. 2013. ArrayExpress update--trends

in database growth and links to data analysis tools. *Nucleic acids research* **41**:D987–90.

Sardiu ME, Cai Y, Jin J, Swanson SK, Conaway RC, Conaway JW, Florens L, Washburn MP. 2008. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proceedings of the National Academy of Sciences of the United States of America* **105**:1454–1459.

Selbach M, Mann M. 2006. Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nature methods* **3**:981–3.

Singh SS, Typas A, Hengge R, Grainger DC. 2011. Escherichia coli σ^{70} senses sequence and conformation of the promoter spacer region. *Nucleic acids research* **39**:5109–18.

Skarra D V., Goudreault M, Choi H, Mullin M, Nesvizhskii AI, Gingras AC, Honkanen RE. 2011. Label-free quantitative proteomics and SAINT analysis enable interactome mapping for the human Ser/Thr protein phosphatase 5. *Proteomics* **11**:1508–1516.

Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Rattei T, Frishman D, Ruepp A. 2009. The Negatome database: A reference set of non-interacting protein pairs. *Nucleic Acids Research* **38**:540–544.

Smith GR, Sternberg MJE. 2002. Prediction of protein–protein interactions by docking methods. *Current Opinion in Structural Biology* **12**:28–35.

Smits AH, Jansen PWTC, Poser I, Hyman A a., Vermeulen M. 2013. Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Research* **41**:1–8.

Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)* **27**:431–2.

Sowa ME, Bennett EJ, Gygi SP, Harper JW. 2009. Defining the Human Deubiquitinating Enzyme Interaction Landscape. *Cell* **138**:389–403.

Sprinzak E, Margalit H. 2001. Correlated sequence-signatures as markers of protein-protein interaction. *Journal of molecular biology* **311**:681–92.

Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O. 2008. OpenMS - an open-source software framework for mass spectrometry. *BMC bioinformatics* **9**:163.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa

M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **101**:6062–7.

Suderman M, Hallett M. 2007. Tools for visually exploring biological networks. *Bioinformatics (Oxford, England)* **23**:2651–9.

Suter B, Kittanakom S, Stagljar I. 2008. Two-hybrid technologies in proteomics research. *Current opinion in biotechnology* **19**:316–23.

Tamames J, Casari G, Ouzounis C, Valencia A. 1997. Conserved Clusters of Functionally Related Genes in Two Bacterial Genomes. *Journal of Molecular Evolution* **44**:66–73.

Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW. 2008. An in vivo map of the yeast protein interactome. *Science (New York, N.Y.)* **320**:1465–70.

Teo G, Liu G, Zhang J, Nesvizhskii AI, Gingras AC, Choi H. 2014. SAINTexpress: Improvements and additional features in Significance Analysis of INTeractome software. *Journal of Proteomics* **100**:37–43.

Terrak M, Kerff F, Langsetmo K, Tao T, Dominguez R. 2004. Structural basis of protein phosphatase 1 regulation. *Nature* **429**:780–4.

Torchala M, Moal IH, Chaleil RAG, Fernandez-Recio J, Bates PA. 2013. SwarmDock: a server for flexible protein-protein docking. *Bioinformatics (Oxford, England)* **29**:807–9.

Tsou C-C, Avtonomov D, Larsen B, Tucholska M, Choi H, Gingras A-C, Nesvizhskii AI. 2015. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods* **12**:258–264.

Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Björling L, Ponten F. 2010. Towards a knowledge-based Human Protein Atlas. *Nature biotechnology* **28**:1248–50.

Varjosalo M, Sacco R, Stukalov A, van Drogen A, Planyavsky M, Hauri S, Aebersold R, Bennett KL, Colinge J, Gstaiger M, Superti-Furga G. 2013. Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nature methods* **10**:307–14.

Vaudel M, Burkhart JM, Zahedi RP, Oveland E, Berven FS, Sickmann A, Martens L, Barsnes H. 2015. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology* **33**:22–24.

Walzer M, Pernas LE, Nasso S, Bittremieux W, Nahnsen S, Kelchtermans P, Pichler P, van den Toorn HWP, Staes A, Vandenbussche J, Mazanek M, Taus T, Scheltema RA, Kelstrup CD, Gatto L, van Breukelen B, Aiche S, Valkenborg D, Laukens K, Lilley KS, Olsen J V, Heck AJR, Mechtler K, Aebersold R, Gevaert K, Vizcaíno JA, Hermjakob H, Kohlbacher O, Martens L. 2014. qcML: an exchange format for quality control metrics from mass spectrometry experiments. *Molecular & cellular proteomics : MCP* **13**:1905–13.

Walzthoeni T, Leitner A, Stengel F, Aebersold R. 2013. Mass spectrometry supported determination of protein complex structure. *Current opinion in structural biology* **23**:252–60.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**:470–6.

Wass MN, Fuentes G, Pons C, Pazos F, Valencia A. 2011. Towards the prediction of protein interaction partners using physical docking. *Molecular systems biology* **7**:469.

Wingender E, Dietze P, Karas H, Knüppel R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic acids research* **24**:238–41.

Wuchty S. 2006. Topology and weights in a protein domain interaction network--a novel way to predict protein interactions. *BMC genomics* **7**:122.

Xia J-F, Han K, Huang D-S. 2010. Sequence-Based Prediction of Protein-Protein Interactions by Means of Rotation Forest and Autocorrelation Descriptor. *Protein & Peptide Letters* **17**:137–145.

Xie Z, Kwoh CK, Li XL, Wu M. 2011. Construction of co-complex score matrix for protein complex prediction from AP-MS data. *Bioinformatics* **27**.

Xue Y, Ren J, Gao X, Jin C, Wen L, Yao X. 2008. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Molecular & cellular proteomics : MCP* **7**:1598–608.

Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H. 2004. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proceedings of the National Academy of Sciences of the United States of America* **101**:5934–9.

Yu C-Y, Chou L-C, Chang DT-H. 2010. Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC bioinformatics*

11:167.

Yu H, Paccanaro A, Trifonov V, Gerstein M. 2006. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics (Oxford, England)* **22**:823–9.

Yu J, Pacifico S, Liu G, Finley RL. 2008. DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC genomics* **9**:461.

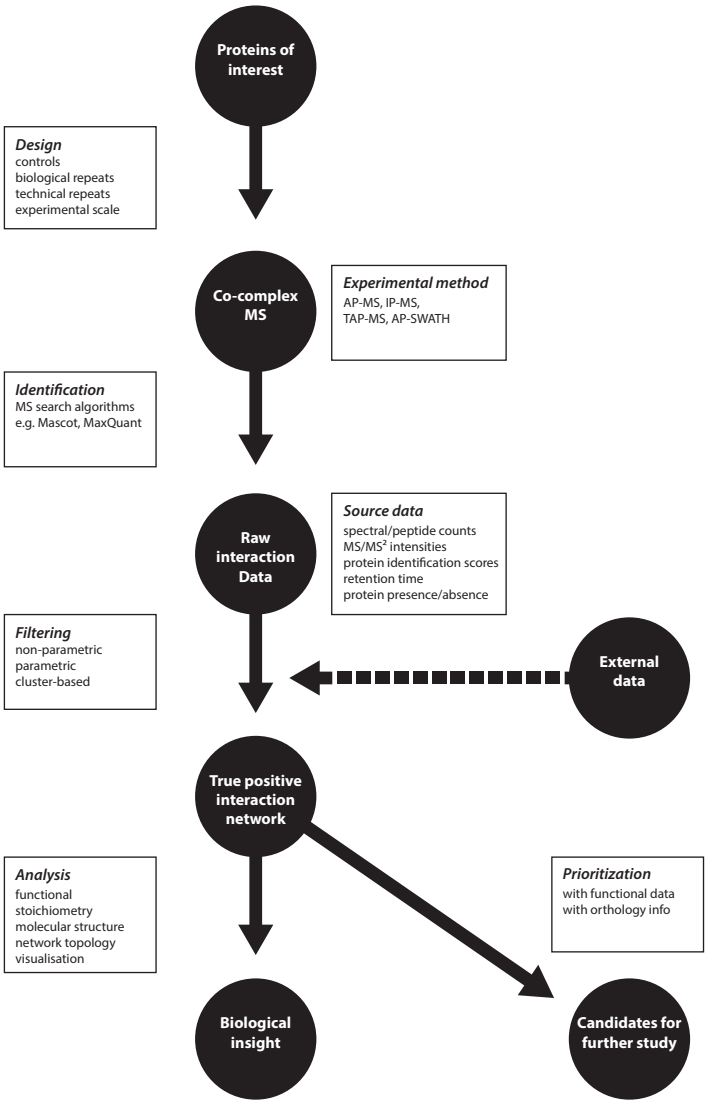
Zhang SJ, Hagenbuchner M, Scarselli F, Tsoi AC. 2010. Supervised Encoding of Graph-of-Graphs for Classification and Regression Problems. *Lecture Notes in Computer Science* **6203**:449–461.

Zhang B, Park BH, Karpinets T, Samatova NF. 2008. From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics* **24**:979–986.

Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B. 2012. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**:556–60.

Figure legend

Figure 1. Overview of the steps in co-complex mass spectrometry data analysis. MS: mass spectrometry, AP: affinity purification, IP: immunoprecipitation, TAP: tandem affinity purification, SWATH: sequential window acquisition of all theoretical spectra.



Tables

Table 1: Overview of commonly used methods for filtering false positive from AP-MS

Name	Type	Model	Input data	Reference
PE score	Parametric	Bayesian model	Presence	(Collins et al., 2007)
SAINT	Parametric	Bayesian model	Spectral counts	(Skarra et al., 2011)
SAINT-MS1	Parametric	Bayesian model	Intensities	(Choi et al., 2012)
SAINT express	Parametric	Bayesian model	Spectral counts	(Teo et al., 2014)
Co-complex score	Parametric	Bayesian model	Presence	(Xie et al., 2011)
HGScore	Parametric	Hypergeometric	Spectral counts	(Guruharsha et al., 2011)
Decontaminator	Parametric	Log ratio	Mascot scores	(Lavallée-Adam et al., 2011)
PP-NSAF	Non-parametric	Bayesian probability	Spectral counts	(Sardiu et al., 2008)
Socio-affinity index	Non-parametric	Log-odds	Presence	(Gavin et al., 2006)
CompPASS Z-Score	Non-parametric	Normal distribution	Spectral counts	(Sowa et al., 2009)
CompPASS D-Score	Non-parametric	Normal distribution	Spectral counts	(Sowa et al., 2009)
E-filter	Non-parametric	Box plots	Spectral counts	(Malovannaya et al., 2011)
SFINX	Non-parametric	Binomial distribution	Peptide counts	(Titeca et al. submitted)
Dice Coefficient	Cluster-based	Dice index	Presence	(Zhang et al., 2010)
MCL	Cluster-based	Clique finding	Presence	(Enright et al., 2002)
Nested clustering	Cluster-based	Mixed model	Spectral counts	(Choi et al., 2010)

Table 2: Overview of some of the major public protein-protein interaction databases with number of interactions (as of 23 February, 2015).

Name	**Organism(s)**	**Source**	**# Interactions**	**Website**
IntAct	Many	Experimental	477 526	http://www.ebi.ac.uk/intact/
BioGRID	Many	Experimental	364 964	http://thebiogrid.org
MINT	Many	Experimental	241 458	http://mint.bio.uniroma2.it
DIP	10	Experimental	78 744	http://dip.doe-mbi.ucla.edu
HPRD	Human	Experimental	41 327	http://www.hprd.org
MIPS	Mammals	Experimental	1 814	http://mips.helmholtz-muenchen.de/proj/ppi/
STRING	Many	Predictions & experimental	>200 000 000	http://string-db.org
I2D	6	Predictions	900 529	http://ophid.utoronto.ca/ophidv2.204/
iRefIndex	Many	Combined	492 588	http://irefindex.org/
DroID	Drosophila	Combined	235 333	http://www.droidb.org
APID	Many	Combined	322 579	http://bioinfow.dep.usal.es/apid

Table 3: Overview of network visualization tools

Name	Platform	Analysis tools	Export options	Website
Cytoscape	Windows, Mac, Linux	Plugins	JPG, PDF, PNG, SVG and HTML	http://cytoscape.org
Biolayout 3D	Windows, Mac, Linux	MCL clustering	PNG, JPG and 3D.js	http://biolayout.org
Gephi	Windows, Mac, Linux	MCODE clustering	PDF, PNG, SVG	http://gephi.github.io
VisANT	Java standalone and online	Topology and annotation analysis	JPG, PNG, SVG	http://visant.bu.edu
Pajek	Windows, Wine	Topology analysis	EPS, SVG, JPG, BMP	http://pajek.imfm.si
GraphViz	Windows, Linux, Solaris, Mac	External tools	Many	http://www.graphviz.org
NAViGaTOR	Windows, Mac, Linux	Topology and annotation analysis	BMP, JPEG, PDF, SVG, TIFF, PNG	http://ophid.utoronto.ca/navigator/