BIOINFORMATICS

Vol. 27 ISMB 2011, pages i159–i166
doi:10.1093/bioinformatics/btr212

Construction of co-complex score matrix for protein complex prediction from AP-MS data

Zhipeng Xie[1,2], Chee Keong Kwoh[2], Xiao-Li Li[2,3,*] and Min Wu[2]

[1]School of Computer Science, Fudan University, Shanghai 200433, China, [2]School of Computer Engineering, Nanyang Technological University, Singapore 639798 and [3]Institute of Infocomm Research, 1 Fusionopolis Way, Singapore 138632

ABSTRACT

Motivation: Protein complexes are of great importance for unraveling the secrets of cellular organization and function. The AP-MS technique has provided an effective high-throughput screening to directly measure the co-complex relationship among multiple proteins, but its performance suffers from both false positives and false negatives. To computationally predict complexes from AP-MS data, most existing approaches either required the additional knowledge from known complexes (supervised learning), or had numerous parameters to tune.

Method: In this article, we propose a novel unsupervised approach, without relying on the knowledge of existing complexes. Our method probabilistically calculates the affinity between two proteins, where the affinity score is evaluated by a co-complexed score or C2S in brief. In particular, our method measures the log-likelihood ratio of two proteins being co-complexed to being drawn randomly, and we then predict protein complexes by applying hierarchical clustering algorithm on the C2S score matrix.

Results: Compared with existing approaches, our approach is computationally efficient and easy to implement. It has just one parameter to set and its value has little effect on the results. It can be applied to different species as long as the AP-MS data are available. Despite its simplicity, it is competitive or superior in performance over many aspects when compared with the state-of-the-art predictions performed by supervised or unsupervised approaches.

Availability: The predicted complex sets in this article are available in the Supplementary information or by sending email to asckkwoh@ntu.edu.sg

Contact: xlli@i2r.a-star.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A protein complex is a group of two or more associated polypeptide chains. Most proteins participate in cellular processes by interacting with other molecules, often with other proteins in the assembly of operational complexes. To better understand and detect the co-complexed relationship among proteins, the screening technique of affinity purification followed by mass spectrometry (AP-MS) has been designed and applied by two research groups to detect the full yeast interactome (Gavin *et al*., 2006; Krogan *et al*., 2006).

In an AP-MS experiment, a tagged protein is expressed in yeast and then 'pulled down' from a cell extract, along with any proteins associated with it, by co-immunoprecipitation or by tandem affinity purification. The set of pulled down proteins is then identified by MS in a laborious and expensive process. This procedure has been systematically applied to large sets of yeast proteins. The tagged protein in AP-MS is also typically called the bait and the proteins it pulls down are called the preys. AP-MS experiments measure complex co-membership, and the fact that a prey is found by certain bait means that there is either a direct physical interaction or an indirect physical interaction (functional interaction) mediated by a protein complex. Since AP-MS data provide us direct information about the co-complex relationships among proteins, it is thus more useful resources for complex detection compared with pairwise protein interaction data. However, AP-MS screening may not be good enough to detect protein complexes directly, because a single bait protein may be involved in more than one complex in a cell, it may therefore capture a set of prey proteins which actually never occur in the same complex. In addition, it is well known that real purification data are noisy and it contains many false positives and false negatives (Gavin *et al*., 2006).

To deal with these problems, several notable computational approaches have recently been proposed to identify protein complexes from AP-MS data, which typically consist of following three steps. The first step is to assess the protein interaction affinities. Krogan *et al*. (2006) assigned a probability to each pair of proteins by using a stacking algorithm (an advanced supervised learning algorithm from machine learning) based on experimental reproducibility and mass-spectrometry scores, with the hand-curated MIPS complexes as the training set. Gavin *et al*. (2006) described the socio-affinity (SA) scores of comparing the number of co-occurrences of two proteins against the random expectation by using a combination of spoke and matrix model. Collins *et al*. (2007) developed the purification enrichment (PE) scores as a modified version of SA scores, in the probabilistic framework of a naïve Bayes classifier. Hart *et al*. (2007) designed a matrix-model scoring algorithm based on the hypergeometric distribution. Zhang *et al*. (2008) proposed the dice coefficient (DC) to measure interaction affinity between two proteins based on similarity of their co-purification patterns. After the affinity scores are calculated, the second step is to construct a protein–protein interaction (PPI) network by applying a threshold or cutoff value. Hart *et al*. (2007) tried all thresholds to determine the one that yielded the set of predicted complexes with the best performance of balanced accuracy and coverage against the set of manually curated MIPS complexes. Zhang *et al*. (2008) tested a series of thresholds and chose the one that produced the best F_1-measure (the harmonic mean of recall and precision) on the MIPS complexes. Finally, the third step is

*To whom correspondence should be addressed.

to mine complexes on the constructed PPI network. A variety of computational algorithms including MCODE (Bader and Hogue, 2003), Markov clustering (Enright *et al*., 2002) and DPClus (Altaf-Ul-Amin *et al*., 2006) to name a few, are qualified for this job. These algorithms can be characterized into two categories according to the outcomes of complex mining. Collins *et al*. (2007), Pu *et al*. (2007), Hart *et al*. (2007) and Friedel *et al*. (2009) exploits the algorithms in the first category (such as Markov clustering and hierarchical clustering) which output non-overlapping clusters, (normally) followed by an optional step that adds shared proteins to the clusters. On the other hand, Zhang *et al*. (2008), Geva and Sharan (2011) make use of the algorithms in the second category which output highly overlapping clusters based on some graph algorithms such as maximal cliques or maximal bi-cliques. The highly overlapping clusters should be merged to reduce the number of clusters. In addition, Gavin *et al*. (2006) utilized iterative hierarchical clustering approach multiple times, each time with a different set of parameters, on the SA scored network, which is best classified as the second category.

Although the task of predicting protein complexes from AP-MS data has been widely studied, there are still challenging research problems. On the one hand, as indicated in Friedel *et al*. (2009), most existing approaches rely on (more or less) supervised information of known reference complexes. It is desirable and has great value that an approach be independent of such supervised information such that it can be applied to large-scale AP-MS datasets of other organisms without a requirement of a sufficient size of known protein complexes. Friedel *et al*. (2009) proposed an unsupervised algorithm for complex identification based on Bootstrap sampling, whose detailed process is quite computationally expensive. On the other hand, only a few approaches (Friedel *et al*., 2009; Hart *et al*., 2007; Pu *et al*., 2007) have dealt with the issue of integrating multiple AP-MS datasets. It will benefit complex mining greatly if a flexible and adaptive mechanism is provided for the combination of multiple different AP-MS datasets. Hart *et al*. (2007) combined two AP-MS datasets by multiplying *P*-values for the same interaction derived from different datasets. Collins *et al*. (2007) and Pu *et al*. (2007) dealt with problem by weighted summing the PE scores from different datasets. Differently, Friedel *et al*. (2009) went into operation by simply pooling their purification experiments. Further, traditional approaches usually construct a PPI network first, and then apply a clustering algorithm [mostly Markov Clustering algorithm (MCL)] to it. During the process of deriving a PPI network from the original AP-MS data, the useful quantity information that two proteins are unlikely to be co-complexed is discarded. However, such kind of information can play an important role in accurately assembling the proteins into complexes, for it can be exploited to determine when to stop clusters from growing further.

To deal with these problems, this paper first proposed a novel scoring method, called co-complexed score (or C2S score in brief), which represents the log likelihood ratio of a protein pair being co-complexed to being randomly drawn, based on four probabilistic parameters. Each AP-MS dataset has its own estimated probabilistic parameters, which can be estimated solely on the AP-MS dataset itself. Our method then integrated two most comprehensive AP-MS datasets from Gavin *et al*. (2006) and Krogan *et al*. (2006) by score matrix merging. Finally, a hierarchical clustering algorithm is applied directly on this merged score matrix (instead of a PPI network with a cutoff threshold). Our method terminates the clustering process automatically (we will stop the cluster merging process if the current merging step does not improve the quality of clustering), and returns the remaining clusters as the predicted complexes. In the experimental section, we will show that our approach is competitive or superior in performance over many aspects, compared with the state-of-the-art supervised or unsupervised methods.

2 METHODS

Let $E=\{e_1, e_2, \ldots, e_N\}$ be an AP-MS dataset of N purifications. For each purification $e \in E$, we use bait(e) to denote the bait used, and preys(e) the set of pulled down preys. The set of all the baits used in E is then denoted by baits(E)$=\{$bait(e)$|e \in E\}$. Further, preys(E)$=\cup_{e\in E}$preys(e) denotes the set of all the proteins that have been pulled down as preys in at least one purification.

2.1 Definition of C2S scores

Two proteins are co-complexed to each other if both of them are members of the same protein complex. In the ideal situation, when a protein is used as bait in a purification experiment, all its co-complexed proteins will be pulled down as preys. However, real purification data are noisy and it contains many false positives and false negatives for a variety of reasons such as tag interference, low protein abundance. Thus, we regard each purification experiment as a piece of evidence about the co-complex relationship among proteins. All the purifications related to a pair of proteins are accumulated to infer the affinity (or the quantitative measurement of the co-complex relationship) of the protein pair. More specifically, a purification is related to a protein pair $\{x, y\}$ if it can be categorized into one of the following four evidence types:

- Type 1: a purification with bait x (or y) pulls down the protein y (or x);
- Type 2: a purification with bait x (or y) does not pull down the protein y (or x);
- Type 3: a purification whose bait is different from x and y pulls down both x and y at the same time;
- Type 4: a purification whose bait is different from x and y pulls down one and only one of the proteins x and y.

Otherwise, or equivalently, if a purification contains neither x nor y, it is not useful for measuring their affinity and thus we will not take it into consideration.

Qualitatively speaking, it is reasonable that the evidence types 1 and 3 should be a plus (positive effect) to the estimated affinity, while evidence types 2 and 4 be a minus (negative effect). Both types 1 and 3 requires co-occurrences of the two proteins in a single purification: bait–prey co-occurrence for type 1 and prey–prey co-occurrence for type 3. The more co-occurrences of two proteins, the more likely they are to be co-complexed. For a pair of proteins (x, y), these four types of evidences are illustrated in Figure 1, where z is a protein different from x and y.

Next, the problem is: to what extent does each unit of evidence (or a purification) contribute to our belief that proteins x and y are co-compexed? To answer this question, we would like to propose the four probabilistic parameters listed as follows: (How to estimate these parameters for a given AP-MS dataset will be addressed later in Section 2.2)

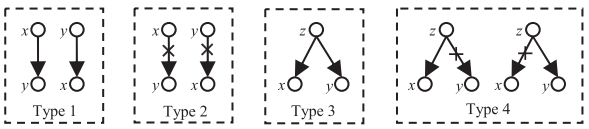

Fig. 1. Four evidence types for a protein pair $\{x, y\}$.

- r_{bp}: the probability that, in a purification, a co-complexed protein of the bait will be detected as a prey.
- ρ_{bp}: the probability that a protein is detected as a prey in a purification experiment for non-specific reason.
- r_{pp}: in a purification where a prey is detected, the probability that a co-complexed protein of the prey will also be detected as another prey in the purification.
- ρ_{pp}: the probability that, in a purification where a prey is pulled down, another protein is detected as prey in the purification for non-specific reason.

Thereupon, a novel measurement called C2S score is defined. Each C2S score is a measurement of the log-likelihood ratio of observed purifications given the hypothesis that the protein pair is co-complexed relative to the likelihood of the same results if the protein pair is drawn randomly. This is similar to the PE score in spirit. However, the proposed C2S score is totally probabilistic in nature, and it is fundamentally different from the existing scoring methods such as SA or PE score in that C2S takes into consideration both positive evidences and negative evidences equally, while SA and PE score concentrate more on positive ones, e.g. PE scores have overlooked the evidence type 4.

For a given pair of proteins $\{x,\ y\}$ in E, its C2S score is calculated by

$$\begin{aligned} C2S^E(x,y) &= \log \frac{P(E|x \text{ and } y \text{ are co-complexed})}{P(E|x \text{ and } y \text{ are randomly related})} \\ &= \log \prod_{e\in E} \frac{P(e|x \text{ and } y \text{ are co-complexed})}{P(e|x \text{ and } y \text{ are randomly related})} \\ &= \sum_{i=1}^{4} \text{para}_i \times \text{count}_{Ti}^E(x,y) \end{aligned} \tag{1}$$

where $\text{count}_{Ti}^E(x,y)$ denotes the number of purifications in E that are of evidence type i with respect to the protein pair $\{x,\ y\}$, while para_i, representing the evidence that each purification of type i contribute to the C2S score of the protein pair, is defined as

$$\text{para}_i = \log \frac{P(\text{an evidence unit of type } i|x \text{ and } y \text{ are co-complexed})}{P(\text{an evidence unit of type } i|x \text{ and } y\text{are randomly related})}.$$

and is in turn calculated as:

$$\text{para}_i = \begin{cases} \log \frac{r_{bp}}{\rho_{bp}}, & i=1; \\ \log \frac{1-r_{bp}}{1-\rho_{bp}}, & i=2; \\ \log \frac{r_{pp}}{\rho_{pp}}, & i=3; \\ \log \frac{1-r_{pp}}{1-\rho_{pp}}, & i=4. \end{cases} \tag{2}$$

However, if any protein in the pair does not occur (as bait or prey) in the AP-MS data, their C2S score will be 0. Finally, we use the following example to illustrate how to calculate the value of $\text{count}_{Ti}^E(x,y)$ from an AP-MS dataset.

EXAMPLE 1. *Let $E=\{e_1,e_2,e_3,\ldots,e_8\}$ denote a dataset of eight purification experiments, which are illustrated as follows. Take e_1 as an example: it has the bait p_1 that pulls down four preys $\{p_2,p_3,p_6,p_{10}\}$.*

$$\begin{array}{lll} e_1: p_1 \to p_2,p_3,p_6,p_{10} & e_2: p_1 \to p_3,p_4,p_6 & e_3: p_{10} \to p_3,p_6 \\ e_4: p_4 \to p_5,p_7 & e_5: p_8 \to p_7,p_4 & e_6: p_9 \to p_1,p_3,p_{10} \\ e_7: p_5 \to p_1,p_4,p_7 & e_8: p_3 \to p_1,p_2,p_9 & \end{array}$$

For a pair of proteins $\{p_3,p_4\}$, we have $\text{count}_{T1}^E(p_3,p_4)=0$ and $\text{count}_{T2}^E(p_3,p_4)=|\{e_4,e_8\}|=2$ because in the only two purifications (e_4 and e_8) whose baits are from $\{p_3,\ p_4\}$, the other protein is not pulled down; $\text{count}_{T3}^E(p_3,\ p_4)=|\{e_2\}|=1$ and $\text{count}_{T4}^E(p_3,\ p_4)=|\{e_1,e_3,e_5,e_6,e_7\}|=5$ because p_3 and p_4 are both detected as preys in e_2, while one and only one of $\{p_3,p_4\}$ is detected as a prey in e_1,e_3,e_5,e_6,e_7.

2.2 Parameter estimation

So far, we have defined the C2S score based on the four probabilistic parameters. In this section, we describe how to estimate them from AP-MS data. For each protein x, we use the notation $PulledBy(x)=\{y|\exists e\in E$ such that $(\text{bait}(e)=y)\wedge(x\in\text{preys}(e))\}$ to denote the set of all the baits that pulled down x, and notation $PulledDown(x)=\cup_{e:\text{bait}(e)=x}\text{preys}(e)$ to denote the set of all the preys that have been pulled down at least once in the purifications with bait x. For example, we have $PulledBy(p_3)=\{p_1,p_9,p_{10}\}$, $PullDown(p_3)=\{p_1,p_2,p_9\}$, $PullDown(p_1)=\{p_2,p_3,p_4,p_6,p_{10}\}$ and $PulledBy(p_1)=\{p_3,p_9\}$ in the data illustrated in Example 1.

It is relatively straightforward to estimate probabilistic parameters ρ_{bp} and ρ_{pp} for a given AP/MS dataset E:

$$\rho_{bp}=\frac{\sum_{e\in E}|\text{preys}(e)|}{\sum_{e\in E}(n-1)}=\frac{\sum_{e\in E}|\text{preys}(e)|}{|E|\times(n-1)}, \tag{3}$$

and

$$\rho_{pp}=\frac{\sum_{x\in\text{preys}(E)}\sum_{e\in E:x\in\text{preys}(e)}(|\text{preys}(e)|-1)}{\sum_{x\in\text{preys}(E)}\sum_{e\in E:x\in\text{preys}(e)}(n-2)}, \tag{4}$$

where n is the number of all distinct proteins that appear in the dataset.

As to the other two parameters r_{bp} and r_{pp}, the technique for unsupervised estimation is a little more complicated. The key point here is to construct an approximate set of co-complexed proteins for each bait protein directly from AP-MS data. For estimating r_{bp}, we use the back-link set as the approximate set of co-complexed proteins, while for r_{pp}, we use the reciprocal set. Their definitions are given in the following description.

For each protein $x\in\text{baits}(E)$, the notation $backlink(x)=PulledBy(x)\cap\text{preys}(E)$ denotes the back-link set of x which consists of all the baits that pulled down x and were also detected as preys in other purifications in E. Each protein in $backlink(x)$ is called a back-link protein of x. In Example 1, we have $PulledBy(p_4)=\{p_1,p_5,p_8\}$, but p_8 has never been detected as a prey in all the eight purifications, so $backlink(p_4)=\{p_1,p_5\}$. Let $e\in E$ be a purification of bait x, the number of proteins in $backlink(x)$ that are detected as preys in e can be expressed as $|backlink(x)\cap\text{preys}(e)|$. In the average, the probability that a back-link protein of x will be detected as a prey in a purification with bait x is estimated as

$$\frac{\sum_{e\in E}|backlink(\text{bait}(e))\cap\text{preys}(e)|}{\sum_{e\in E}|backlink(\text{bait}(e))|}.$$

Let tpr be the true positive rate (or the probability that a protein being pulled down by a bait is co-complexed with the bait) with 0.6 as the default value. The sensitivity analysis of the values of tpr will be illustrated in the experimental section. Thus, the parameter r_{bp} is estimated approximately by averaging the percentages over all experiments as follows:

$$r_{bp}=\frac{\sum_{e\in E}|backlink(\text{bait}(e))\cap\text{preys}(e)|}{\sum_{e\in E}|backlink(\text{bait}(e))|}\times\frac{1}{tpr}. \tag{5}$$

For each protein $x\in$ baits(E), we also define $\text{reciprocal}(x)=PulledBy(x)\cap PullDown(x)$ to be the reciprocal set of proteins that not only are pulled down by x, but also pull down x. Each protein in reciprocal(x) is called a reciprocal protein of x. As in Example 1, we have $PulledBy(p_3)=\{p_1,p_9,p_{10}\}$ and $PullDown(p_3)=\{p_1,p_2,p_9\}$, and thus we have $\text{reciprocal}(p_3)=PulledBy(p_3)\cap\ PullDown(p_3)=\{p_1,p_9\}$. Then, if a protein x is detected as a prey in a purification e, the number of the reciprocal proteins of x that are also detected in e can be expressed as $|\text{reciprocal}(x)\cap\text{preys}(e)|$, and the number of all its reciprocal proteins that are possible to be detected as preys in e is $|\text{reciprocal}(x)-\{\text{bait}(e)\}|$. In the average, the probability that a reciprocal protein of x will be co-purified with x as preys in a purification is estimated as

$$\frac{\sum_{x\in\text{baits}(E)}\sum_{e\in E:x\in\text{preys}(e)}|\text{reciprocal}(x)\cap\text{preys}(e)|}{\sum_{x\in\text{baits}(E)}\sum_{e\in E:x\in\text{preys}(e)}|\text{reciprocal}(x)-\{\text{bait}(e)\}|}.$$

Furthermore, the probability that a reciprocal protein of x is also a co-complexed protein of x can be estimated as $\sigma=1-(1-tpr)^2$. Therefore, the probabilistic parameter r_{pp} can be estimated approximately as:

$$r_{pp}=\frac{\sum_{x\in\text{baits}(E)}\sum_{e\in E:x\in\text{preys}(e)}|\text{reciprocal}(x)\cap\text{preys}(e)|}{\sum_{x\in\text{baits}(E)}\sum_{e\in E:x\in\text{preys}(e)}|\text{reciprocal}(x)-\{\text{bait}(e)\}|}\times\frac{1}{\sigma}. \tag{6}$$

2.3 Score matrix merging

In Sections 2.1 and 2.2, we have discussed how to compute the C2S score given an AP-MS dataset. We are now ready to extend it to multiple AP-MS datasets. In particular, let $\{E_i | 1 \le i \le L\}$ be a set of AP-MS datasets, the score matrices can be merged as follows:

$$C2S(x,y) = \sum_{i=1}^{L} w_i \times C2S^{E_i}(x,y), \quad (7)$$

where w_i denotes the weight of the i-th dataset.

In this article, we treat Gavin and Krogan data equally, so each dataset is assigned with the unit weight. While for Krogan's dataset, we divided it into two subsets: Krogan-LC and Krogan-MALDI according to the mass spectrometry techniques used, which are also treated equally—each subset is assigned the weight of 0.5 so that their total weight is equal to 1.

2.4 Hierarchical clustering for complex mining

Once a C2S score matrix has been constructed, a simple hierarchical clustering algorithm can be applied directly to cluster the set of proteins into multiple clusters. It starts from the set of all singleton protein clusters. It then merges the two clusters with the highest similarity at each iteration, where the definition about the similarity between two clusters is given below. The detailed procedure is illustrated in Algorithm 1.

Note that a cluster is a subset of proteins collective. For any two clusters c_i and c_j, their similarity is defined as the C2S score value averaged over all protein pairs between the clusters, that is:

$$\mathrm{sim}(c_i, c_j) = \frac{1}{|c_i| \times |c_j|} \sum_{x \in c_i, y \in c_j} C2S(x,y). \quad (8)$$

For any two singleton clusters $c_i = \{x\}$ and $c_j = \{y\}$, it is evident that $\mathrm{sim}(c_i, c_j) = C2S(x,y)$.

In our algorithm, this merging process will terminate automatically only if it cannot find any pair of clusters with positive similarity, or in other words, if the similarity between any two nearest clusters is less than or equal to zero. It is attributed to the fact that evidence types (2) and (4) has introduced a lot of negative values into the C2S score matrix, which prevents any further cluster merging. This automated stop process is theoretically sound and better than the existing methods such as Gavin *et al*. (2006) where a merging threshold has to be manually set by a user.

Once terminated, the set of remaining clusters will be outputted as the set of predicted complexes, which are sorted in the descending order of their confidence scores expressed by the averaged co-complex score over all pairs within a predicted complex:

$$\mathrm{confidence}(c) = \frac{\sum_{x \in c, y \in c} C2S(x,y)}{|c| \times (|c| - 1)} \text{ for any cluster } c. \quad (9)$$

3 RESULTS

In this section, we first elaborate the AP-MS datasets we used and list the complex sets predicted by current state-of-the-arts (for comparison). Then, we compare our predicted complex sets on the integrated Gavin and Krogan's data with those existing ones in terms of widely used evaluation metrics such as accuracy, recall, co-localization and functional co-annotation. Finally, the result of applying our method on Gavin's data alone is also presented.

3.1 The AP-MS datasets used and the predicted complex sets compared

Gavin *et al*. (2006) used only one mass spectrometry method (MALDI-TOF) to identify proteins co-purified with a bait, while Krogan *et al*. (2006) used two separate methods [MALDI-TOF and Liquid chromatography-mass spectrometry (LC-MS)/MS]. Therefore, the Krogan's dataset can be divided into two subsets Krogan-MALDI and Krogan-LC, each corresponding to the mass spectrometry method used. In addition, confidence scores were assigned for protein identification by mass spectrometry, so we prune Krogan's dataset with the cut-off thresholds (99.6 for LC-MS/MS protein identification, and 3.4 for MALDI-TOF protein identification) used by Hart *et al*. (2007) and thus yield a more reliable dataset (called 'Krogan-HighConf') from Krogan's raw data. The number of purifications, the number of distinct baits and the number of distinct preys for each dataset are summarized in Table 1.

By applying our unsupervised method, two predicted complex sets are generated:

- C2S: on Gavin dataset, Krogan-MALDI data subset and Krogan-LC data subset
- C2S-HighConf: on Gavin dataset, Krogan-HighConf-MALDI data subset and Krogan-HighConf-LC data subset.

Algorithm 1. Hierarchical clustering on the merged C2S score matrix

Step 1 (Initialization): Initialize the cluster set $C = \{\{x\} | x$ is a protein appearing in the AP-MS datasets$\}$ by creating a cluster $\{x\}$ for each single protein x. The similarity matrix is initialized according to the C2S score metric, that is, $\mathrm{sim}(\{x\}, \{y\}) = C2S(x,y)$.

Step 2 (Identification of the two most similar clusters): the similarity matrix is scanned and the highest value $\mathrm{sim}_{\mathrm{best}}$ is identified, with corresponding pair of clusters denoted as (c_i, c_j).

Step 3 (Termination or not?): If $\mathrm{sim}_{\mathrm{best}}$ is larger than 0, go to step 4; otherwise, return the current set C of clusters (with at least two proteins) as the set of predicted complexes.

Step 4 (Merging clusters and update the similarities): A new cluster c_{new} is created by merging the two clusters c_i and c_j. The cluster c_{new} is then added into C, while c_i and c_j are removed. Then, for each $c_k \in C$ (different from c_{new}), calculate its similarity with the new cluster c_{new} as follows:

$$\mathrm{sim}(c_{\mathrm{new}}, c_k) = \frac{|c_i| \times |c_k| \times \mathrm{sim}(c_i, c_k) + |c_j| \times |c_k| \times \mathrm{sim}(c_j, c_k)}{(|c_i| + |c_j|) \times |c_k|}$$

Step 5 (Loop): Go to Step 2

Table 1. Details of AP-MS datasets

AP-MS datasets	#Purifications	#Baits	#Preys
Gavin	2166	1993	2671
Krogan	4332	2294	5333
Krogan-HighConf	3575	2143	2567

Table 2. Statistics of predicted complex sets to be compared

Predicted complex set	#Complexes	Avg. complex size	#distinct proteins
C2S	1039	4.93	5121
C2S-HighConf	679	3.79	2571
Hart	390	4.33	1689
Pu	400	5.14	1913
BT-893	893	6.25	5187

The existing complex sets predicted by combining Gavin and Krogan data for comparison include:

- Hart: the predicted complex set from Hart *et al*. (2007), using a supervised method;
- Pu: the predicted complex set from Pu *et al*. (2007), using a supervised method; and
- BT-893: the predicted complex set from Friedel *et al*. (2009), using an unsupervised Bootstrap method.

The statistics of these predicted complex sets to be compared are summarized in Table 2.

3.2 Comparative evaluation on reference complexes

To evaluate the accuracy of the predicted complex set, sensitivity (Sn) and positive predictive value (PPV) were calculated with regard to the following two widely used benchmark complex reference sets:

- CYC2008 contains 408 manually curate complexes (Pu *et al*., 2008); and
- MIPS contains 214 manually curated complexes from the MIPS database (Mewes *et al*., 2004).

The sensitivity and the PPV between a reference complex r_i and a predicted complex c_j are calculated from the number $T_{i,j}$ of proteins shared between them (Brohee and Helden, 2006):

$$Sn = \frac{\sum_i \max_j T_{i,j}}{\sum_i |r_i|}, \text{ and}$$

$$\text{PPV} = \frac{\sum_j \max_i T_{i,j}}{\sum_j |\cup_i (r_i \cap c_j)|} = \frac{\sum_j \max_i T_{i,j}}{\sum_j |(\cup_i r_i) \cap c_j|}.$$

Please note that the definition of PPV here is a little bit different from the original definition given by Brohee and Helden (2006). We propose this new definition is due to the fact that the original definition cannot evaluate overlapping clusters properly as reported by Li *et al*. (2010). For example: if the known gold standard MIPS complex set is taken to match with itself, then the resulting PPV value is 0.77 instead of 1; when CYC2008 is taken to match with itself, the resulting PPV is only 0.68. The new proposed PPV can address this issue and yield PPV value of 1 when a complex set is matched with itself. In addition, it is worth mentioning that our definition is equivalent to the original definition when all the reference complexes do not overlap with each other.

As a summary metric, the accuracy of a prediction (*Acc*) can then be defined as the geometric average of sensitivity and PPV,

$$\text{Acc} = \sqrt{Sn \times \text{PPV}}$$

Table 3. Sensitivities (Sn), PPVs and Acc compared on reference complexes

Predicted complex set	CYC2008 (408)			MIPS (214)		
	Sn	PPV	Acc	Sn	PPV	Acc
C2S	0.680	0.837	0.755	0.582	0.821	0.692
C2S-HighConf	0.643	0.889	0.756	0.5294	0.889	0.686
Pu (400)	0.691	0.789	0.738	0.593	0.795	0.686
Hart (390)	0.610	0.863	0.725	0.514	0.846	0.660
BT-893	0.720	0.759	0.740	0.582	0.773	0.671

From Table 3, we observed that C2S and S2S-HighConf achieved the best accuracy compared with the state-of-the-arts which used the benchmark complexes to select reliable protein interactions. Our proposed unsupervised method has an advantage to be directly applied to mine protein complexes from other less well-studied species, even without the known protein complexes existing.

When evaluating the predicted complex set over a reference set, other commonly used evaluation metrics include precision, recall and *F*-measure. Let r be a reference complex, c be a predicted complex. The matching degree between r and c is used to measure how well they match with each other (Bader and Hogue, 2003):

$$\text{MD}(r,c) = \frac{|r \cap c|^2}{|r| \times |c|}.$$

Given a threshold ω, we say that r and c match each other if $\text{MD}(r, c) \geq \omega$. Let M_{ref} be the number of reference complexes that match at least one predicted complex, and M_{pre} be the number of predicted complexes that match at least one reference complex. Precision, recall and *F*-measure are then defined as (Chua *et al*., 2008):

$$\text{Precision} = \frac{M_{\text{pre}}}{|C|}, \qquad \text{Recall} = \frac{M_{\text{ref}}}{|R|},$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

However, due to the fact that the reference complex set CYC2008 is far from complete, an unmatched predicted complex may be also a true complex, so we do not think precision is a good way to measure the quality of a predicted complex set. In addition, we have the following Disjoint property, which is easy to prove.

[Disjoint Property]: For a given predicted complex set C and a reference complex set R, if the predicted complexes in C and the complexes in R are disjoint, respectively, then each reference complex in R matches at most one predicted complex in C, and each predicted complex in C matches at most one reference complex in R when ω is larger than 0.5. In this situation, it is evident that $|M_{\text{ref}}| = |M_{\text{pre}}|$.

Proof. See the Supplementary Material for detailed information.

Since all the compared approaches here rely on a partitioning-based clustering method, the predicted complexes can be considered as (approximately) disjoint. Therefore, we only compare recall values for these predicted complex sets, with different values of ω.

It can be seen from Figure 2 that C2S and C2S-HighConf have similar recall with the supervised methods (Hart and Pu), and

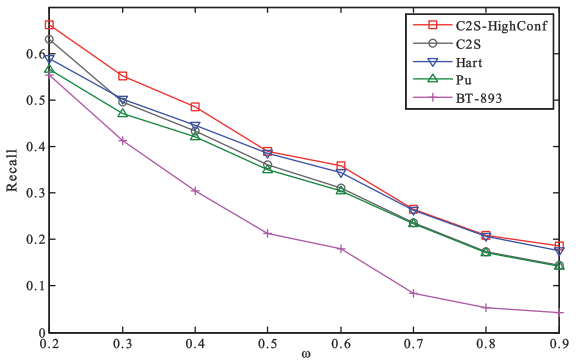

Fig. 2. The comparison of recall with different values of ω.

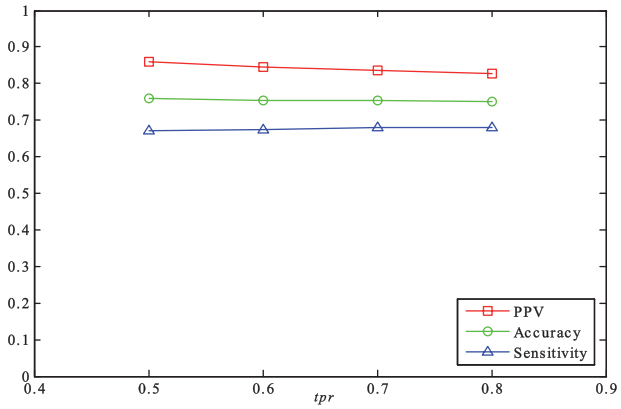

Fig. 3. The effect of varying *tpr* on sensitivity, PPV and accuracy.

they evidently outperform the unsupervised method (BT-893) which drops faster with the increased value of ω.

3.3 The effect of varying tpr

Recall that in the parameter estimation, we have to set the value of *tpr*, which stands for the true positive rate. Here, we investigate how the variation of *tpr* affects the performance of our approach. Figure 3 shows the sensitivity, PPV and accuracy on the combined Gavin and Krogan data under different values of *tpr*.

As we change the value of *tpr*, the performance of our approach remains quite stable. The accuracy drops slightly from 0.76 to 0.75 when the value of *tpr* is increased from 0.5 to 0.8, which is consistently better than all the other approaches compared in Table 2.

3.4 Comparative evaluation on co-localization and functional co-annotation within complexes

Since a protein complex is usually assembled by the proteins with same/similar localization to carry out a specific function, the similarity of co-localization and that of functional co-annotation among proteins in the same complex provide indirect evidences about the quality of the predicted complexes.

As to the similarity of functional co-annotation, we use the relevance similarity described by Schlicker *et al.* (2006) based on the protein annotations of the Gene Ontolgoy (GO) (Ashburner *et al.*, 2000). As in Friedel *et al.* (2009), the GO score of a complex set is the weighted mean over all complex scores, and in turn, the score of a complex is the average relevance similarity of all protein pairs in the complex. The GO scores are calculated for the 'biological process (GO-BP)' and 'molecular function (GO-MF)' ontologies separately. As to the similarity of co-localization, we use the protein localizations derived by Huh *et al.* (2003). The co-localization score (COLOC) for a complex is defined as the maximum fraction of proteins in this complex which share the same localization, and it is the weighted average over all complexes for a predicted complex set.

Table 4. Comparison of co-localization and functional co-annotation within complexes

Predicted complex set	COLOC (%)	GO-BP (%)	GO-MF (%)
C2S-HighConf-405	89.2	85.9	80.3
BT-409	89.1	86.5	79.3
Pu	84.6	85.8	77.7
Hart	88.1	87.5	78.0

To make a fair evaluation on co-localization and functional co-annotation within complexes, we extract the first 405 predicted complexes from C2S-HighConf ranked by their confidences measured by Equation (9), and denote the complex set as 'C2S-HighConf-405' which covers 1725 distinct proteins. This is because we want all the compared complex sets to be of comparable size. The compared complex sets by Pu and Hart contain 400 and 390 complexes, respectively, and the BT method used 408 complexes (denoted as BT-408) which have good performance measured on co-localization information.

From Table 4, we observe that C2S-HighConf-405 has similar (slightly higher) quality to Hart and BT-409 on the three scores, and they have substantially higher scores than Pu.

3.5 Evaluating predicted complex set from the Gavin data alone

We also applied our C2S approach to the Gavin dataset alone, with the predicted complex set (called C2S-Gavin) of 474 complexes covering 1942 distinct proteins. The previous predictions used for comparison include the Bootstrap predictions (Friedel *et al.*, 2009), the predictions based on the DCs (Zhang *et al.*, 2008), the complete Gavin complexes (Gavin *et al.*, 2006), the CODEC-w0 and the CODEC-w1 (Geva and Sharan, 2011).

The sensitivities, PPVs and accuracies of these predicted complex sets against the two reference complex sets are summarized in Table 5. It can be seen that C2S-Gavin has the highest accuracy among all these predictions. Its PPV value is only slightly lower than Gavin-Core, but the sensitivity is much higher.

We also measured the Recall with different values of ω. From Figure 4, it is evident that our approach remains the best among the compared predictions when ω varies from 0.2 to 0.9.

From Table 6, it can be seen that the C2S-Gavin has achieved substantially higher co-localization and functional co-annotations than those previous predictions. Furthermore, to make comparison against the high confidence Gavin-Core predictions, 231 predicted complexes with the highest scores (denoted as 'C2S-Gavin-231') are extracted from C2S-Gavin. The complex set C2S-Gavin-231 covers

Table 5. Comparisons of Sensitivities (Sn), PPVs and Acc of predictions on Gavin data alone against two reference complex sets

Predicted complex set	CYC2008 (408)			MIPS (214)		
	Sn	PPV	Acc	Sn	PPV	Acc
C2S-Gavin (474)	0.588	0.884	0.721	0.500	0.895	0.669
BT-Gavin (381)	0.631	0.756	0.691	0.547	0.774	0.650
Zhang-Gavin (851)	0.607	0.679	0.642	0.547	0.699	0.618
Gavin-Core (478)	0.392	0.914	0.598	0.350	0.907	0.564
Gavin-All (491)	0.570	0.552	0.561	0.517	0.605	0.559
CODEC-w0-Gavin (1082)	0.552	0.506	0.528	0.486	0.535	0.510
CODEC-w1-Gavin (1005)	0.549	0.542	0.546	0.484	0.600	0.539

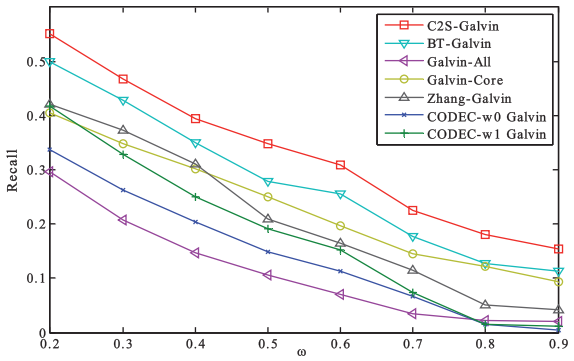

Fig. 4. Comparison of recall by varying ω for predictions on Gavin data alone.

Table 6. Comparison of co-localization, functional co-annotation for predictions on Gavin data alone

Predicted complex set	COLOC (%)	GO-BP (%)	GO-MF (%)
C2S-Gavin	85.29	81.64	76.05
BT-Gavin	78.93	78.35	71.92
Zhang-Gavin	75.51	75.64	70.92
Gavin-All	70.27	74.36	68.19
CODEC-w0-Gavin	68.83	69.75	66.15
CODEC-w1-Gavin	76.78	78.71	72.87

1128 distinct proteins, which is identical with Gavin-Core. Even though C2S-Gavin-231 contain fewer complexes than Gavin-Core of 478 complexes, it gets the co-localization score of 90.8%, the 'GO-Biological Process' co-annotation score of 91.3% and the 'GO-Molecular Function' co-annotation score of 81.8%, which is 2.3, 10.0, 5.3% higher than Gavin-Core, respectively, indicating C2S-Gavin can predict protein complexes significantly better than the existing techniques.

We also observed that the results obtained by using Gavin data alone are around 2–3% lower than those using integrated Gavin and Krogan's data, indicating that our proposed method is able to effectively incorporate multiple biological evidences and achieve better results.

3.6 Running time

We implemented the C2S algorithm using Java programming language under the Eclipse framework. On a Lenovo X200 laptop with Intel core 2 duo P8400 (2.26 GHz) and 2 G memory, it takes about 4 min to run the C2S algorithm on the integrated Gavin and the Krogan's raw data.

4 CONCLUSIONS AND DISCUSSIONS

In order to predict protein complexes from AP-MS data, we have proposed a novel scoring method (C2S scores) for evaluation of the log-likelihood ratio of a protein pair being co-complexed based on four probabilistic parameters that are learned solely on the AP-MS dataset. Multiple AP-MS datasets can then be integrated by merging their corresponding C2S score matrices. On the merged C2S score matrix, we developed a hierarchical clustering algorithm which is capable of terminating the clustering process automatically and the final clusters will be treated as the predicted protein complexes. Experimental comparisons have shown that our approach is better than or competitive to other existing ones in many aspects. Furthermore, compared with existing approaches, our approach has the following advantages:

- It is easy to implement and runs efficiently. Even on the combined dataset of Gavin and Krogan, our program only takes 4 min to finish on a laptop.
- The hierarchical clustering process can be easily visualized and easy to understand. It could provide more biological insights for the complex formation. In addition, it can terminate automatically without the requirement of a user predefined threshold.
- There is only one parameter in our algorithm and our results are not sensitive to its value.
- It is unsupervised, and do not require the knowledge of existing complexes. As such, we can be directly applied to other newly generated AP-MS data in yeast or other species.
- All the four probabilistic parameters can be estimated solely from the AP-MS dataset itself.

In future work, we plan to propose more accurate methods for co-complex score measurement, and new mechanisms for data integration:

- More accurate scoring methods: the scoring method in this article is actually an average of measurement across all the purifications in a given AP-MS dataset. However, different purifications should have their own characteristics such as different precisions or recalls. It will lead to a more accurate scoring method by taking these factors into consideration, and further lead to a predicted complex set with higher quality.
- New mechanisms for data integration: here, multiple AP-MS datasets get integrated by first calculating the core matrices for individual datasets, and then merging these individual matrices into a single consolidated matrix, which can be called the score-matrix-level integration. However, there are several other locations in the Figure 5 where the integration of multiple AP-MS dataset may be included. The first possibility is to merge all the purifications from multiple dataset together, which is called the raw dataset level integration. And another option is to

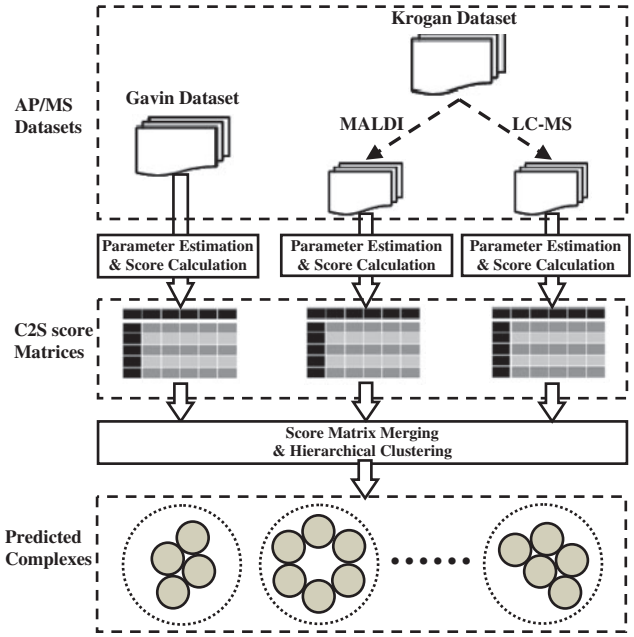

Fig. 5. The flowchart of protein complex prediction based on C2S score matrix.

integrate the result sets mined from individual score matrices, following by a post-processing step to merge the redundant ones. We will leave them as our future work.

Furthermore, we plan to design a post-processing phase, similar to that of Pu *et al*. (2007) and that of Friedel *et al*. (2010), such that proteins can be contained in more than one complex.

ACKNOWLEDGEMENTS

We are grateful to our fellow researchers who have graciously shared with us the prediction results or source codes of their systems, which greatly facilitate our research work. We also thank the anonymous reviewers for their valuable suggestions.

Funding: Singapore MOE AcRF (grant no. MOE2008-T2-1-074).

Conflict of Interest: none declared.

REFERENCES

Altaf-Ul-Amin,M. *et al.* (2006) Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, **7**, 207.

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Brohee,S. and van Helden,J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **7**, 488.

Chua,H.N. *et al.* (2008) Using indirect protein-protein interactions for protein complex prediction. *J. Bioinform. Comput. Biol.*, **6**, 435–466.

Collins,S.R. *et al.* (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell. Proteomics*, **6**, 439–450.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584

Friedel,C.C. *et al.* (2009) Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. *J. Comput. Biol.*, **16**, 1-17

Gavin, A.-C. *et al*. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636

Geva,G. and Sharan,R. (2011) Identification of protein complexes from co-immunoprecipitation data. *Bioinformatics*, **27**, 111–117

Hart,G.T. *et al.* (2007) A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, **8**, 236.

Huh,W.-K. *et al*. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

Krogan,N.J. *et al*. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643

Kumar,A. *et al*. (2002) Subcellular localization of the yeast proteome. *Genes Dev.*, **16**, 707–719.

Li,X. *et al.* (2010) Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics*, **11** (Suppl. 1), S3.

Mewes,H.W. *et al*. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*., **32**, D41–D44.

Pu,S. *et al.* (2007) Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics*, **7**, 944–960.

Pu,S. *et al.* (2008) Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.*, **37**, 825–831.

Schlicker,A. *et al.* (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, **7**, 302.

Zhang,B. *et al.* (2008) From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics*, **24**, 979–986.