Protein complex identification from AP-MS data

7

7.1 An introduction to protein complex identification

In Chapter 6, we discussed some novel computational algorithms for predicting interactions that draw the raw bait–prey data one step closer to the target interactomes. However, each protein often carries out its functions by working as a member of a protein complex. A *protein complex* is a collection of proteins that interact with each other at the same time and location, and which has essential roles in regulatory processes, cellular functions, and signaling cascades. Therefore, it is critical to organize the filtered (or even unfiltered) interaction data into protein complexes. As a result, the detection of protein complexes is another important problem in the analysis of the AP-MS data.

Computationally, the problem of protein complexes detection is to find collections of protein groups, where proteins in each group tightly cooperate with each other. Intuitively, if two proteins belong to the same complex, then they should frequently cooccur together (directly or indirectly) in the AP-MS data.

7.2 Data collection and data preprocessing

Existing protein complex detection methods can be divided into two categories according to the input data [1]. First, some methods detect protein complexes from an established PPI network, which is generated by connecting protein pairs that interact with each other. Here the interaction relationship between different proteins is obtained from the interaction prediction methods listed in Chapter 6. Alternatively, other methods detect complexes from original AP-MS data directly. These methods usually model the AP-MS data as a bipartite graph in which the two sets of vertices are the set of baits and the set of preys, respectively.

7.3 Modeling: A graph clustering framework

From the viewpoint of data mining, the protein complex detection problem is essentially a cluster analysis problem. Despite the seeming difference among existing methods with respect to the input data, clustering criterion, and algorithmic principle, these methods can be unified under a common *graph clustering* framework.

A *graph* is a data structure that is formed by a set of *vertices* and a set of *edges*, where each edge connects two vertices. Graph clustering is to find groups of related

Data Mining for Bioinformatics Applications. http://dx.doi.org/10.1016/B978-0-08-100100-4.00007-7

vertices in a graph, where the vertices within each group are highly connected whereas there are only few edges between different groups [2].

To solve the graph clustering issue, scholars from different disciplines have already proposed numerous methods. These methods can be divided into several different categories [3]: graph partitioning, hierarchical clustering, partitional clustering, spectral clustering, modularity-based methods, dynamic algorithms (e.g., random walk), clique percolation, methods based on statistical inference, among others. Roughly, each of existing protein complex detection methods can find its corresponding category in the area of graph clustering. In the following, several typical protein complex detection methods are used to illustrate this point.

7.3.1 The clique percolation approach

The clique percolation method is based on the concept that the internal edges of a cluster are likely to form cliques due to their high density [4]. This method has been widely applied to solve the protein complex identification problem, where protein complexes generally correspond to dense subgraphs or cliques of the network.

CACHET [5] is a typical clique percolation method for identifying protein complexes from AP-MS data. It models 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. Each edge between a bait node and a prey node represents the bait–prey relationship in the original data. This method mainly has three steps:

1. First, all maximal bicliques are generated from the bipartite graph based on the assumption that proteins within protein-complex cores are highly interactive.
2. Next, potential false positive bait–prey edges are removed from each biclique to generate a set of reliable bicliques. Because these reliable bicliques may share many common proteins, a maximal independent set of them are used as the so-called protein-complex cores.
3. Finally, attachment proteins are included into the cores to form protein complexes with core-attachment substructures.

Figure 7.1 presents an example of a bait–prey bipartite graph, where $\{B1, B2, B3, B4\}$ is the set of bait proteins and $\{P1, P2, P3, P4, P5, P6\}$ is the set of prey proteins. The number associated with each edge denotes the reliability score of the corresponding bait–prey interaction, which is generated from the interaction prediction methods in Chapter 6, such as SA. This example graph is used to depict how the CACHET method identifies protein complexes via clique percolation.

First, three maximal bicliques, $G1$, $G2$, and $G3$, are extracted from the bipartite graph, as shown in Figure 7.2. If a clique is considered to be reliable when the average reliability score of its edges is larger than a threshold, then only two bicliques, $C1$ and $C2$, are considered to be reliable (suppose the threshold is 0.7), and $C3$ is discarded. Because $C1$ and $C2$ overlap with each other significantly, only $C1$ is finally reported as a protein-complex core.

Finally, the protein complex is identified by including bait protein $B3$ as the attachment of protein-complex core $C1$, as shown in Figure 7.3.

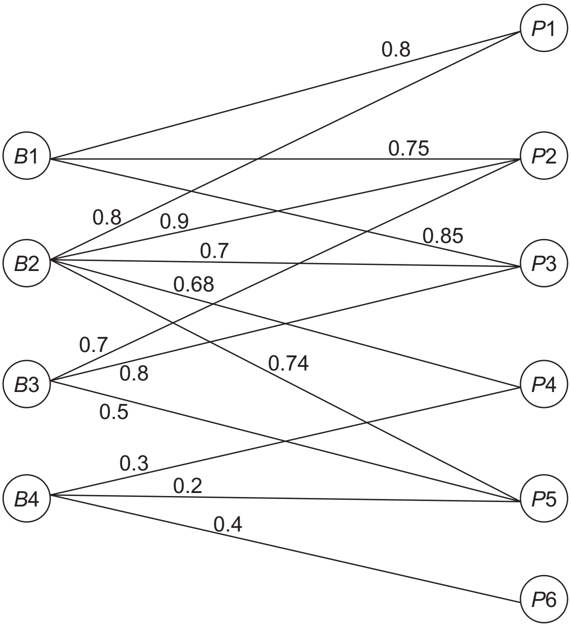
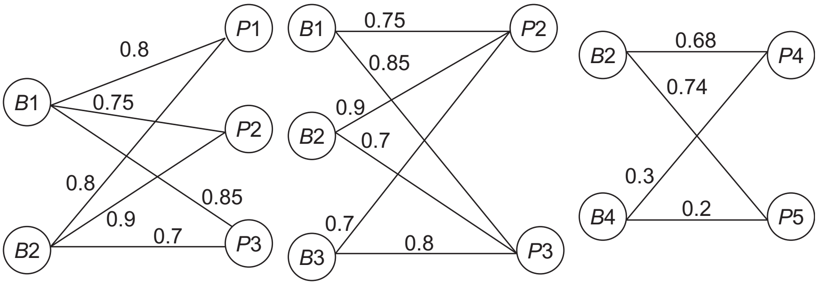


Figure 7.1 An example bait–prey graph. In this figure, each Bi ($i = 1, 2, 3, 4$) denotes a bait protein and each Pi ($i = 1, 2, 3, 4, 5, 6$) represents a prey protein. The score that measures interaction strength between a bait–prey pair is provided as well.



(1) *C*1 (2) *C*2 (3) *C*3

Figure 7.2 Three maximal bicliques are identified. Among these three bicliques, $C1$ and $C2$ are reliable and only $C1$ is finally reported as a protein-complex core.

7.3.2 The statistical inference method

Statistical inference deduces properties of data sets from a set of observations and hypotheses. When the data set is a graph, the clustering objective is to find a partition model that best fits the graph based on the connectivity patterns of vertices. In the context of protein complex identification, Bayesian inference is widely adopted in which observations (bait–prey graph) are used to estimate the probability that a given hypothesis is true.

There are two basic ingredients in Bayesian inference: the observed evidence and a statistical model with some parameters. Bayesian inference starts by writing the

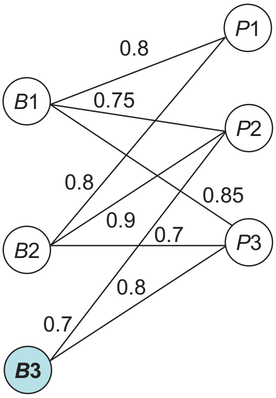


Figure 7.3 The final protein complex by including both the protein complex core $C1$ and an attachment $B3$.

likelihood that the observed evidence is generated by the model for a given set of parameters. The inference is performed to find parameters that maximize the posterior distribution of the parameters given the model and the evidence. Graph clustering can be considered to be a specific example of Bayesian inference problem, where the evidence is the graph structure and a hidden partition model that one wishes to infer along with some parameters.

In Ref. [6], a Bayesian approach is proposed to identify protein complexes from AP-MS data. To illustrate this method, we use the sample data set in Chapter 6 (Figure 6.2) as an example. As shown in Table 7.1, the original AP-MS data are first transformed into a binary purification matrix $\mathbf{U}$ with the size of $R \times N$, where R is the number of bait proteins and N is the number of all proteins that have once appeared in the purifications. Then, the corresponding adjacency matrix $\mathbf{M}$ in Table 7.2 is defined by $\mathbf{M}=\mathbf{U}^{\mathbf{T}}\mathbf{U}$, which is a symmetric $N \times N$ matrix. The ijth element, $\mathbf{M}_{ij}$, is the number of purifications in which protein i and protein j cooccur.

The element $\mathbf{M}_{ij}$ in the adjacency matrix can be regarded as the number of distinct "paths" between protein i and protein j discovered by the AP-MS experiment. For example, there are three paths between protein A and protein C and no path that

Table 7.1 **The binary purification matrix**

Protein	A	B	C	D	E	F	G	H
Bait A	1	1	1	1	1	0	0	0
Bait B	1	1	1	0	0	0	1	0
Bait C	1	0	1	1	1	0	0	1
Bait D	0	0	0	1	1	1	1	1
Bait E	0	1	1	1	1	1	0	0
Bait F	1	0	0	1	1	1	1	0

Table 7.2 **The adjacency matrix derived from Table 7.1**

	A	B	C	D	E	F	G	H
A	4	2	3	3	3	1	2	1
B	2	3	3	2	2	1	1	0
C	3	3	4	3	3	1	1	1
D	3	2	3	5	5	3	2	2
E	3	2	3	5	5	3	2	2
F	1	1	1	3	3	3	2	1
G	2	1	1	2	2	2	3	1
H	1	0	1	2	2	1	1	2

directly connects protein B and protein H. However, it is possible to reach protein H indirectly from protein B through their neighbors. For instance, B can connect with H via the path B→D→H. The number of distinct paths between two proteins via another protein can be directly obtained by the matrix product **MM**. More generally, the number of paths from protein i to protein j of length l on the graph corresponds to the ijth element of the matrix $\boldsymbol{M}^l$. Therefore, the number of distinct paths with different lengths can be used to measure the "similarity" between two proteins. Based on this observation, the von Neumann diffusion kernel is used to evaluate the likelihood of two proteins belonging to the same complex in Ref. [6]:

$$K = \sum_{l=1}^{\infty} \gamma^{l-1} M^l = M(1-\gamma M)^{-1}, \tag{7.1}$$

where γ is a parameter (the diffusion factor) to make the effect of longer paths decay exponentially. The kernel can be normalized into [0, 1] in the following way:

$$S_{ij} = \frac{K_{ij}}{\sqrt{K_{ii} K_{jj}}}. \tag{7.2}$$

Because the elements of von Neumann kernel matrix are between 0 and 1, this makes $\mathbf{S}_{ij}$ suitable as a probabilistic measure for evaluating the likelihood of two proteins belonging to the same complex.

To identify protein complexes, a binary matrix **Z** for protein complex membership is defined as well. Each entry z_{ci} in **Z** is a random variable, which indicates the membership of the ith protein in the cth complex. Note that the number of protein complexes is unknown in advance and one protein may belong to multiple complexes. The task here is to infer the unknown protein membership matrix **Z** from the observed AP-MS data.

Because the actual number of protein complexes is unknown, an infinite latent feature model is employed for protein complex membership identification [6]. Initially, the method starts with a finite model of C complexes, and then takes the limit as $C \to \infty$ to obtain the prior distribution over the binary matrix **Z**.

If each protein belongs to a complex c with probability π_c, then the conditional probability $P(Z|\pi)$ is a product of binomial distributions:

$$P(Z|\pi) = \prod_{c=1}^{C}\prod_{i=1}^{N} P(z_{ci}|\pi_c) = \prod_{c=1}^{C} \pi_c^{n_c}(1-\pi_c)^{N-n_c}, \tag{7.3}$$

where $n_c = \sum_{i=1}^{N} z_{ci}$ is the number of proteins in the cth complex.

If the prior distribution of π is a beta distribution $beta(\alpha/C, 1)$ with a model parameter α, then conditional distribution for any z_{ci} is:

$$P(z_{ci}|Z_{-i,c}) = \frac{n_{-i,c} + \alpha/C}{N + \alpha/C}, \tag{7.4}$$

where $Z_{-i,c}$ represents the set of all entries in **Z** except z_{ci}, and $n_{-i,c}$ is the number of proteins (excluding the ith protein) in the cth complex.

If we let $C \to \infty$, then the conditional distribution of z_{ci} becomes

$$P(z_{ci}|Z_{-i,c}) = \frac{n_{-i,c}}{N} \tag{7.5}$$

for any c such that $n_{-i,c} > 0$.

For the c's with $n_{-i,c} = 0$, the number of new complexes associated with this protein has a Poisson distribution:

$$P(v_i|Z_{-i,c}) = \left(\frac{\alpha}{N}\right)^{v_i} \frac{\exp\left(-\frac{\alpha}{N}\right)}{v_i!}, \tag{7.6}$$

where v_i is the expected number of new complexes.

For a given protein complex membership matrix **Z**, the inner product of two protein column vectors $z_i^{\mathrm{T}} z_j$ can be used to check if protein i and protein j belong to the same complex. That is, two proteins are in the same complex if $z_i^{\mathrm{T}} z_j > 0$. Then, the likelihood can be evaluated as:

$$P(S|Z) = \prod_{\{ij:z_i^{\mathrm{T}} z_j > 0\}} (S_{ij})^{z_i^{\mathrm{T}} z_j} \prod_{\{ij:z_i^{\mathrm{T}} z_j = 0\}} (1 - S_{ij}), \tag{7.5}$$

where **S** is the normalized von Neumann kernel matrix obtained from the AP-MS data.

According to the Bayes' theorem, the posterior distribution of the protein complex membership matrix **Z** is $P(Z|S)$, which is proportional to $P(S|Z)P(Z)$, where $P(S|Z)$ is given in (7.5), and $P(Z)$ is defined by the infinite latent feature model [6]. To carry out the inference, a Gibbs sampler with the following steps is used.

```
(1) Initialize **Z** randomly.
(2) For $t=1$ to T
    (a) According to $P(z_{ci}|Z_{-i,c},S) \propto P(S|Z)P(z_{ci}|Z_{-i,c})$, sample $z_{ci}$ for each $i$ and each $c$ with $n_{-i,c} > 0$.
    (b) According to $P(v_i|Z_{-i,c},S) \propto P(S|Z)P(v_i|Z_{-i,c})$, sample the number of new complexes for each $i$.
    (c) Save the sample **Z**
```

7.4 Validation

Essentially, the validation of detected protein complexes is a special issue of cluster validation. Due to the popularity of cluster analysis, the data mining literature has proposed many cluster validation techniques. These techniques fall into two categories: external validation methods and internal validation methods [7].

External validation methods evaluate a clustering result based on the knowledge of the correct class labels. These methods are only applicable when the true cluster structure is known in advance. In the context of protein complex detection, this means that the set of true protein complexes is available. The database-based validation approach presented in Chapter 6 is an external validation method, which can be applied to evaluate the protein complex identification results as well.

If no class label is available, the internal validation methods become appropriate. Internal validation techniques measure how well a given partition corresponds to the true cluster structure of the data based on the information intrinsic to the data alone. The database-free method (i.e., randomization method) in Chapter 6 does not use additional knowledge of known protein–protein interactions and protein complexes, which is a special internal validation method in the context of cluster analysis.

In addition to the randomization method discussed in Chapter 6, other types of internal measures can be used to assess the quality of detected protein complexes as well. These measures differ in their particular notions of clustering quality that they employ [7].

- *Compactness*: Intracluster homogeneity measures such as the average or maximum pairwise intracluster distances.
- *Separation*: Measures that quantify the degree of separation between individual clusters such as the average intercluster distance.
- *Compliance between a partition and distance information*: The degree to which distance information in the original data is preserved in the clustering result.

Note that the above internal measures are presented for general cluster analysis. In the context of graph clustering for protein complex detection, a novel notion of "distance" is needed to be clarified for a protein pair. For this purpose, the similarity score for measuring the affinity of protein–protein interactions such as the SA score in Chapter 6 can be used.

7.5 Discussion and future perspective

High-throughput technologies for generating large-scale AP-MS data have already become routine in most laboratories, while the accurate detection of protein complexes remains relatively immature. Hence, the development of effective computational approaches for detecting protein complexes is still needed. Although many new methods have been proposed to identify protein complexes from AP-MS data, many challenging problems need to be further investigated.

To date, computational methods for detecting protein complexes from the AP-MS data mostly focus on the high-density cluster, and protein complexes with low-density are always neglected. New clustering algorithms that are capable of finding such kinds of protein complexes should be developed.

During the past years, numerous algorithms have been developed in the literature of graph clustering. Unfortunately, only several types of graph clustering methods have been tested and applied to solve the protein complex detection problem. It would be plausible to investigate the feasibility of using other unexplored graph clustering algorithms so as to obtain more accurate protein complex identification result.

Existing algorithms for protein complex detection either focus on binary bait–prey data or quantitative AP-MS data. There are still no computational methods that can handle both types of data sets in a unified framework.

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.E. Schaeffer, Graph clustering, Comput. Sci. Rev. 1 (1) (2007) 27–64.

[3] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3–5) (2010) 75–174.

[4] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814–818.

[5] M. Wu, X. Li, C.K. Kwoh, et al., Discovery of protein complexes with core-attachment structures from tandem affinity purification (TAP) data, J. Comput. Biol. 19 (9) (2012) 1027–1042.

[6] W. Chu, Z. Ghahramani, R. Krause, D.L. Wild, Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model, Pac. Symp. Biocomput. 11 (2006) 231–242.

[7] J. Handl, J. Knowles, D.B. Kell, Computational cluster validation in post-genomic data analysis, Bioinformatics 21 (15) (2005) 3201–3212.