PPI network inference from AP-MS data

6

6.1 Introduction to protein–protein interactions

Proteins mediate their functions physically by interacting with each other in stable or transient multiprotein complexes of distinct composition. To understand the diverse and dynamic proteome, it is necessary to construct the underlying networks of physical interactions. Moreover, proteins can interact with other molecules, such as metabolites, lipids, and nucleic acids. These complexes have essential roles in regulatory processes and cellular functions. Therefore, the construction and analysis of such interactome networks will provide important insights into the global organization of cellular systems.

There are two main types of high-throughput experimental methods for identifying protein–protein interactions (PPIs): yeast two-hybrid (Y2H) and affinity purification mass spectrometry (AP-MS).

Y2H primarily detects direct PPIs and can quickly screen large numbers of such interactions. Users often regard data from Y2H experiments as the proof of a direct interaction. However, the quality of Y2H data sets has been controversial because different Y2H systems can generate markedly different interactions in the same interactome. This is because false positives do occur when yeast proteins act as a bridge between two indirectly interacting proteins.

Recently, the combination of large-scale affinity purification (AP) with mass spectrometry (MS) is widely used to detect and characterize protein complexes. In AP-MS, a protein of interest ("bait") is affinity captured and then followed by MS to identify its interaction partners ("prey"). AP-MS can delineate the dynamics of interactions at different conditions to determine protein complex compositions.

6.2 AP-MS data generation

In a typical AP/MS experiment, selected proteins of interest ("baits") are purified along with their interactors ("preys") through one or more AP steps, as illustrated in Figure 6.1. Proteins in the affinity purified sample are digested into peptides with protease such as trypsin. The mixtures of peptides are separated using liquid chromatography coupled online to a mass spectrometer. Eluting peptides are ionized, transferred into a gas phase, and selected peptide ions are further fragmented to generate tandem mass spectra. These acquired spectra are used to identify peptides and proteins. The output data from such an AP/MS experiment is the bait protein and its possible interaction partners (prey proteins), and each protein can also be associated with its abundance information that is provided by the protein quantification method.

Data Mining for Bioinformatics Applications. http://dx.doi.org/10.1016/B978-0-08-100100-4.00006-5

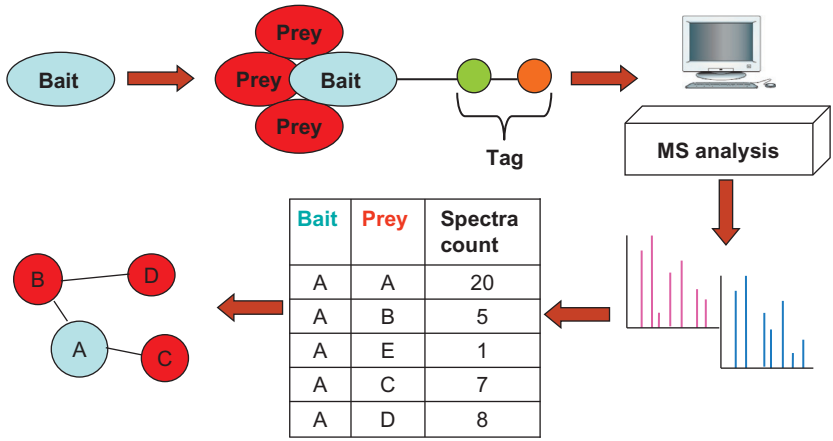

Figure 6.1 A typical AP-MS workflow for constructing PPI network. A typical AP-MS study performs a set of experiments on bait proteins of interest, with the goal of identifying their interaction partners. In each experiment, a bait protein is first tagged and expressed in the cell. Then, the bait protein and their potential interaction partners (prey proteins) are affinity purified using AP. The resulting proteins (both bait and prey proteins) are digested into peptides and passed to tandem mass spectrometer for analysis. Peptides are identified from the MS/MS spectra with peptide identification algorithms and proteins are inferred from identified peptides with protein inference algorithms. In addition, the label-free quantification method such as spectral counting is typically used to estimate the protein abundance in each experiment. Such pull-down bait–prey data from all AP-MS runs are used to filter contaminants and construct the PPI network.

Intuitively, each bait–prey protein pair should interact directly with others. In practice, however, there are a large number of false-positive interactions, where a prey protein can be either a contaminant or can interact with the bait indirectly.

6.3 Data collection and preprocessing

To conduct the PPI network, raw AP-MS data or preprocessed data need to be collected first. To date, many such AP-MS data sets are available online for academic use. Most of these public data sets have been preprocessed, in which the identified bait–prey pairs from raw MS data are provided to users. That is, the users need not perform preprocessing steps such as peptide identification and protein inference to obtain the bait–prey list. Table 6.1 lists some online AP-MS data resources that have been used in some recent studies.

6.4 Modeling with different data mining techniques

Generally, PPIs can be divided into two major types [1]: co-complex interactions and physical interactions. A protein complex is a group of proteins that interact with each other at the same location and time. The protein pair of co-complex interaction

Table 6.1 Some AP-MS data sets available online

Reference	URL
[2]	http://kroganlab.ucsf.edu/links.html
[3]	http://www.nature.com/nmeth/journal/v8/n1/abs/nmeth.1541.html
[4]	http://www.nature.com/nature/journal/v481/n7381/abs/nature10719.html
[5]	http://pubs.acs.org/doi/suppl/10.1021/pr201185r
[6]	http://www.sciencedirect.com/science/article/pii/S0092867409005030

interacts with each other in the formation of a complex. That is, if two proteins belong to the same complex (they do not have direct physical contact), then they have the co-complex interaction relationship. Unlike co-complex interactions, the physical interaction represents a direct biophysical interaction between two proteins, that is, two proteins are linked by an edge in the PPI network.

To construct the PPI network, either co-complex interactions or physical interactions can be used as the edges to connect protein nodes. Therefore, the key problem in the network inference from AP-MS data is how to accurately predict co-complex interactions and physical interactions. From the viewpoint of data mining, there are several different modeling approaches for such an interaction prediction task.

6.4.1 A correlation mining approach

The problem of interaction prediction from bait–prey data can be modeled as a correlation/association mining problem. The following computational framework can be used to illustrate most existing algorithms in this category.

First, the original AP-MS data are transformed into a type-value table. Here we use one simple example to illustrate this procedure, where a preprocessed bait–prey data set is given in Figure 6.2. In this sample data, there are six purifications where each selected bait protein captures a list of prey proteins. The transformed type-value table is shown in Table 6.2, where rows correspond to purifications, and each column represents the presence status of a specific protein in the corresponding purification. Note that the protein type (bait vs. prey) and the protein abundance information (e.g., spectral count) can also be recorded as the associated information with each cell in the table.

Based on the transformed table, the interaction prediction between different proteins could be modeled as the problem of pair-wise correlation mining among different features. These methods differ in several ways: the correlation measure used and the information used in the correlation calculation.

In Ref. [7], the Dice coefficient (DC) is used to measure the correlation between two proteins. More precisely, the interaction score between two proteins i and j is defined as: $2O_{ij}/(2O_{ij}+O_i+O_j)$, where O_{ij} is the number of times that both proteins i and j are 1s in the table, and O_i and O_j represent the number of times that only protein i is 1 and only protein j

Purification 1

Bait	Prey	Spectra count
A	A	41
A	B	16
A	C	31
A	D	19
A	E	15

Purification 2

Bait	Prey	Spectra count
B	B	20
B	F	5
B	G	1
B	C	7
B	A	18

Purification 3

Bait	Prey	Spectra count
C	C	5
C	A	8
C	D	15
C	E	2
C	H	1

Purification 4

Bait	Prey	Spectra count
D	D	30
D	E	67
D	F	1
D	G	33
D	H	3

Purification 5

Bait	Prey	Spectra count
E	E	22
E	F	15
E	D	11
E	B	3
E	C	1

Purification 6

Bait	Prey	Spectra count
F	F	25
F	A	18
F	G	5
F	E	12
F	D	18

Figure 6.2 A sample AP-MS data set with six purifications.

Table 6.2 The transformed type-value table of the sample data in Figure 6.2

	A	B	C	D	E	F	G	H
1	1 (b, 41)	1 (p, 16)	1 (p, 31)	1 (p, 19)	1 (p, 15)	0	0	0
2	1 (p, 18)	1 (b, 20)	1 (p, 7)	0	0	0	1 (p, 1)	0
3	1 (p, 8)	0	1 (b, 5)	1 (p, 15)	1 (p, 2)	0	0	1 (p, 1)
4	0	0	0	1 (b, 30)	1 (p, 67)	1 (p, 1)	1 (p, 33)	1 (p, 3)
5	0	1 (p, 3)	1 (p, 3)	1 (p, 11)	1 (b, 22)	1 (p, 15)	0	0
6	1 (p,18)	0	0	1 (p, 18)	1 (p, 12)	1 (b, 25)	1 (p, 5)	0

In the new data set, there are six samples and eight features. In each cell, 1 (0) denotes that the protein is present (absent) in the corresponding purification. In addition, the protein type and protein abundance information can be associated with each cell if its value is 1. Here b and p denote that the protein in the purification is a bait protein and a prey protein, respectively.

is 1, respectively. For instance, the interaction score based on DC is 4/7 for protein A and protein B in Table 6.2. Clearly, this method ignores both the bait/prey information and the protein abundance information in its scoring function.

Figure 6.3 shows the predicted PPI network when DC is used as the correlation measure with a cut-off threshold of 0.5.

In Ref. [8], the socio-affinity (SA) scoring method is proposed, which incorporates the protein type into the correlation calculation procedure. More precisely, the

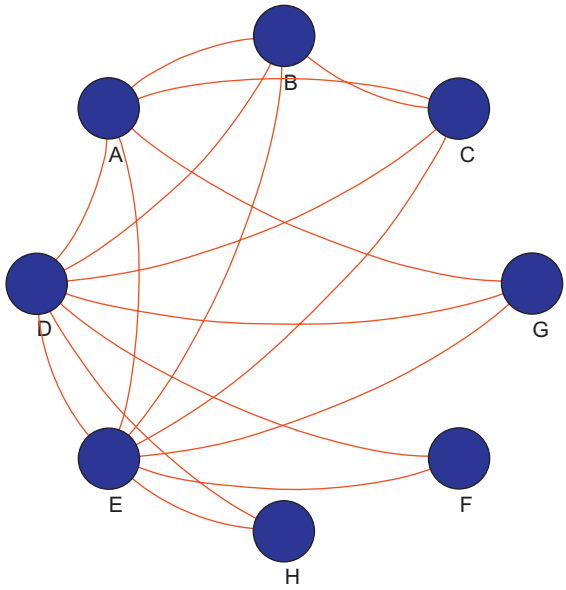

Figure 6.3 The PPI network constructed from the sample data. Here DC is used as the correlation measure and the score threshold is 0.5, that is, a protein pair is considered to be a true interaction if the DC score is above 0.5. In the figure, the width of the edge that connects two proteins is proportional to the corresponding DC score.

correlation score between two proteins i and j is composed of three parts: (1) the correlation score when protein i is the bait; (2) the correlation score when protein j is the bait; and (3) the correlation score when both protein i and protein j are not the bait. For example, if SA is used to calculate the association between protein A and protein B in Table 6.2, then the samples in Table 6.2 are divided into three sets: {1}, {2}, and {3,4,5,6}. The correlation score is first calculated for each subset and then the final score is the sum of three component scores.

The protein abundance information can be used in calculating the interaction score as well. In Ref. [9], the correlation of two proteins is defined in terms of the cosine-distance between their columns:

$$\mathrm{sim}(i,j) = \frac{\sum_{k=1}^{n} x_{ki}x_{kj}}{\sqrt{\sum_{k=1}^{n} x_{ki}^2}\sqrt{\sum_{k=1}^{n} x_{kj}^2}}, \tag{6.1}$$

where $x_{ki}(x_{kj})$ is the protein abundance of the ith (jth) protein in the kth purification and n is number of total purifications. For instance, the cosine-distance between protein A and protein B in Table 6.2 is calculated as:

$$\frac{(41\times 16+18\times 20+8\times 0+0\times 0+0\times 1+1\times 0)}{\sqrt{(41^2+18^2+8^2+0^2+0^2+1^2)}\sqrt{16^2+20^2+0^2+0^2+1^2+0^2}} = 0.87$$

6.4.2 A discriminative pattern mining approach

In addition to the experimental data for bait proteins, AP-MS data often contain negative controls, in which the preys are likely to be contaminants because the tagged bait is not expressed. When the negative controls are available, it is possible to directly identify noninteracting bait–prey pairs from the perspective of discriminative pattern mining.

In Ref. [10], the protein expressions obtained from specific experiments are compared with the protein expressions obtained from negative control experiments to determine if a prey protein is a contaminant, that is, if the corresponding bait–prey pair is not a true interaction pair.

For m identified proteins and n purifications, let x_{ki} represent the normalized spectral abundance factor (NSAF) of the kth protein in the ith experiment. The vector $[x_{k1}, x_{k2}, \ldots, x_{kn}]$ is defined as "protein vector." Similarly, let y_{ki} represent the NSAF value of kth identified protein in the ith control purification of negative controls. Then, the vector $[y_{k1}, y_{k2}, \ldots, y_{kn}]$ represents the negative control protein vector. For each protein with two protein vectors, the vector magnitude (α) is calculated as:

$$\alpha = \sqrt{\frac{\langle y, y\rangle}{\langle x, x\rangle}} = \sqrt{\frac{y_{k1}^2 + y_{k2}^2 + \cdots + y_{kn}^2}{x_{k1}^2 + x_{k2}^2 + \cdots + x_{kn}^2}}. \tag{6.2}$$

If $\alpha \geq 1$, then the protein was more abundant in the negative controls than in the specific experiments. Hence, a protein with $\alpha \geq 1$ can be considered as a contaminant since it is "over-expressed" in the negative controls.

6.5 Validation

Based on whether one is using the additional reference database, there are two commonly used strategies for validating the constructed PPI networks. In the first strategy, the predicted network and the gold standard database are used as the input. Usually the PPIs in the reference database are collected from multiple sources and are postprocessed to remove erroneous interactions. In some databases, the interactions are further classified according to the type of interactions: physical interactions and co-complex interactions. Table 6.3 shows some available databases that are used for interaction validation. According to these databases, the predicted interactions can be evaluated through some standard performance indices such as accuracy and false discovery rate.

The database-based approach is probably the most widely used method for validating the interaction prediction results. This method is very accurate if the database is complete and all stored interactions in the database are valid. However, most databases are still not complete so far. Meanwhile, there may be some invalid interaction entries in the database. For some species or organisms, such a database-based validation approach is not applicable because there is still no gold standard database available in the literature.

Table 6.3 Some high-quality protein–protein interaction database available online

Database	URL
MIPS	http://mips.helmholtz-muenchen.de/proj/ppi/
BioGrid	http://biodata.mshri.on.ca/grid/servlet/Index
HPRD	http://www.hprd.org/
IntAct	http://www.ebi.ac.uk/intact/
DIP	http://dip.doe-mbi.ucla.edu/dip/Download.cgi
MINT	http://cbm.bio.uniroma2.it/mint/
HINT	http://hint.yulab.org

Alternatively, the database-free method does not need a gold standard database, which uses the predicted interactions and original AP-MS data set as the input. Generally, such database-free approaches have the following steps, as described in Figure 6.4:

1. The first step is to generate multiple simulated data sets that have the same characteristics as the original AP-MS data. Here the null hypothesis is: each bait captures a list of preys randomly. Under this null hypothesis, many simulated data sets can be created by randomly swapping two prey proteins from different baits. Hence, these simulated data sets are statistically comparable to the original one.
2. Then, PPIs are predicted on these random data sets with the algorithm to be evaluated. That means the original interaction prediction method should be repetitively executed many times.
3. Finally, the family-wise error rate or false discovery rate can be obtained by comparing the original prediction result with those generated from the simulated data sets.

6.6 Discussion and future perspective

The fast generation of AP-MS data makes it possible to find true PPIs with data mining methods. However, the data analysis problem has some special characteristics that are different from standard data mining issues and poses some new computational challenges, as we discuss below.

It is a very natural idea to model the interaction prediction problem as a pair-wise correlation mining issue. That is, the cooccurrence correlation between two proteins across the purifications is used to measure their interaction strength. However, such a strategy relies on the assumption that there are a sufficient number of purifications so that interacting proteins can cooccur frequently. Currently, small-to-intermediate scale AP-MS data sets are becoming increasingly popular, in which we may observe the cooccurrence between a tagged bait protein and a prey protein only once. This makes it unfeasible to employ traditional correlation mining techniques in the context of small-to-intermediate scale AP-MS experiments. Therefore, some new correlation mining formulations should be presented to handle such a distinct data analysis problem.

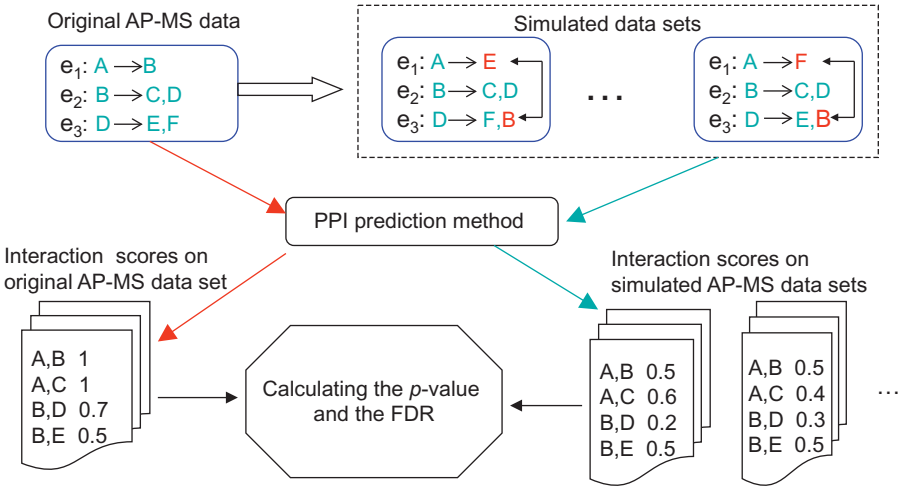

Figure 6.4 An illustration of database-free method for validating the interaction prediction results. Under the null hypothesis that each bait protein captures a prey protein is a random event, some simulated data sets are generated such that they are comparable to the original one. Then, an empirical *p*-value representing the probability that an original interaction score for a protein pair would occur in the random data sets by chance can be calculated. Finally, the false discovery rate is calculated according to these *p*-values.

Even for large-scale AP-MS data sets, it is still a challenging problem to determine how to assign weights to different protein types automatically in the correlation calculation procedure. The SA method [8] solves this problem by dividing the data into different parts according to if one of two proteins has been tagged as bait. This strategy is very effective in practice. However, it is probably not the best solution. New methods that can determine the weights of different protein types in a mathematically sound manner should be developed in the future.

Furthermore, current interaction prediction algorithms can only detect interaction between proteins that cooccur in the same purification. Those proteins, which never occur in the same purification but may interact with each other, are ignored. To fulfill this void, one feasible solution is to use the indirect association mining technique to find such kinds of protein interactions.

References

[1] B. Teng, C. Zhao, X. Liu, Z. He, Network inference from AP-MS data: computational challenges and solutions. Brief. Bioinform. (2014), http://dx,doi.org/10.1093/bib/bbu038.

[2] S.R. Collins, P. Kemmeren, X.C. Zhao, et al., Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*, Mol. Cell. Proteomics 6 (3) (2007) 439–450.

[3] H. Choi, B. Larsen, Z.Y. Lin, et al., SAINT: probabilistic scoring of affinity purification-mass spectrometry data, Nat. Methods 8 (2011) 70–73.
[4] S. Jager, P. Cimermancic, N. Gulbahce, et al., Global landscape of HIV-human protein complexes, Nature 481 (7381) (2012) 365–370.
[5] H. Choi, T. Glatter, M. Gstaiger, et al., SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments, J. Proteome Res. 11 (4) (2012) 2619–2624.
[6] M.E. Sowa, E.J. Bennett, S.P. Gygi, et al., Defining the human deubiquitinating enzyme interaction landscape, Cell 138 (2) (2009) 389–403.
[7] B. Zhang, B.H. Park, T. Karpinets, et al., From pull-down data to protein interaction networks and complexes with biological relevance, Bioinformatics 24 (7) (2008) 979–986.
[8] A.C. Gavin, P. Aloy, P. Grandi, et al., Proteome survey reveals modularity of the yeast cell machinery, Nature 440 (7084) (2006) 631–636.
[9] J. Kutzera, H.C.J. Hoefsloot, A. Malovannaya, et al., Inferring protein–protein interaction complexes from immunoprecipitation data, BMC Res. Notes 6 (2013) 468.
[10] M.E. Sardiu, Y. Cai, J. Jin, et al., Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics, Proc. Natl. Acad. Sci. U.S.A. 105 (5) (2008) 1454–1459.