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

Zhipeng Xie\textsuperscript{1,2}, Chee Keong Kwoh\textsuperscript{2}, Xiao-Li Li\textsuperscript{2,3,*} and Min Wu\textsuperscript{2}

\textsuperscript{1}School of Computer Science, Fudan University, Shanghai 200433, China, \textsuperscript{2}School of Computer Engineering, Nanyang Technological University, Singapore 639798 and \textsuperscript{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 xllili@i2r.a-star.edu.sg

Contact: xllili@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 naive 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 F1-measure (the harmonic mean of recall and precision) on the MIPS complexes. Finally, the third step is...
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 information quantity 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 \( E = \{e_1, e_2, \ldots, e_k\} \) be an AP-MS dataset of \( N \) purifications. For each purification \( e \in E \), we use \( \text{bait}(e) \) to denote the bait used, and \( \text{prey}(e) \) the set of pulled down preys. The set of all the baits used in \( E \) is then denoted by \( \text{bait}(E) = \{\text{bait}(e) \mid e \in E\} \). Further, \( \text{prey}(e) = \{\text{prey}(e) \mid e \in E\} \) 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-complexed? 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)

\[ P_{12} = \frac{1}{2} (1 + \delta(x,y)) \]
\[ P_{12} = \frac{1}{2} (1 - \delta(x,y)) \]
\[ P_{23} = \frac{1}{2} \delta(x,y) \]
\[ P_{43} = \frac{1}{2} (1 - \delta(x,y)) \]

Fig. 1. Four evidence types for a protein pair \([x, y]\).
• $\rho_p$: the probability that, in a purification, a co-complexed protein of the bait will be detected as a prey.

• $\rho_p$: the probability that a protein is detected as a prey in a purification experiment for non-specific reason.

• $C_{2S}$: 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 measure 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 experiments, which are illustrated as follows. Take e

| $e_8$ | whose baits are from $\{p_1, p_2, p_3\}$, and $E = \{e_8\}$ is detected as a prey in E.

\[
\rho_p = \frac{\sum_{|y| \times \text{Preys}(e)} \text{Preys}(y)}{\sum_{|\backslash y| \times \text{Preys}(e)} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}
\]

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 $\text{PulledBy}(x) = \{y \in E \mid \text{such that } \rho_p(x, y), \rho_p(y, x)\}$ to denote the set of all the baits that pulled down $x$, and notation $\text{PulledDown}(x) = \{\text{Preys}(e) \mid \rho_p(x, y)\}$ to denote the set of all the prey that have been pulled down at least once in the purifications with bait $x$. For example, we have $\text{PulledBy}(p_1) = \{p_2, p_3, p_4\}$, $\text{PulledDown}(p_1) = \{p_2, p_3, p_4\}$, $\text{PulledBy}(p_2) = \{p_1, p_3\}$, and $\text{PulledBy}(p_3) = \{p_1, p_4\}$ in the data illustrated in Example 1.

It is relatively straightforward to estimate probabilistic parameters $\rho_p$ and $\rho_p$ for a given AP-MS dataset E.

\[
\rho_p = \frac{\sum_{|\text{Preys}(e)} \text{Preys}(y)}{\sum_{|\text{Preys}(e)} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Preys}(e)} \text{Preys}(y)}{\sum_{|\text{Preys}(e)} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Preys}(e)} \text{Preys}(y)}{\sum_{|\text{Preys}(e)} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Preys}(e)} \text{Preys}(y)}{\sum_{|\text{Preys}(e)} \text{Preys}(y)}
\]

Each protein in $\text{PullDown}(x)$ is called a back-link protein of $x$. In Example 1, we have $\text{PulledBy}(p_1) = \{p_2, p_3, p_4\}$, but $p_3$ has never been detected as a prey in all the eight purifications, so $\text{Backlink}(p_3) = \{p_1, p_2\}$. Let $e$ be a purification of bait $x$, the number of proteins in $\text{Backlink}(x)$ that are detected as preys in $e$ can be estimated as $\rho_p(x, y)$. The sensitivity analysis of the values of $\rho_p$ will be illustrated in the experimental section. Thus, the parameter $\rho_p$ is estimated approximately by averaging the percentages over all experiments as follows:

\[
\rho_p = \frac{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}
\]

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 $\text{PulledBy}(x) = \{y \in E \mid \text{such that } \rho_p(x, y), \rho_p(y, x)\}$ to denote the set of all the baits that pulled down $x$, and notation $\text{PulledDown}(x) = \{\text{Preys}(e) \mid \rho_p(x, y)\}$ to denote the set of all the prey that have been pulled down at least once in the purifications with bait $x$. For example, we have $\text{PulledBy}(p_1) = \{p_2, p_3, p_4\}$, $\text{PulledDown}(p_1) = \{p_2, p_3, p_4\}$, $\text{PulledBy}(p_2) = \{p_1, p_3\}$, and $\text{PulledBy}(p_3) = \{p_1, p_4\}$ in the data illustrated in Example 1.

It is relatively straightforward to estimate probabilistic parameters $\rho_p$ and $\rho_p$ for a given AP-MS dataset $E$:

\[
\rho_p = \frac{\sum_{|\text{Preys}(e)} \text{Preys}(y)}{\sum_{|\text{Preys}(e)} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Preys}(e)} \text{Preys}(y)}{\sum_{|\text{Preys}(e)} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Preys}(e)} \text{Preys}(y)}{\sum_{|\text{Preys}(e)} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Preys}(e)} \text{Preys}(y)}{\sum_{|\text{Preys}(e)} \text{Preys}(y)}
\]

Each protein in $\text{PullDown}(x)$ is called a back-link protein of $x$. In Example 1, we have $\text{PulledBy}(p_1) = \{p_2, p_3, p_4\}$, but $p_3$ has never been detected as a prey in all the eight purifications, so $\text{Backlink}(p_3) = \{p_1, p_2\}$. Let $e$ be a purification of bait $x$, the number of proteins in $\text{Backlink}(x)$ that are detected as preys in $e$ can be estimated as $\rho_p(x, y)$. The sensitivity analysis of the values of $\rho_p$ will be illustrated in the experimental section. Thus, the parameter $\rho_p$ is estimated approximately by averaging the percentages over all experiments as follows:

\[
\rho_p = \frac{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}
\]

\[
\rho_p = \frac{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}{\sum_{|\text{Backlink}(\text{Preys}(e))} \text{Preys}(y)}
\]
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 \( \{E_i|1 \leq i \leq L\} \) be a set of AP-MS datasets, the score matrices can be merged as follows:

\[
C2S(x,y) = \sum_{i=1}^L w_i \times C2S^i(x,y),
\]

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:

\[
sim(c_i, c_j) = \frac{1}{|x_i| \times |y_j|} \sum_{x_i \in x_i, y_j \in y_j} C2S(x,y)
\]

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

In our algorithm, this merging process will terminate automatically 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.

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 spectometry, 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, \( \sim(x,\{\}|y) = C2S(x,y) \).

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

Step 3 (Termination or not?): If \( \sim_{\text{best}} \) 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 \( c_{\text{new}} \) is created by merging the two clusters \( c_i \) and \( c_j \). The cluster \( c_{\text{new}} \) is then added into \( C \), while \( c_i \) and \( c_j \) are removed. Then, for each \( c_k \in C \) (different from \( c_{\text{new}} \)), calculate its similarity with the new cluster \( c_{\text{new}} \) as follows:

\[
\sim(c_{\text{new}}, c_k) = \frac{|x_i| \times |x_k| \times \sim(x_i, x_k) + |x_j| \times |x_k| \times \sim(x_j, x_k)}{|x_i| + |x_j| + |x_k|}
\]

Step 5 (Loop): Go to Step 2

Table 1. Details of AP-MS datasets

<table>
<thead>
<tr>
<th>AP-MS datasets</th>
<th>#Purifications</th>
<th>#Baits</th>
<th>#Preys</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gavin</td>
<td>2166</td>
<td>1993</td>
<td>2671</td>
</tr>
<tr>
<td>Krogan</td>
<td>4332</td>
<td>2294</td>
<td>5333</td>
</tr>
<tr>
<td>Krogan-HighConf</td>
<td>3575</td>
<td>2143</td>
<td>2567</td>
</tr>
</tbody>
</table>

Krogan et al. (2006) used two separate methods [MALDI-TOF and Liquid chromatography-mass spectrometry (LC-MS)/MS].
We propose this new definition is due to the fact that the original
to evaluate the accuracy of the predicted complex set, sensitivity
The existing complex sets predicted by combining Gavin and
statistics of predicted complex sets to be compared
be defined as the geometric average of sensitivity and PPV,
reference complexes do not overlap with each other.
address this issue and yield PPV value of 1 when a complex set
itself, the resulting PPV is only 0.68. The new proposed PPV can
complex set is taken to match with itself, then the resulting PPV
by Li et al. (2010). For example: if the known gold standard MIPS
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

Table 2. Statistics of predicted complex sets to be compared

<table>
<thead>
<tr>
<th>Predicted complex set</th>
<th>#Complexes</th>
<th>Avg. complex size</th>
<th>#distinct proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2S</td>
<td>1039</td>
<td>4.93</td>
<td>5121</td>
</tr>
<tr>
<td>C2S-HighConf</td>
<td>679</td>
<td>3.79</td>
<td>2571</td>
</tr>
<tr>
<td>Hart</td>
<td>390</td>
<td>4.33</td>
<td>1689</td>
</tr>
<tr>
<td>Pu</td>
<td>400</td>
<td>5.14</td>
<td>1913</td>
</tr>
<tr>
<td>BT-893</td>
<td>893</td>
<td>6.25</td>
<td>5187</td>
</tr>
</tbody>
</table>

The existing complex sets predicted by combining Gavin and
Brohee and Helden (2006). Given a threshold $\omega$, we say that $r$
and $c$ match each other if $\text{MD}(r,c) \geq \omega$. Let $M_{\text{ref}}$ be
the number of reference complexes that match at least one predicted
complex, and $M_{\text{pre}}$ 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):

$$\text{Precision} = \frac{M_{\text{pre}}}{|C|}$$

$$\text{Recall} = \frac{M_{\text{ref}}}{|R|}$$

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

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

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):

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

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

<table>
<thead>
<tr>
<th>Predicted complex set</th>
<th>CYC2008 (408)</th>
<th>MIPS (214)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sn</td>
<td>PPV</td>
</tr>
<tr>
<td>C2S</td>
<td>0.680</td>
<td>0.837</td>
</tr>
<tr>
<td>C2S-HighConf</td>
<td>0.643</td>
<td>0.889</td>
</tr>
<tr>
<td>Pu (400)</td>
<td>0.691</td>
<td>0.789</td>
</tr>
<tr>
<td>Hart (390)</td>
<td>0.610</td>
<td>0.863</td>
</tr>
<tr>
<td>BT-893</td>
<td>0.720</td>
<td>0.759</td>
</tr>
</tbody>
</table>

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 curated 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):

$$\text{Sn} = \frac{\sum \max T_{i,j}}{\sum |r_i|}, \quad \text{and} \quad \text{PPV} = \frac{\sum \max T_{i,j}}{\sum |(U(r_i) \cap r_j)|}$$

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,

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

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,
they evidently outperform the unsupervised method (BT-893) which drops faster with the increased value of $\omega$.

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 Ontology (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.

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.

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 $\omega$. From Figure 4, it is evident that our approach remains the best among the compared predictions when $\omega$ 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

<table>
<thead>
<tr>
<th>Predicted complex set</th>
<th>CYC2008 (408)</th>
<th>MIPS (214)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sn</td>
<td>PPV</td>
</tr>
<tr>
<td>C2S-Gavin (474)</td>
<td>0.588</td>
<td>0.884</td>
</tr>
<tr>
<td>Zhang-Gavin (851)</td>
<td>0.607</td>
<td>0.679</td>
</tr>
<tr>
<td>Gavin-Core (478)</td>
<td>0.392</td>
<td>0.914</td>
</tr>
<tr>
<td>Gavin-All (491)</td>
<td>0.570</td>
<td>0.552</td>
</tr>
<tr>
<td>CODEC-w0-Gavin (1082)</td>
<td>0.552</td>
<td>0.506</td>
</tr>
<tr>
<td>CODEC-w1-Gavin (1005)</td>
<td>0.549</td>
<td>0.542</td>
</tr>
</tbody>
</table>

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

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.

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.

REFERENCES


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

Conflict of Interest: none declared.