BRIEFINGS IN BIOINFORMATICS. VOL 16. NO 4. 658–674
Advance Access published on 5 November 2014
doi:10.1093/bib/bbu038

Network inference from AP-MS data: computational challenges and solutions

Ben Teng, Can Zhao, Xiaoqing Liu and Zengyou He

Submitted: 10th July 2014; Received (in revised form): 30th September 2014

Abstract

Protein–protein interaction is of primary importance to understand protein functions. In recent years, the high-throughput AP-MS experiments have generated a large amount of bait–prey data, posing great challenges on the computational analysis of such data for inferring true interactions and protein complexes. To date, many research efforts have been devoted to developing novel computational methods to analyze these AP-MS data sets. In this article, we review and classify the key computational methods developed for the inference of protein–protein interactions and the detection of protein complexes from the AP-MS experiments. We hope that our review as well as the challenges highlighted in the article will provide valuable insights into driving future research for further advancing the state-of-the-art technologies in computational prediction, characterization and analysis of protein–protein interactions and protein complexes from the AP-MS data.

Keywords: *AP-MS data; protein–protein interactions; protein complexes; validation*

INTRODUCTION

Protein–protein interactions (PPIs) play an important role in the biological process and metabolic functions in the cell [1, 2]. Recently, high-throughput experimental techniques such as Affinity Purification/Mass Spectrometry (AP-MS) have generated large data sets of experimentally detected protein–protein interactions. In the AP-MS experiments, a tagged bait protein is used to capture the prey proteins in a purification, where the preys are candidate interacting partners of the bait. Figure 1 gives an illustration of the entire experimental process, and Table 1 lists some online AP-MS data resources that have been used in the methods reviewed in this article. The resulting AP-MS data can be either qualitative or quantitative. The quantitative protein abundance can be estimated from different types of information generated in MS experiments such as the peptide count, the spectral count and the sequence coverage [8–11]. From such AP-MS data, protein–protein interactions can be inferred, and then protein complexes are detected.

However, there are still several challenges to overcome for accurately detecting protein interactions and complexes from the AP-MS data. First, a single bait protein may be involved in more than one complex and therefore it can capture a set of prey proteins which actually never occur in the

Corresponding author: Zengyou He, School of Software, Dalian University of Technology, China. Tel.: +86-411-87571630; E-mail: zyhe@dlut.edu.cn

Ben Teng received the BS degree in software engineering from Dalian University of Technology, China, in 2013. He is currently working towards the MS degree in the School of Software at Dalian University of Technology. His research interests include bioinformatics and data mining.

Can Zhao received the BS degree in software engineering from Dalian University of Technology, China, in 2014. She is currently working towards the MS degree in the School of Software at Dalian University of Technology. Her research interests include bioinformatics and data mining.

Xiaoqing Liu received the BS degree in software engineering from Dalian University of Technology, China, in 2013. She is currently working towards the MS degree in the School of Software at Dalian University of Technology. Her research interests include bioinformatics and data mining.

Zengyou He received the BS, MS, and PhD degrees in computer science from Harbin Institute of Technology, China, in 2000, 2002, and 2006, respectively. He was a research associate in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology from February 2007 to February 2010. Since March 2010, he has been an associate professor in the School of Software at Dalian University of Technology. His research interests include data mining and computational mass spectrometry.

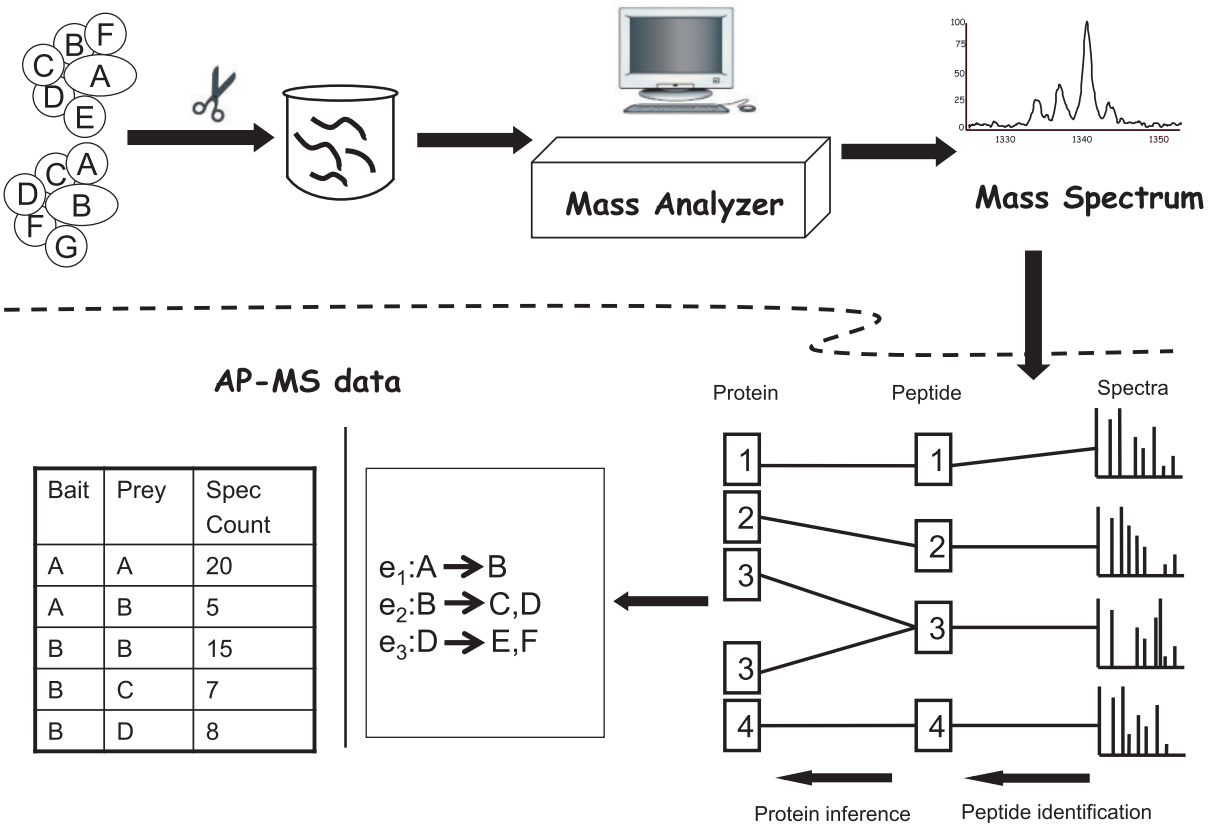


Figure 1: The entire workflow of an AP-MS experiment. Firstly, the bait proteins are used to capture the sets of prey proteins. Then, all proteins are digested into peptides, and these peptides are forwarded into a mass spectrometer. Thirdly, peptides are identified from the tandem mass spectra and thereafter proteins are inferred from identified peptides. Finally, we obtain a list of bait–prey pairs, which may contain many false-positive interactions.

Table 1: Some AP-MS data available online

Name	URL
Saccharomyces cerevisiae [3]	http://kroganlab.ucsf.edu/links.html
TIP49, DUB and CDC23 [4]	http://www.nature.com/nmeth/journal/v8/n1/abs/nmeth.1541.html
HEK293 and Jurkat [5]	http://www.nature.com/nature/journal/v481/n7381/abs/nature10719.html
CDC23 [6]	http://pubs.acs.org/doi/suppl/10.1021/pr2011185r
Human Deubiquitinating Enzyme [7]	http://www.sciencedirect.com/science/article/pii/S0092867409005030

same complex. Second, it is well known that the real purification data sets are noisy and contain many false-positive interactions [12]. Third, in the processes of peptide identification and protein inference, new errors may be introduced.

Thus, effective computational methods for network inference from AP-MS data must be developed. To illustrate the characteristics and uniqueness of the AP-MS data, here we discuss several related network inference problems from different application domains, which are most similar to the AP-MS data analysis problem reviewed in this article.

- Diffusion network inference [13]. There is an unknown network over which contagions propagate. The objective is to recover the true network from a collection of observed cascades, where each cascade is created by the diffusion of a particular contagion. This problem is very similar to the network inference problem from AP-MS data in that we can regard each 'purification' as a 'cascade',

where identified proteins in each purification correspond to the infected nodes in the cascade. However, the problem of network inference from AP-MS data is more difficult because proteins in each purification are unordered. In contrast, the infected nodes in each cascade are ordered by their infection time. In other words, the collections of cascades contain more information than the set of purifications, making it unfeasible to directly apply models and algorithms in diffusion network inference to solving the network inference problem from AP-MS data.

- Network inference from co-occurrences [14]. The problem is to infer the network structure from a set of observations, where each observation consists of a set of related entries/nodes that carry the information transmission together. The observations only reflect which subsets of nodes are involved, but do not indicate the order in which they handle the transmission. Despite of the seeming similarity between the co-occurrence data and purification data, there are at least two key differences. Firstly, the inference objective in [14] is to establish a directed graph where each observation corresponds to an ordered path through the network. This is clearly different from the target network desired in AP-MS data analysis. In addition, proteins in a purification are not equally important, i.e. the bait protein pulls down other prey proteins. If we model each purification as an observation in the co-occurrence data, such valuable information is lost.
- Network inference from perturbation data. Perturbation experiments in system biology generate large amounts of data. Numerous algorithms have been proposed to analyze these perturbation data sets in order to reveal the true underlying network structures (e.g. [15, 16]). In our opinion, the 'pull-down' AP-MS data can be regarded as a special kind of perturbation data. However, the AP-MS data are distinct from other types of perturbation data in several perspectives. Firstly, the noise rate of the AP-MS data is very high, where the true interaction partners of the bait are often <10% of all identified preys in a single purification [17]. Secondly, the perturbation data sets for other network inference problems (e.g. signal transduction network) usually contain some extra information that can facilitate the inference. Finally, the aim of analyzing the AP-MS data is to build an undirected PPI network, while the algorithms such as those methods for inferring signal transduction networks focus on establishing a directed network.

The above analysis shows that the problem of network inference from the AP-MS data is distinct from the existing network inference problems in different domains. As a result, there have been many research efforts on developing novel computational methods to analyze such AP-MS data. However, the problem of inferring true interactions and protein complexes from the high-throughput AP-MS data is far from being resolved.

Generally, there are four main computational issues in the analysis of AP-MS data:

- **Predicting co-complex interactions between two proteins.** Generally, PPIs can be divided into two major types: physical interactions and co-complex interactions. Co-complex interactions are those which take place at such distances that the average properties of the protein or its surface may be used in calculating interaction energies [18]. The co-complex interaction means that the interacting protein pair does not need to have a direct physical contact, but interacts in the formation of a complex. In other words, co-complex interactions indicate the relationships among the members of the protein complexes: two proteins in the same complex share one co-complex interaction. Because raw AP-MS data sets contain many false-positive interactions, it is obliged to reduce the number of false positives. Co-complex interactions provide the co-membership information in a complex such that predicting co-complex interactions can be used as the pre-processing step for some protein complex detection methods. Meanwhile, it is also feasible to use the predicted co-complex interactions as the input for further inferring physical interactions. Therefore, in this article, we put the problem of predicting co-complex interactions as a research issue alone.
- **Inferring physical interactions between two proteins.** Physical interactions are those in which the interacting molecules approach closely and thus bring into play local forces [18]. Unlike co-complex interactions, one physical interaction represents a direct biophysical interaction between two proteins, that is, they are linked by an edge in the protein–protein interaction network. Thus, to construct the real PPI network, it is vital to infer

the physical interactions. Generally, it is recognized that yeast two-hybrid (Y2H) method [19] detects direct binary interactions while the AP-MS experiment captures co-complex associations. Therefore, it is a challenging issue to infer physical interactions from the AP-MS data.

- **Detecting protein complexes.** Protein does not act individually, but works as a member of a module to carry out its functions. There are two types of cellular modules [20]: protein complexes and functional modules. A protein complex is a group of proteins that interact with each other at the same location and time, while proteins in a functional module do not necessarily interact with each other at the same time and location. As discussed in [21], here we also do not distinguish these two concepts and use the term of protein complex to refer to both of them in this article. Because protein complexes are the key molecular entities to perform many essential biological functions, detecting protein complexes is another important goal in the analysis of the AP-MS data. Complex detection methods can use two types of inputs: the raw AP-MS data and the PPI network obtained through pre-processing the raw AP-MS data, as shown in Figure 2.
- **Assessing the results.** Validating the inference results is as important as developing new inference methods. The validation techniques are essential to develop new algorithms and verify the inference results.

These four issues penetrate the whole process of analyzing the AP-MS data, as described in Figure 2.

Several recent reviews have focused on some computational issues and methods listed above in detail [22–26]. However, there are still no available reviews that cover all these subjects from an algorithmic perspective. Here, we try to bridge this gap by reviewing major computational methods for analyzing the AP-MS data in a breadth-first manner. Our main objective is to highlight the data analysis challenges and typical computational strategies for each problem with a special focus on the algorithmic nature of underlying problem. To this end, we will not cover all methods in each category. Instead, we discuss the basic principle first and then just list some typical methods for illustration. In the rest of this article, we will briefly describe the computational problems and concisely discuss recent developments in different categories.

PREDICTING CO-COMPLEX INTERACTIONS

Since proteins recovered in a given purification rarely correspond to a single complex, but to a mixture of multiple complexes, which leads to the fact that the AP-MS data cannot be converted into co-complex interactions in a straightforward manner. Therefore, computational methods should be used to remove false-positive interactions and measure the confidence that a pair of proteins belongs to the same complex. As shown in Figure 2, there are two popular models for fitting the AP-MS data: the spoke model and the matrix model [27]. In the spoke model, the bait proteins act as hub nodes and the prey proteins are connected with the baits. In the matrix model, all the proteins that occur in the same purification experiment, either bait–prey or prey–prey pairs, are recognized as connected with each other. Until now, many methods have been proposed to score the reliability of a co-complex interaction between a pair of proteins that occur in the same purification experiment under these two models. In Figure 3, we provide a categorization on co-complex prediction algorithms and list some representative methods as well. Now, we briefly describe some typical scoring methods for predicting co-complex interactions.

Before we present the algorithms, we define some notations that are used in this section. Let $E = \{e_1, e_2, ..., e_n\}$ be an AP-MS data set of n purifications. For a pair of proteins, i and j, let $n_{i\to j}$ be the number of times j is observed among preys in purifications performed with i as the bait and $m_{i,j}$ be the number of times i and j are observed as preys in purifications performed with a third protein as a bait. Some scoring schemes combine these terms into one number $o_{i,j} = n_{i\to j} + n_{j\to i} + m_{i,j}$.

The methods for predicting co-complex interactions without using quantitative information can be further classified into two categories according to their scoring strategies. The scoring methods in the first class treat bait–prey pairs and prey–prey pairs separately. The socio-affinity (SA) [12] scoring method is one representative scoring method that adopts this strategy, whose interaction score can be written as a sum of direct bait–prey components (S) and an indirect prey–prey component (M). Thus, for a potential interaction between proteins i and j, the SA score reads as:

$$SA_{i,j} = S_{i\to j} + S_{j\to i} + M_{i,j}. \quad (1)$$

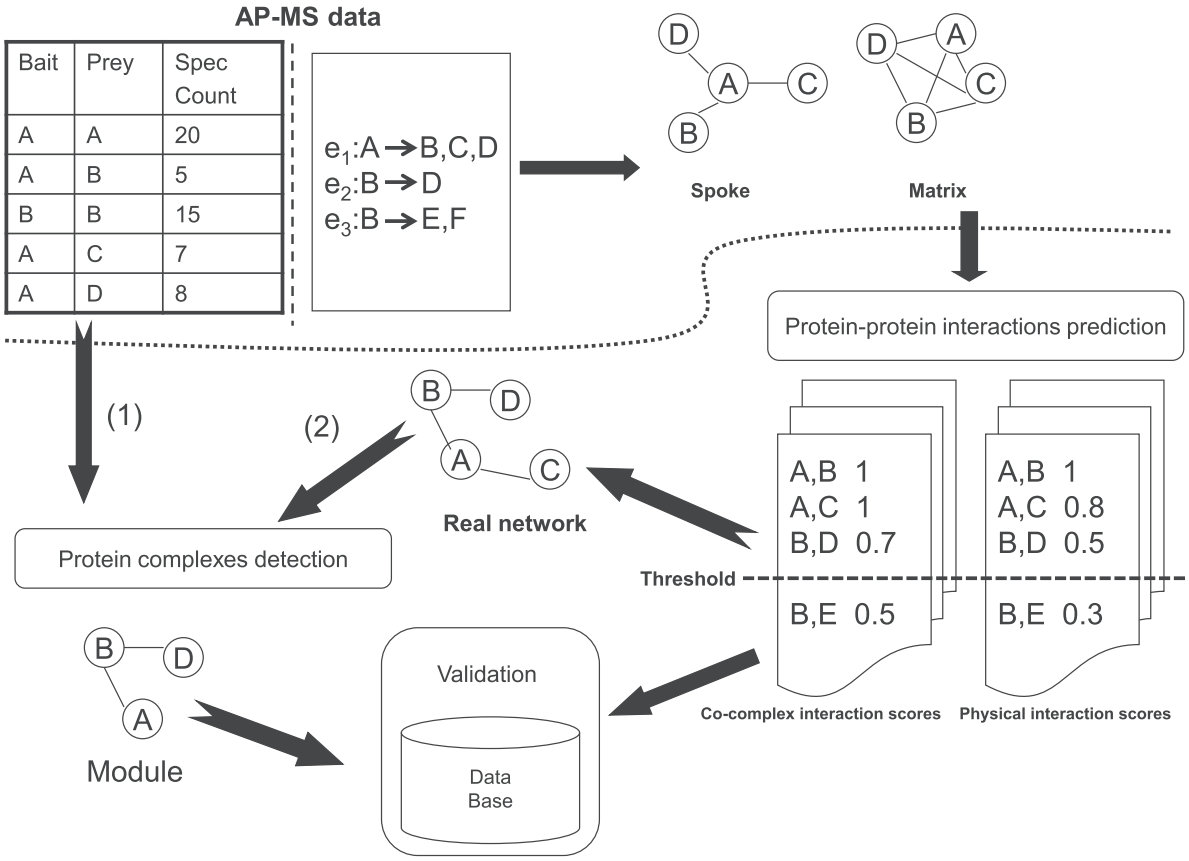


Figure 2: The brief process of dealing with AP-MS data. The spoke model and matrix model are widely used for interpreting the AP-MS data. Based on these two models, we can evaluate the interaction strength between two proteins with interaction prediction algorithms. With a given threshold, interactions with low scores are removed and interactions with high scores are used to construct the PPI network. Complex detection methods can use two types of inputs: (i) the raw AP-MS data and (ii) the PPI network obtained from the raw AP-MS data. Both the predicted interactions and detected complexes are assessed with some validation methods. There are two commonly used validation methods: the database-based approach and the reference-free approach.

Here, both $S_{i\to j}$ and $M_{i,j}$ use the log-ratios of actual co-occurrences relative to what would be expected based upon protein purification frequencies. That is,

$$S_{i\to j} = log\frac{n_{i\to j}}{f_i^{bait} nf_j^{prey} n_{bait=i}^{prey}}, \quad (2)$$

$$M_{i,j} = log\frac{m_{i,j}}{f_i^{prey} f_j^{prey} \sum_{allbaits} n_{prey}(n_{prey}-1)/2}, \quad (3)$$

where f_i^{bait} is the percentage of purifications where protein i is bait, f_j^{prey} is the percentage of all captured preys that are protein j and $n_{bait=i}^{prey}$ is the number of preys obtained with protein i as the bait. For the matrix model term (M), n_{prey} is the number of times that proteins i and j are observed in purifications with baits other than i or j.

Similar to the socio-affinity scores, the purification enrichment (PE) [3] scoring scheme also uses log-ratios of the actual co-occurrences relative to the expected ones based upon protein purification frequencies. The difference is that the PE method uses a more sophisticated statistical model to score each pair of proteins.

Another class of scoring methods without using quantitative abundance treat bait–prey pairs and prey–prey pairs equally in the process of interaction assessment. Here we use the IDBOS algorithm [28] and a correlation-based method [29] as examples for illustration.

IDBOS is a scoring method that exploits constrained randomized simulations. This method compares observed co-occurrences of protein pairs with those from random simulations where the latter are realized by shuffling prey proteins randomly. For each unique protein pair i and j, the total number of times they co-occur in the same purification, $o_{i,j}$, is first counted. Their average shuffled

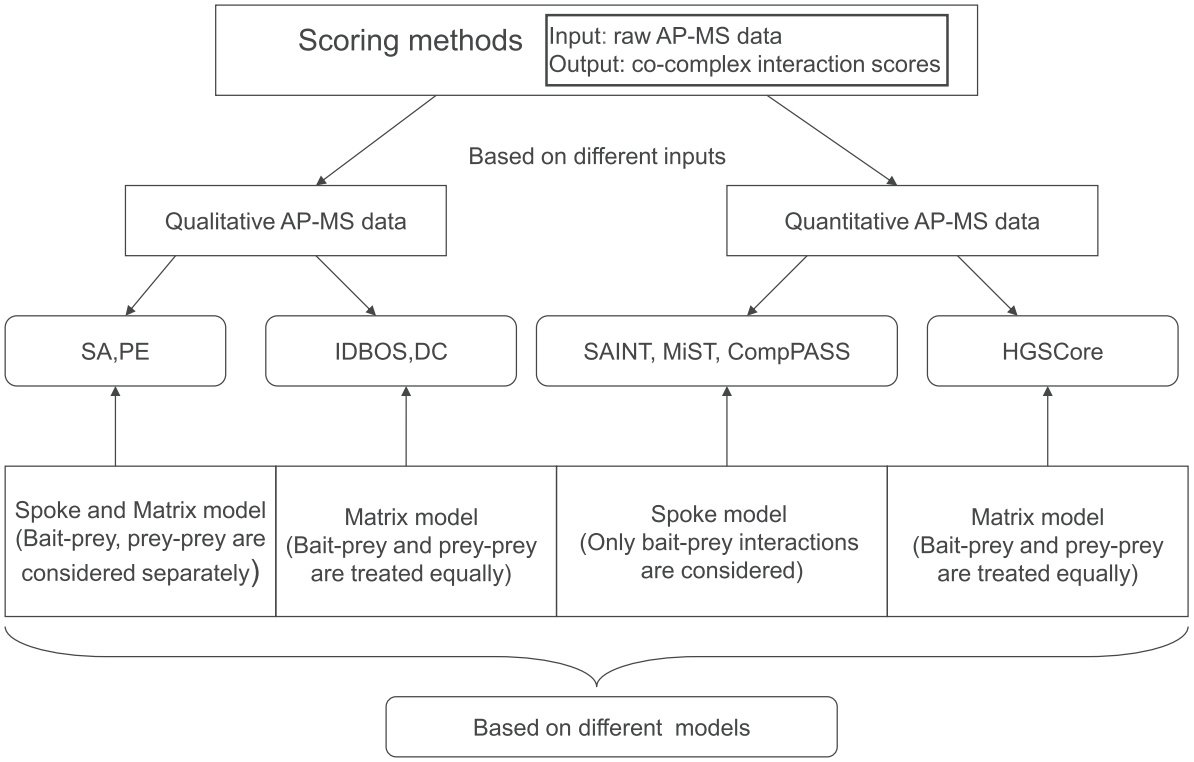


Figure 3: The categorization of co-complex interaction prediction methods. The co-complex interaction prediction methods can be divided into two categories based on whether using the quantitative information as the input. The methods in the first category that ignore the protein abundance information use two different strategies to score the candidate protein pairs. The first strategy evaluates the bait–prey and prey–prey interactions in a different manner and the methods in the second category treats all candidate protein pairs uniformly. On the other hand, the methods that incorporate quantitative protein abundance use either the spoke model or the matrix model for interaction prediction.

co-occurrence, $\hat{o}_{i,j}$, and associated standard deviation, $\delta_{i,j}$, are then obtained from the randomized purification sets. The *CS* score for this protein pair, *i* and *j*, is computed as:

$$CS_{i,j} = \frac{o_{i,j} - \hat{o}_{i,j}}{\delta_{i,j}}. \quad (4)$$

Figure 4 describes the workflow for computing the CS score.

In [29], the authors use the Dice coefficient (DC) to measure the correlation between two proteins. For two proteins, *i* and *j*, the Dice coefficient is defined as:

$$DC_{i,j} = \frac{2o_{i,j}}{2o_{i,j} + r + s}, \quad (5)$$

where *r* denotes the times that only protein *i* occurs, but protein *j* does not occur. Likewise, *s* denotes the times that only protein *j* occurs, but protein *i* does not occur.

Protein abundance information is helpful to distinguish true interactions from false positives. Therefore, many effective scoring methods have been developed to predict co-complex interactions from quantitative AP-MS data. These methods adopt either the spoke model or the matrix model in their scoring schemes. We describe some representative methods of each type as follows.

The Significance Analysis of Interactome (SAINT) method [4] uses spectral counts to facilitate the detection of bait–prey interactions. It employs a mixture of Poisson distributions to compute the posterior probability of a true interaction between a bait–prey pair.

The mass spectrometry interaction statistics (MiST) method [5] is designed for identifying AP-MS-derived host–pathogen interactions. MiST uses three measures in its scoring function: abundance (protein abundance measured by peak intensities), reproducibility (the invariability of abundance over replicated experiments) and specificity (the uniqueness of an observed interaction across all purifications). These three measures are combined into a single composite score, the

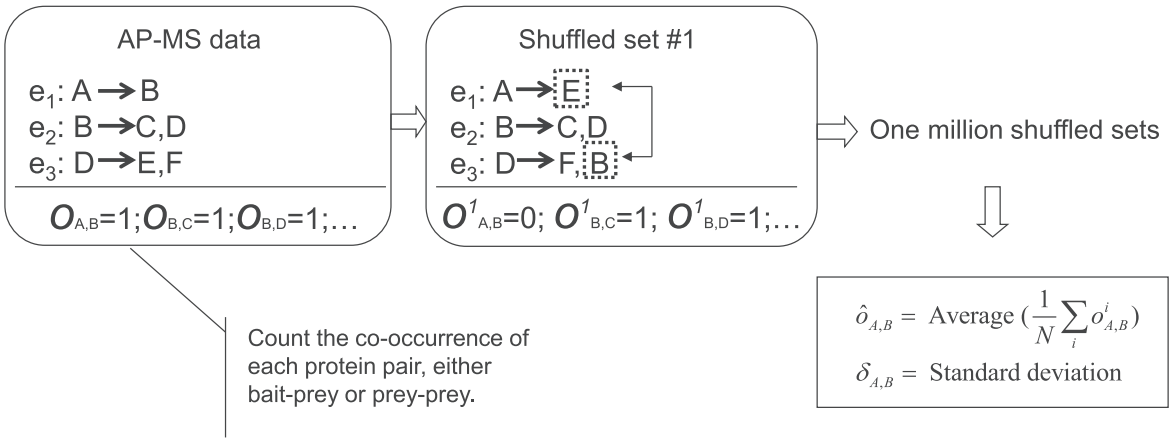


Figure 4: The workflow of the IDBOS method. This method first generates many random data sets by shuffling the original AP-MS data and then compares observed co-occurrences with those from random simulations. For example, in the original data, the number of co-occurrence of the protein pair *A* and *B*, $o_{A,B}$ is 1. While in the first shuffled set, the prey *B* and the prey *E* are exchanged. Then the number of co-occurrence of protein *A* and *B*, $o^1_{A,B}$ becomes 0. By calculating the average shuffled co-occurrence, $\hat{o}_{A,B}$, and associated standard deviation, $\delta_{A,B}$, the final CS score of protein pair *A* and *B* can be obtained.

MiST score, by maximizing the variance using the standard principal component analysis.

CompPASS [6] assigns scores to interactions based on the quantitative AP-MS matrix, where the columns are individual purifications, the rows are prey proteins, and each element is populated with the corresponding total spectral counts (TSC). This method employs two scoring metrics: the conventional Z score and the D score, as described in Figure 5. As the Z score equally weights unique preys regardless of their TSC, the D score is proposed to address this drawback. The D score incorporates the uniqueness, the prey abundance and the reproducibility of the interaction to assign a score to each bait–prey pair.

Different from the methods introduced above which only consider bait–prey pairs, the HGSCore method [30] is based on the matrix model that treats bait–prey and prey–prey interactions equally. This method uses the transformed and normalized TSC to measure the protein abundance, which is denoted by T_N. Because the determination of specificity of interaction between each pair of proteins hinges on the smaller one of the two T_N values, the HGSCore calculates the hypergeometric probability of observing an interaction between protein *i* and protein *j* as:

$$P_{hygeo;i,j} = P\left(\sum min(T_N) > k|w, q, T\right) = \sum_{x=k}^{min(w,q)} P_{hygeo}(x|w, q, T) = \sum_{x=k}^{min(w,q)} \frac{\binom{w}{x}\binom{T-w}{q-x}}{\binom{T}{q}}, \quad (6)$$

where

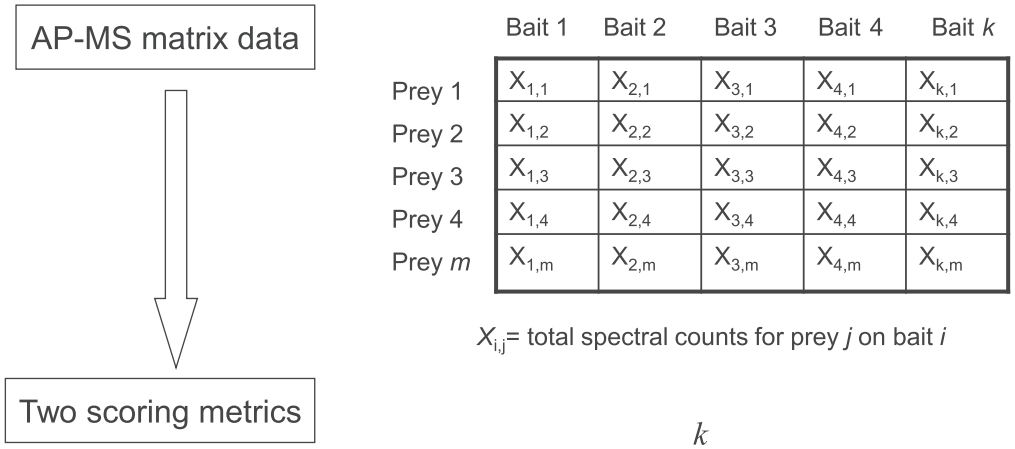
$k = \sum min(T_N)$ for purifications with $T_{N;i} > 0$ and $T_{N;j} > 0$,
$w = \sum min(T_N)$ for purifications with $T_{N;i} > 0$,
$q = \sum min(T_N)$ for purifications with $T_{N;j} > 0$,
$T = \sum min(T_N)$ for all purifications.

The final interaction score between protein *i* and protein *j* is calculated as:

$$HGSCore_{i,j} = -log(P_{hygeo;i,j}). \quad (7)$$

Tucker et al. [31] propose a sampling-based approach for scoring interactions, as described in Figure 6. This method first converts the spectral count into the probability that the observed interaction is true. Then, a specified number of random qualitative data sets are created by sampling bait–prey interactions according to their probabilities. Finally, each randomized data set is used as the input to an interaction prediction method of choice that operates on binary data, and the results from different binary matrices are aggregated to produce an ensemble score.

Although some recent breakthroughs have been achieved for predicting co-complex interactions, there are still certain drawbacks in the existing methods, as shown in Table 2. The methods such as SA, PE, IDBOS and DC, rely on a large-scale data set for statistical validity. These methods can effectively identify candidate interacting proteins from



Z score: $$Z_{i,j} = \frac{X_{i,j} - \bar{X}_j}{\delta_j} \quad \text{where} \quad \bar{X}_j = \frac{\sum_{i=1}^{k} X_{i,j}}{k}; j = 1,2,...m \qquad \text{(Equation 1)}$$

D score: $$D_{i,j}^{R} = \sqrt{(\frac{k}{\sum_{i=1} f_{i,j}})^{p_j} X_{i,j}} \quad \text{where} \quad \begin{cases} f_{i,j} = \begin{cases} 1; & X_{i,j} > 0 \\ X_{i,j} \end{cases} \\ p_j = \text{number of experiments that the prey } j \text{ is present.} \end{cases} \qquad \text{(Equation 2)}$$

Figure 5: The CompPASS method. In the quantitative AP-MS matrix, the columns are individual purifications, the rows are prey proteins, and each element is populated with the corresponding total spectral counts, as shown in the upper part. Based on the quantitative AP-MS matrix, CompPASS employs two scoring metrics: the conventional Z score and the D score. The Z score is calculated with Equations I, where $\bar{X}_j$ and δ_j denote the average and the standard deviation of total spectral counts for prey *j*, respectively. The D score is calculated according to Equation 2, where *k* is the total number of purifications in the AP-MS matrix.

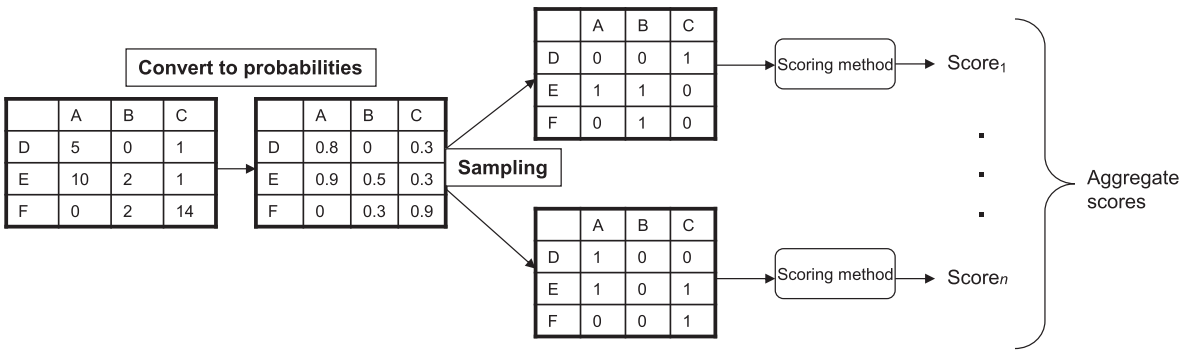


Figure 6: The process of the sampling method with quantitative information. The method first converts the spectral count into a probability that measures the interaction strength between a bait–prey pair. Then, it generates a specified number of binary data sets by sampling bait–prey interactions according to their probabilities. Thereafter, existing scoring methods that operate on binary data are employed to infer interactions from these binary matrices. Finally, the results from different binary data sets are aggregated to produce the final scores.

large-scale data sets but may not be appropriate with small-scale data sets or when baits do not share common preys. Additionally, they do not fully utilize the quantitative protein abundance. SAINT, MiST, CompPASS and HGSCore, are designed for the AP-MS data sets with quantitative information, and these methods can perform better when the quantitative information is sufficient and accurate.

Table 2: Summary of methods for predicting co-complex interactions from the AP-MS data.

Method	Data model	Input	Advantages	Disadvantages
SA [12]	Spoke and matrix model	Qualitative AP-MS data	Simple and no need for additional quantitative information.	Relying on large-scale data set; may not be appropriate with small-scale data sets or when baits do not share common preys.
PE [3]				
IDBOS [28]	Matrix model			
DC [29]				
SAINT [4]	Spoke model	Quantitative AP-MS data	More accurate and can be applied to small-scale data sets.	Only can be applied to AP-MS data with quantitative information.
CompPASS [7]				
MiST [5]				
HGScore [30]	Matrix model			
Tucker et al. [31]	Spoke and matrix model			

The CompPASS method can perform very well for data sets having a large number of unrelated baits but will filter out some true interactions with higher detection frequency when all baits belong to the same protein pathway [32]. The SAINT method over-penalizes true interactions that have high average total spectral counts but are not captured in all replicates [32]. On the other hand, the methods developed for special purpose such as MiST, can be applied to the general AP-MS data sets but their experimental performance needs to be further investigated. Therefore, the prediction of co-complex associations remains an open problem.

INFERRING PHYSICAL INTERACTIONS

The methods reviewed in the section of *Predicting co-complex interactions* mostly concentrate on two aspects: (i) separating the true co-complex interactions from experimental noises such as protein misidentification [29, 33, 34] and contaminants [22, 35, 36], and (ii) measuring the strength of a co-complex interaction between two proteins. However, the complex associated proteins may not interact with each other directly in the PPI networks. And the scoring methods for predicting co-complex interactions can assign high scores to the pairs of proteins occurring in the same complex, but probably they are not interacted directly with each other. In order to construct the real PPI networks, it is important to infer the physical interactions. Accurately separating direct interactions from indirect ones can also help to detect protein complexes from given purifications which correspond to a mixture of multiple complexes. To date, three computational methods have been proposed, indicating that physical PPIs can also be inferred from the AP-MS data.

The ISA algorithm [37] extends the SA scoring method to predict physical interactions from the AP-MS data. It adopts the null model used in the SA to derive the ISA score as:

$$ISA_{i,j} = -logPr(n_{i\to j}^{null} \geq n_{i\to j}) - logPr(n_{j\to i}^{null} \geq n_{j\to i}), \quad (8)$$

where $n_{i\to j}^{null}$ denotes the times that protein j is retrieved with protein i as the bait when the null hypothesis is true. The results on two high-throughput yeast data [12, 38] show that ISA is good at inferring physical interactions from AP-MS data.

The Sets2Networks algorithm [39] is designed for inferring networks from repeated observations of sets, which can be applied to discovering the direct PPIs from the AP-MS data as well. This method first generates an ensemble of networks consistent with the observed data. Then the presence probability of a given link in the underlying real network conditioned on the data is estimated from the occurrence frequency of the link throughout the ensemble. Specially, it calculates the mean adjacency matrix over the ensemble, each element of which corresponds to the probability of the edge being present in a uniformly random draw from the ensemble. This matrix quantifies the confidence of each direct physical PPI given solely the information on the connectivity of the underlying network derived from the experimental data.

Kim et al. [40] present a physical interaction prediction method with quantitative abundance as the input. This method models the unknown direct interaction network as a probabilistic graph under the following two assumptions:

- All direct interactions survive with the same probability p, and fail independently with probability $1 - p$.

- All direct interactions take place with the same frequency at the same time, irrespective of the presence of other interactions.

In order to measure the connectivity of all pairs of proteins in the direct PPI network G_{direct}, P_{direct} is used to denote the connectivity matrix of G_{direct}. For a pair of proteins, i and j, the corresponding $p_{direct}(i, j)$ in P_{direct} means the probability that there exists at least one path between i and j after each edge in G_{direct} is broken with probability p. Figure 7 shows an example of a direct interaction network and its connectivity matrix.

Under the above assumptions, the following formula is further derived:

$$A_{i,j} \propto \chi(i, j) * p_{direct}(i, j). \quad (9)$$

Here $A_{i,j}$ denotes the abundance of the prey protein j when protein i is selected as the bait and $\chi(i, j)$ denotes the number of pairs of proteins i and j that interact directly in the cell considered. This formula means that the amount of protein j that will be obtained when protein i is the bait is proportional to the probability that i and j remain connected after each edge in G_{direct} is broken with probability p. In the AP-MS experiments, $m_{direct}(i, j)$, which is an estimate of $p_{direct}(i, j)$, can be gained from $A_{i,j}$ through appropriate normalization. Therefore, the problem of inferring direct PPIs is formulated as finding the best P_{direct} such that for each pair of proteins i and j, $|m_{direct}(i, j) - p_{direct}(i, j)| < \epsilon$, where ϵ represents the error tolerance.

The study on the prediction of physical PPIs from the AP-MS data is still in its infant stage. Each of the existing three methods has its strengths and weaknesses. On one hand, both ISA and Sets2Networks can be applied to the AP-MS data with or without quantitative information while the method proposed in [40] needs protein abundances as the input. On the other hand, ISA is more efficient than the other two algorithms, as Sets2Networks needs to conduct a large-scale simulation and the algorithm in [40] has to solve a hard optimization problem.

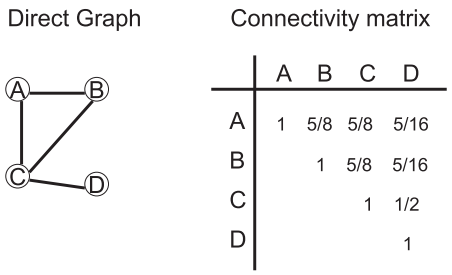


Figure 7: An example of a direct interaction network (left) and its connectivity matrix (right). In the calculation of connectivity matrix, each edge is assumed to survive with the probability of 1/2. The connectivity probability between two proteins can be estimated via sampling the probabilistic network. For example, the probability of connectivity between protein A and protein B is calculated as $1 - (1/2 * 1/2 + 1/2 * 1/2 * 1/2) = 5/8$.

DETECTING PROTEIN COMPLEXES

Protein complexes are key units to carry out metabolic functions in the cell. To date, some notable computational approaches have been proposed to identify protein complexes from AP-MS data. These methods can be characterized into two categories, as shown in the right part of Figure 8.

As shown in the left part of Figure 8, there are three steps for the complex detection methods in the first category. The first step is to assess the protein interaction affinities with methods such as SA and PE. The second step is to construct a protein–protein interaction network by applying a threshold or a cutoff value. Finally, the third step is to mine complexes on the constructed PPI network. Because the detection of protein complexes from PPI networks has been investigated for a long time, a variety of computational algorithms including MCODE [27], Markov clustering [41] and DPClus [42] can be employed directly in the third step. Recently, the hierarchical clustering algorithm is applied to detecting protein complexes from the PPI network that is constructed with a newly developed co-complex interaction prediction algorithm [43].

Converting the raw AP-MS data into a PPI network not only introduces errors but also loses useful information about the underlying multi-protein relationships that can be exploited to detect the internal organization of protein complexes [44]. Therefore, another alternative strategy is to detect complexes from the AP-MS data directly without constructing the protein–protein interaction networks. In this strategy, the AP-MS data set is modeled as a bipartite graph in which the two vertex sets are composed of bait proteins and prey proteins, respectively. The edges between the two vertex sets represent bait-prey connections. CACHET [45] and CODEC [46] are two examples in the category. Both of them focus on detecting high-quality protein-complex cores from the bipartite graph. To minimize the effects of false-positive

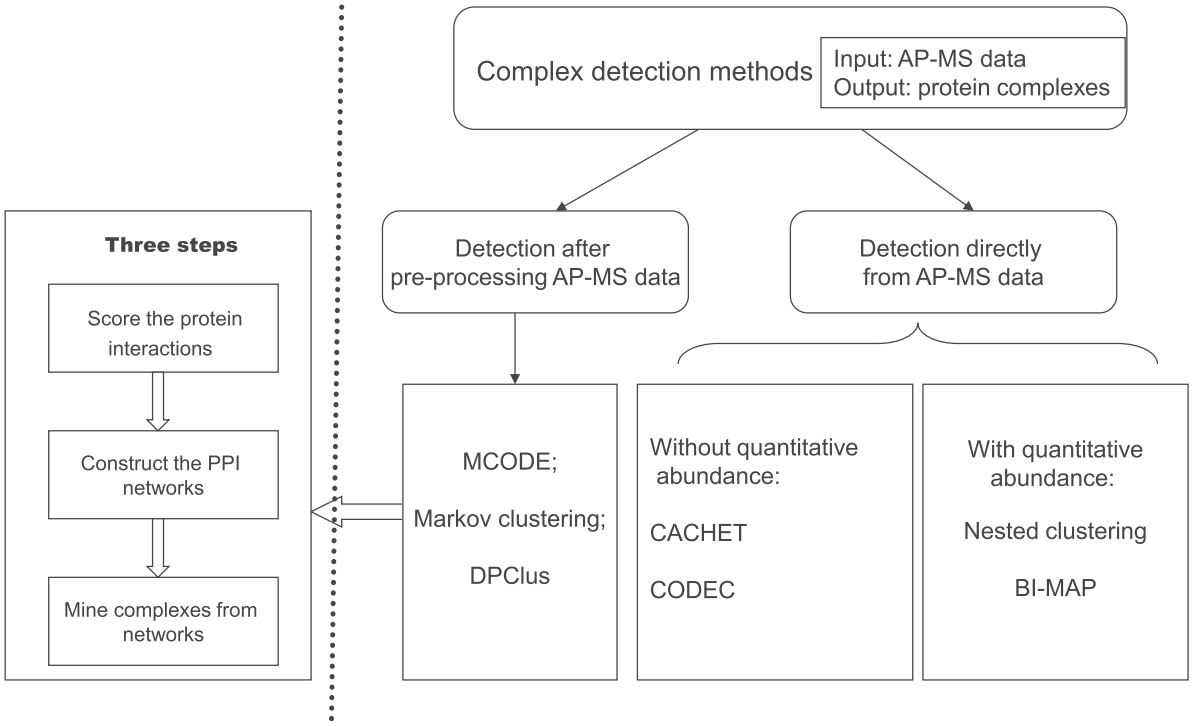


Figure 8: The classification of protein complex detection methods. According to the different inputs of the algorithms, the methods for detecting protein complexes can be characterized into two categories. In the first category, the methods mine the protein complexes from the PPI networks that are constructed with the interaction prediction methods listed in previous sections. In the second category, the methods detect complexes from AP-MS data directly by modeling the AP-MS data as a bipartite graph in which the two vertex sets are the set of baits and the set of preys, respectively.

interactions, the CACHET method uses the reliability scores such as SA, PE and DC scores to measure the bait–prey similarity while CODEC approximates the reliability of bait–prey links by the degree of the vertices. Only non-redundant, reliable bicliques obtained from the bipartite graph are regarded as protein-complex cores. Then both methods construct protein complexes by including attachment proteins into the cores. Figure 9 gives an illustration on the key idea of CODEC and CACHET. The difference between CODEC and CACHET is that CODEC detects dense bipartite subgraph by iteratively adding proteins into seed subgraphs while CACHET selects complex cores directly from the bipartite graph.

With the development of quantitative proteomics, more methods that use quantitative AP-MS matrix as the input have been proposed. In [47], the hierarchical clustering algorithm is used to cluster the rows and the columns of the AP-MS matrix independently so as to obtain a better organized matrix. On the other hand, the nested clustering algorithm [48] is a two-step sequential clustering (biclustering) procedure, as described in Figure 10. It first forms bait clusters based on the similarity of prey abundance vectors. Subsequently, the preys in each bait cluster are grouped independently from the other bait clusters. Unlike the method in [47], this method uses the Markov Chain Monte Carlo (MCMC) algorithm to identify biclusters by stochastically drawing samples of bait and prey cluster configurations from a posterior distribution. The biclustering configuration yielding the highest posterior probability is selected as the solution.

Similar to the nested clustering method, BI-MAP [49] is also a Bayesian statistical model for complexes identification. The major difference between the nested clustering approach and the BI-MAP method is that they have different target clusters [49]. The nested clustering approach aims at finding larger parts of protein complexes (sub-complexes), whereas BI-MAP aims at identifying stable and usually smaller structures (modules) that are preserved across all the bait clusters.

So far, there is still no method available that is capable of clustering all kinds of AP-MS data for detecting protein complexes. Existing complex

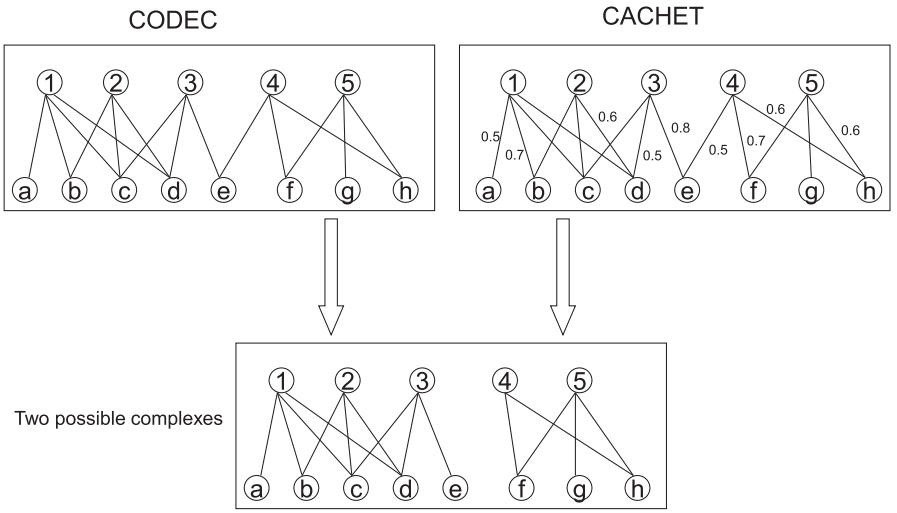


Figure 9: Illustrations of the complex generation procedure in CODEC and CACHET. To minimize the effects of false positive interactions, the CACHET method uses the reliability scores such as SA, PE and DC scores to measure the bait-prey similarity while CODEC approximates the reliability of bait–prey links just by the degree of the vertices.

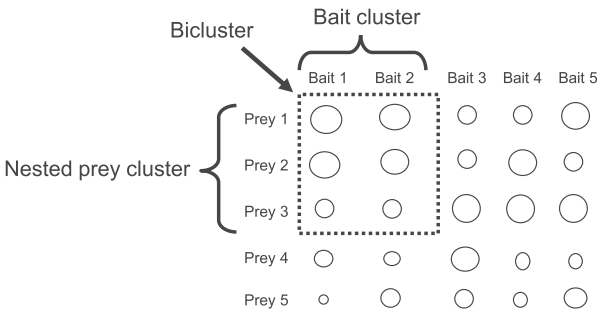


Figure 10: The nested clustering method. The nested clustering algorithm has two steps: (i) grouping baits into clusters based on prey profile similarities, and (ii) identifying nested clusters of preys that share similar abundance level in each bait cluster. Each bicluster corresponds to a submatrix consisting of a bait cluster and an associated nested prey cluster. For example, the nested clustering first groups bait 1 and bait 2 into the bait cluster, and then detects the nested prey cluster that is composed of prey 1, prey 2 and prey 3.

detection algorithms have their own advantages and shortcomings. We give a short summary of methods for detecting the protein complexes from the AP-MS data in Table 3. The methods with multiple phases in the first category are very flexible since we have many different choices in each phase. However, some useful information will be lost when the original AP-MS data is converted into the binary PPI network. In contrast, there will be no such kind of information loss for the methods in the second category. However, both CACHET and CODEC focus on finding clusters with high density so that they may ignore clusters with relatively low density. And the useful quantitative information on protein abundance is not incorporated into these two methods. The methods such as nested clustering and BI-MAP can yield better results, as they fully make use of the quantitative information. However, they are very complicated, as there are many parameters to be specified. In addition, it is very time-consuming to obtain the detection results, as we need to solve hard optimization problems in both methods. Overall, no algorithms can always be the best under all scenarios; therefore, detecting protein complexes from the AP-MS data is still an open problem.

VALIDATING THE RESULTS

Validating the constructed PPI networks, which is as important as developing advanced methods for inferring PPIs and detecting protein complexes, has not received much attention. The validation techniques are conducive not only to the development of new algorithms but also to the verification of the results. Based on whether using additional reference database, the validating methods can be classified into two categories, as shown in Figure 11.

In the first category, the inference result and the gold standard database are used as the input. Because there are no appropriate benchmark data sets, it is a challenge to create a database with high quality. Usually, the PPIs in the database are collected from

Table 3: Summary of methods for detecting protein complexes from the AP-MS data

Method	Input	Advantages	Disadvantages
MCODE [27] Markov clustering [41] DPClus [42]	PPIs from pre-processing the AP-MS data	Flexible since there are many different choices in each phase.	Losing some useful information when the original AP-MS data is converted into the binary PPI network.
CODEC [46] CACHET [45]	Raw AP-MS data (qualitative)	Preserving the information contained in the AP-MS data.	Only focusing on the clusters with high-density; failing to use the quantitative information.
Nested clustering [48] BI-MAP [49]	Raw AP-MS data (quantitative)		Complicated and time-consuming.

multiple sources. These PPIs are filtered out both systematically and manually to remove low-quality/erroneous interactions. In some databases, the collected interactions are classified by the type, such as binary physical interactions and co-complex associations. We list some available PPI databases in Table 4. Using these databases, we can assess the predicted interactions and complexes through some standard performance indices. Some commonly used indices for evaluating the prediction performance are listed in Table 5.

The database-based approach is simple and easy to deploy. Meanwhile, it will be very accurate if the database is complete and all entries in the database are valid. Therefore, for some organisms or species with high-quality PPI database, the database-based approach is the best choice. However, no database is complete and false-positive PPIs may be contained. Anyway, this method is currently the most widely used approach in performance comparison, both for protein–protein interaction inference and protein complex detection.

Alternatively, the reference-free method only uses the inference result and does not need a database as the input. However, this approach requires the repetitive execution of the original PPI inference methods over some simulated data sets, as shown in Figure 12. In [5, 7, 30], under the null hypothesis that a bait captures a list of preys randomly, many shuffled data sets are first generated from a distribution of the abundance of prey proteins. Hence, the simulated data sets are statistically comparable to the original one. Then a p-value representing the likelihood that a given score for the interaction from the original data would occur in the random data sets by chance can be calculated. Finally, the filtering criterion such as false discovery rate (FDR) $< 5\%$ can be used to control the quality of reported interactions. In [57], a permutation framework is adopted as well. The interactions are firstly evaluated with a two-stage Poisson model, and then the procedure of Westfall and Young [58] and the method of Benjamini–Hochberg [59] are employed to estimate the family-wise error rate (FWER) and FDR, respectively.

As the reference-free methods often need to run the original inference algorithm multiple times, it may be time-consuming or prohibitive. However, in the case that resources are limited, i.e. there is no available reference data set, this strategy can be used as an alternative method for validating the analysis results.

Overall, the problem of validating the inference results from the AP-MS data is only partially solved. We suggest to use multiple validation methods if possible. Inevitably, each validation method has certain bias. The utilization of more than one validation method in the evaluation can provide more convincing and comprehensive performance assessment.

CONCLUSIONS

The fast generation of large-scale AP-MS data makes it possible to study protein–protein interactions in a computation-intensive and high-throughput manner. Apparently, the need for computational methods is inevitable. During the past years, we have witnessed the rapid advances in developing the effective algorithms for analyzing the AP-MS data. However, the data analysis problems in this area are far from resolved and there are still many computational challenges to overcome:

- **Scoring the co-complex interactions.** Scoring the co-complex interactions remains an open problem. On one hand, many scoring algorithms are developed only for one special type of data set and do not have strong universality. On the other hand, only the protein pairs that occur in the same purification are considered to be the candidates

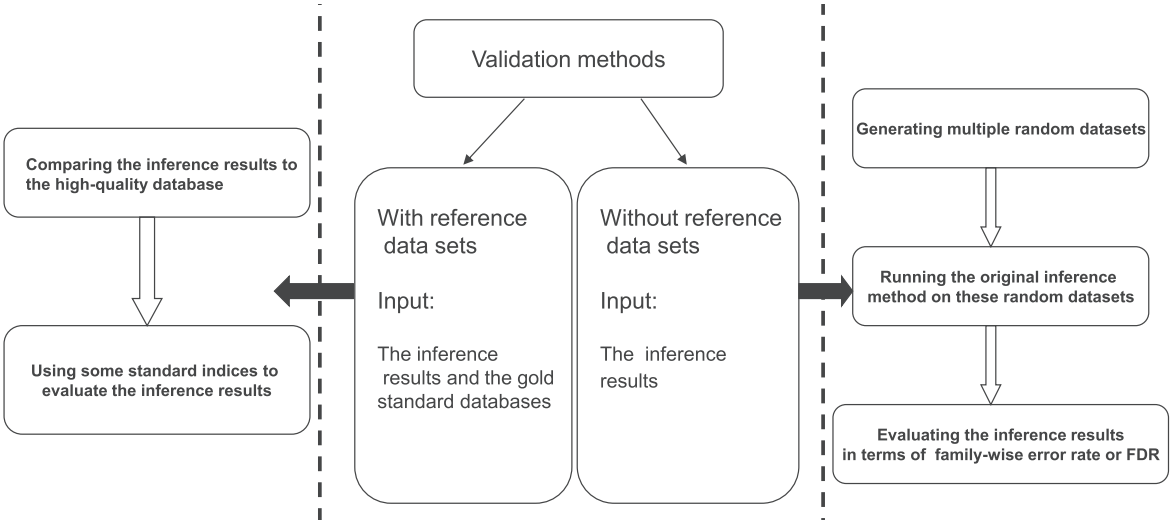


Figure 11: The categorization of validation methods. Based on whether using additional reference data sets, the validation methods can be classified into two categories. In the first category, both the inference result and the gold standard database are taken as the input for performance evaluation. We can evaluate the inference result with some standard performance indices by comparing the prediction result to the reference database. In the second category, the validation methods generally have the following steps: (i) creating multiple simulated data sets that have the same characteristics as the original AP-MS data, (ii) performing interaction prediction or complex detection on these random data sets with the same algorithm, and (iii) calculating the family-wise error rate or false-discovery rate by comparing the original inference result with those generated from the simulated data sets.

Table 4: PPI databases available online

PPI database	URL
MIPS [50]	http://mips.helmholtz-muenchen.de/proj/ppi/
BioGrid [51]	http://biodata.mshri.on.ca/grid/servlet/Index
HPRD [52]	http://www.hprd.org/
IntAct [53]	http://www.ebi.ac.uk/intact/
DIP [54]	http://dip.doe-mbi.ucla.edu/dip/Download.cgi
MINT [55]	http://cbm.bio.uniroma2.it/mint/
HINT [56]	http://hint.yulab.org

Table 5: Indexes for evaluating prediction performance

Index	Formula
Sensitivity	$TP/(TP+FN)$
Specifity	$TN/(TN+FP)$
Positive predicted value (PPV)	$TP/(TP+FP)$
Negative predicted value (NPV)	$TN/(TN+FN)$
Accuracy	$(TP+TN)/(TP+FP+TN+FN)$
F-measure	$2*(PPV*Sensitivity)/(PPV+Sensitivity)$

TP and *TN* stand for true positives and true negatives, respectively. *FP* and *FN* mean false positives and false negatives, respectively.

while those ones which never occur in the same purification but may interact with each other are ignored.

- **Inferring the direct interactions from the indirect ones.** Inferring the direct/physical PPIs from the AP-MS data is far from solved. To date, the methods to solve this problem usually take the raw AP-MS data as the input. An alternative way is to infer the direct PPIs from the co-complex interactions (the indirect ones). In this way, co-complex interactions are first predicted from the AP-MS data and then they are used to construct the networks. The network deconvolution methods such as [60] and [61] are finally employed to recover direct interactions in networks.
- **Detecting the protein complexes with low-density.** Until now, methods of detecting protein complexes from the AP-MS data mostly mine the clusters with high-density, and protein complexes with low-density are always neglected. Novel advanced algorithms are needed to solve this problem.
- **Improved utilization of the quantitative abundance.** With the development of technology, more accurate quantitative information can be gained. It is important to make full use of these resources to build more accurate models.

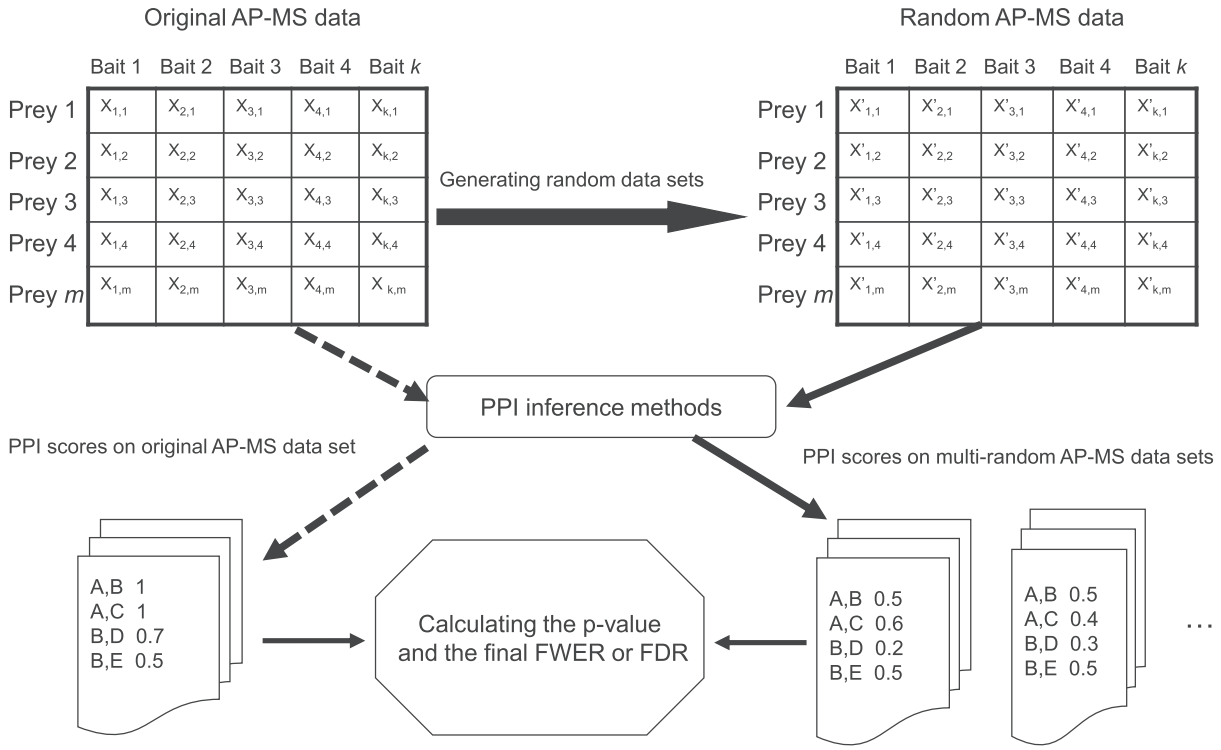


Figure 12: The workflow of reference-free method for validating the prediction results. Under the null hypothesis that each bait protein captures a prey protein at random, the shuffled data sets are generated from a distribution which makes the simulated data sets are statistically comparable to the original one. Then a *p*-value representing the probability that a given score for the interaction from the original data would occur in the random data sets by chance can be calculated. Finally, the FWER or FDR can be obtained.

- **Validating the results.** To date, no PPI database is complete and false-positive PPIs may be contained, which results in that the assessment of the results is not accurate. More complete database with high-quality should be created in the future and the validation methods without using the database also need to be improved.
- **Integrated analysis by incorporating protein identification.** The analysis of the bait–prey data is only one part of the entire AP-MS experiment. Because many new errors may be introduced during the processes of the peptide identification and the proteins inference, analyzing the AP-MS data from a systematic perspective may be a new choice.

The challenges listed above are not complete but provide some interesting research problems. We hope that our review as well as the challenges highlighted here will provide valuable insights into driving future research for further advancing the state-of-the-art technologies in computational prediction and analysis of protein–protein interactions and protein complexes from the AP-MS data.

Key Points

- Raw AP-MS data sets are noisy and error-prone; computational methods should be used to analyze these data sets. There are at least four key computational issues in the analysis of AP-MS data: prediction of co-complex interactions, inference of physical interactions, detection of protein complexes and validation of the results.
- We review some representative algorithms for each computational issue and classify them into different categories according to whether using quantitative information and the underlying scoring strategies used.
- We list some computational challenges in the analysis of the AP-MS data: scoring the co-complex interactions, inferring the direct interactions, detecting the protein complexes with low-density, improved utilization of the quantitative abundance, validating the results and integrated analysis by incorporating protein identification.

FUNDING

This work was partially supported by the Natural Science Foundation of China under Grant No.

61003176, the Fundamental Research Funds for the Central Universities of China (DUT14QY07).

References

1. Alberts B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* 1998;**92**(3):291–4.
2. Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. *Cell* 2011;**144**(6):986–98.
3. Collins SR, Kemmeren P, Zhao XC, *et al*. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* 2007;**6**(3): 439–50.
4. Choi H, Larsen B, Lin ZY, *et al*. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat Methods* 2011;**8**:70–3.
5. Jäger S, Cimermancic P, Gulbahce N, *et al*. Global landscape of HIV-human protein complexes. *Nature* 2012;**481**(7381): 365–70.
6. Choi H, Glatter T, Gstaiger M, *et al*. SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J Proteome Res* 2012;**11**(4):2619–24.
7. Sowa ME, Bennett EJ, Gygi SP, *et al*. Defining the human deubiquitinating enzyme interaction landscape. *Cell* 2009; **138**(2):389–403.
8. Gao J, Opiteck GJ, Friedrichs MS, *et al*. Changes in the protein expression of yeast as a function of carbon source. *J Proteome Res* 2003;**2**(6):643–9.
9. Liu H, Sadygov RG, Yates JR. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 2004;**76**(14):4193–201.
10. Florens L, Washburn MP, Raine JD, *et al*. A proteomic view of the Plasmodium falciparum life cycle. *Nature* 2002; **419**(6906):520–6.
11. Ishihama Y, Oda Y, Tabata T, *et al*. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cel Proteomics* 2005;**4**(9):1265–72.
12. Gavin AC, Aloy P, Grandi P, *et al*. Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006; **440**(7084):631–6.
13. Gomez-Rodriguez M, Leskovec J, Krause A. Inferring networks of diffusion and influence. *ACM Trans on Knowl Discov Data* 2012;**5**(4):21.
14. Rabbat MG, Figueiredo MA, Nowak RD. Network inference from co-occurrences. *IEEE Tran on Inf Theory* 2008; **54**(9):4053–68.
15. Knapp B, Kaderali L. Reconstruction of cellular signal transduction networks using perturbation assays and linear programming. *PLoS One* 2013;**8**(7):e69220.
16. Molinelli EJ, Korkut A, Wang W, *et al*. Perturbation biology: inferring signaling networks in cellular systems. *PLoS Comput Biol* 2013;**9**(12):e1003290.
17. Trinkle-Mulcahy L, Boulon S, Lam YW, *et al*. Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J Cell Biol* 2008;**183**(2): 223–39.
18. Waugh DF. Protein-protein interactions. *Adv in Protein Chem* 1954;**9**:325–437.
19. Ito T, Chiba T, Ozawa R, *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 2001;**98**(8):4569–74.
20. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* 2003;**100**(21):12123–8.
21. Li X, Wu M, Kwoh CK, *et al*. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics* 2010;**11(Suppl 1)**:S3.
22. Wodak SJ, Pu S, Vlasblom J, *et al*. Challenges: rewards of interaction proteomics. *Mol Cell Proteomics* 2009;**8**:3–18.
23. Nesvizhskii AI. Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. *Proteomics* 2012;**12**(10):1639–55.
24. Choi H. Computational detection of protein complexes in AP-MS experiments. *Proteomics* 2012;**12**(10):1663–8.
25. Armean IM, Lilley KS, Trotter MW. Popular computational methods to assess multiprotein complexes derived from label-free affinity purification: mass spectrometry (AP-MS) experiments. *Mol Cell Proteomics* 2013;**12**:1–13.
26. Pardo M, Choudhary JS. Assignment of protein interactions from affinity purification/mass spectrometry data. *J Proteome Res* 2012;**11**(3):1462–74.
27. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003;**4**:2.
28. Yu X, Ivanic J, Wallqvist A, *et al*. A novel scoring approach for protein co-purification data reveals high interaction specificity. *PLoS Comput Biol* 2009;**5**(9):e1000515.
29. Zhang B, Park BH, Karpinets T, *et al*. From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics* 2008;**24**(7):979–86.
30. Guruharsha K, Rual JF, Zhai B, *et al*. A protein complex network of Drosophila melanogaster. *Cell* 2011;**147**(3): 690–703.
31. Tucker G, Loh PR, Berger B. A sampling framework for incorporating quantitative mass spectrometry data in protein interaction analysis. *BMC Bioinformatics* 2013;**14**:299.
32. Sun X, Hong P, Kulkarni M, *et al*. PPIRank-an advanced method for ranking protein-protein interations in TAP/MS data. *Proteome Sci* 2013;**11(Suppl 1)**:S16.
33. Gilmore JM, Auberry DL, Sharp JL, *et al*. A Bayesian estimator of protein–protein association probabilities. *Bioinformatics* 2008;**24**(13):1554–5.
34. Nesvizhskii AI. Protein identification by tandem mass spectrometry and sequence database searching. *Methods in Molecular Biology* 2007;**367**:87–119.
35. Breitkreutz A, Choi H, Sharom JR, *et al*. A global protein kinase and phosphatase interaction network in yeast. *Science* 2010;**328**(5981):1043–6.
36. Saito R, Suzuki H, Hayashizaki Y. Construction of reliable protein–protein interaction networks with a new interaction generality measure. *Bioinformatics* 2003;**19**(6): 756–63.
37. Schelhorn SE, Mestre J, Albrecht M, *et al*. Inferring physical protein contacts from large-scale purification data of protein complexes. *Mol Cell Proteomics* 2011; **10**(6):M110.004929.

38. Krogan NJ, Cagney G, Yu H, *et al*. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 2006;**440**(7084):637–43.
39. Clark NR, Dannenfelser R, Tan CM, *et al*. Sets2Networks: network inference from repeated observations of sets. *BMC Syst Biol* 2012;**6**:89.
40. Kim E, Sabharwal A, Vetta A, *et al*. Predicting direct protein interactions from affinity purification mass spectrometry data. *Algorithms Mol Biol* 2010;**5**:34.
41. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;**30**(7):1575–84.
42. Altaf-Ul-Amin M, Shinbo Y, Mihara K, *et al*. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics* 2006;**7**:207.
43. Xie Z, Kwoh CK, Li XL, *et al*. Construction of co-complex score matrix for protein complex prediction from AP-MS data. *Bioinformatics* 2011;**27**(13):i159–66.
44. Leung HC, Xiang Q, Yiu SM, *et al*. Predicting protein complexes from PPI data: a core-attachment approach. *J Comput Biol* 2009;**16**(2):133–44.
45. Wu M, Li Xl, Kwoh CK, *et al*. Discovery of protein complexes with core-attachment structures from tandem affinity purification (TAP) data. *J Comput Biol* 2012;**19**(9): 1027–42.
46. Geva G, Sharan R. Identification of protein complexes from co-immunoprecipitation data. *Bioinformatics* 2011;**27**: 111–17.
47. Sardiu ME, Cai Y, Jin J, *et al*. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proc Natl Acad Sci USA* 2008;**105**(5): 1454–9.
48. Choi H, Kim S, Gingras AC, *et al*. Analysis of protein complexes through model-based biclustering of label-free quantitative AP-MS data. *Mol Syst Biol* 2010;**6**:385.
49. Stukalov A, Superti-Furga G, Colinge J. Deconvolution of targeted protein-protein interaction maps. *J Proteome Res* 2012;**11**(8):4102–9.
50. Pagel P, Kovac S, Oesterheld M, *et al*. The MIPS mammalian protein-protein interaction database. *Bioinformatics* 2005;**21**(6):832–4.
51. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, *et al*. The BioGRID interaction database: 2011 update. *Nucleic Acids Res* 2011;**39(Suppl 1)**:D698–704.
52. Peri S, Navarro JD, Amanchy R, *et al*. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 2003; **13**(10):2363–71.
53. Aranda B, Achuthan P, Alam-Faruque Y, *et al*. The IntAct molecular interaction database in **2010**. *Nucleic Acids Res* 2010;**38(Suppl 1)**:D525–31.
54. Salwinski L, Miller CS, Smith AJ, *et al*. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004; **32(Suppl 1)**:D449–51.
55. Licata L, Briganti L, Peluso D, *et al*. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 2012; **40**(D1):D857–61.
56. Das J, Yu H. HINT: high-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol* 2012;**6**:92.
57. Fischer M, Zilkenat S, Gerlach RG, *et al*. Pre-and post-processing workflow for affinity purification mass spectrometry data. *J Proteome Res* 2014;**13**(5):2239–49.
58. Westfall PH, Young SS. *Resampling-based multiple testing: examples and methods for P-value adjustment. Wiley series in probability and mathematical statistics (Applied probability and statistics)*. New York: John Wiley & Sons, 1993.
59. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995;**57**:289–300.
60. Feizi S, Marbach D, Médard M, *et al*. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat Biotechnol* 2013;**31**(8): 726–33.
61. Barzel B, Barabási AL. Network link prediction by global silencing of indirect correlations. *Nat Biotechnol* 2013;**31**(8): 720–5.