Integrating experimental and literature protein-protein interaction data for protein complex prediction

Yijia Zhang*, Hongfei Lin, Zhihao Yang, Jian Wang

From The Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015)
HsinChu, Taiwan. 21-23 January 2015

Abstract

Background: Accurate determination of protein complexes is crucial for understanding cellular organization and function. High-throughput experimental techniques have generated a large amount of protein-protein interaction (PPI) data, allowing prediction of protein complexes from PPI networks. However, the high-throughput data often includes false positives and false negatives, making accurate prediction of protein complexes difficult.

Method: The biomedical literature contains large quantities of PPI data that, along with high-throughput experimental PPI data, are valuable for protein complex prediction. In this study, we employ a natural language processing technique to extract PPI data from the biomedical literature. This data is subsequently integrated with high-throughput PPI and gene ontology data by constructing attributed PPI networks, and a novel method for predicting protein complexes from the attributed PPI networks is proposed. This method allows calculation of the relative contribution of high-throughput and biomedical literature PPI data.

Results: Many well-characterized protein complexes are accurately predicted by this method when apply to two different yeast PPI datasets. The results show that (i) biomedical literature PPI data can effectively improve the performance of protein complex prediction; (ii) our method makes good use of high-throughput and biomedical literature PPI data along with gene ontology data to achieve state-of-the-art protein complex prediction capabilities.

Background

Protein complexes are formed from two or more associated polypeptide chains, and accurate determination of protein complexes is of great importance for understanding cellular organization and function. Many proteins are only functional after assembly into protein complexes. Even in the relatively simple model organism Saccharomyces cerevisiae, protein complexes include many subunits that assemble and function in a coherent fashion. A key task of system biology is to understand proteins and their interactions in terms of protein complexes [1].

Recent advances in high-throughput experimental techniques such as yeast two-hybrid and mass spectrometry have generated a large amount of protein-protein interaction (PPI) data for numerous organisms [2,3]. These high-throughput PPI data facilitate the development and testing of computational methods for protein complex prediction. The molecular complex detection (MCODE) algorithm proposed by Bader and Hogue [4] was one of the first computational methods reported. The Markov clustering algorithm [5] was also applied to predict protein complexes by simulating random walks within PPI networks. Adamcsek et al. developed the CFinder tool [6] that found functional modules in PPI networks using the clique percolation method [7] to detect k-clique percolation clusters. Liu et al. proposed a clustering method based on maximal cliques (CMC) to detect protein complexes [8]. Wu et al. developed the COACH algorithm [9] based on core-attachment structural features [10]. COACH initially

* Correspondence: zhyj@dlut.edu.cn
College of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning, China

http://www.biomedcentral.com/qc/1471-2164/16/S2/S4
© 2015 Zhang et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
identifies protein-complex cores at the heart of protein complexes, then attaches other proteins to these cores. Since proteins may have multiple functions, they may belong to more than one protein complex. Nepusz et al. proposed the ClusterONE algorithm [11] which detected overlapping protein complexes in PPI networks.

One major problem with high-throughput experimental PPI data is the high incidence of both false positives and false negatives [12]. Computational methods that only use high-throughput PPI data do not generally predict protein complexes accurately. This situation is improved if gene expression and gene ontology (GO) data are included. Feng et al. used microarray data to weight PPI networks, and this markedly improved the initial binary PPI networks [13]. Zhang et al. proposed the COAN algorithm based on ontology augmentation networks constructed with high-throughput PPI and GO annotation data. COAN takes into account the topological structure of the PPI network, as well as similarities in GO annotations [14].

The biomedical literature contain a large amount of potentially valuable PPI data that can be used to further improve protein complex prediction algorithms. In this study, we attempt to use this resource by first employing a natural language processing technique to extract PPI data from the biomedical literature. This is then integrated with high-throughput PPI and GO data by constructing attributed PPI networks that can be used for protein complex prediction. This novel approach automatically calculates the relative contributions of high-throughput and biomedical literature PPI data. The method is compared with current protein complex prediction tools. The advantages of the method, potential applications and improvements are discussed.

Methods

Extracting PPI data from the literature

The biomedical literature contains lots of potentially valuable PPI data, and extraction of this data is an important research topic in the field of biomedical natural language processing [15,16].

Our method of extracting PPI data consists of three phases: (i) named entity recognition (NER); (ii) normalization; (iii) extracting PPI data (Figure 1). NER aims to identify protein names in the biomedical literature. In our method, we use the FCG model [17] which is a semi-supervised learning strategy. FCG involves learning a novel feature representation from the co-occurrence of class-distinguishing features (CDFs) and example-distinguishing features (EDFs). CDFs and EDFs refer to strong indicators for classes and for examples, respectively. Their co-occurrence in large unlabeled datasets captures information that can not be obtained from labeled training data due to data sparseness.

Protein name normalization is used to determine the unique protein identifiers mentioned in the literature, linking these entities to biological databases. It is difficult to choose between ambiguous protein names based on context and short textual descriptions. We decide to adopt a disambiguation method [16] based on extended semantic similarity, which enriches gene descriptions in databases with information extracted from GO and PubMed abstracts. This allow us to exploit context and extend semantic information.

In the three phases, we extract PPI data from biomedical literature based on NER and normalization. Pattern-based methods is an established methodology for PPI extraction that usually uses defined lexical patterns and retrieves text segments that match the patterns. Because this approach is too rigid to capture semantic/syntactic paraphrases or distant relationships, such pattern-based methods always suffer from low recall rates. Instead, publicly accessible annotated PPI corpora such as GENIA [18] and Almed [19] allow automatic extraction of PPI data using machine learning methods. Recent studies [15,16] have established the power of machine learning methods, which handle PPI extraction as a classification problem. The major challenge is in supplying the learner with the semantic/syntactic information-containing features in order to distinguish between interactions and non-interactions.

Initial filtering of sentences that contain at least two protein entities is performed, and using the Stanford lexical parser generates syntactic information. At this stage, including a dependency graph or syntactic parse tree of candidate sentences maximizes the chances of efficient and accurate data extraction. However, syntactic information such as syntactic parse tree is not easily represented.
Construction of attributed PPI networks

Most computational methods for complex prediction are clearly limited by the poor quality of high-throughput PPI data. Further improvements for complex prediction can be obtained by integrating biomedical literature PPI data. GO is another useful resource for protein complex prediction, which is currently one of the most comprehensive ontology databases in the bioinformatics community [20]. GO aims to standardize the annotation of genes and gene products across species and provides a controlled vocabulary of terms for describing gene product biological properties. Due to the inherent biological properties of protein complexes [10], GO provides valuable PPI data for protein complex prediction. An example of a simple PPI network in which a vertex represents a protein and an edge represents the interaction between two proteins is shown in Figure 2 (a). Due to the presence of noise and the complex connectivity of PPI data, it is hard to predict protein complexes from this type of network. Figure 2c shows that two protein complexes can be predicted reasonably accurately when the PPI network is annotated by GO slims (Figure 2b). Therefore, an accurate method for protein complex prediction should generate similar clusters based on topological structure and GO annotation. In this study, we integrate high-throughput experimental PPI data, biomedical literature PPI data, and GO to predict protein complexes using attributed PPI networks.

We define an attributed PPI network as a 6-tuple \( G = (V, E, A_v, A_e, F_v, F_e) \) where \( V \) is the set of protein vertices, \( E \) is the set of PPIs, \( A_v = \{G_{S1}, G_{S2},..., G_{Sn}\} \) is the set of GO slim attributes for protein vertices, and \( F_v \) is a function that returns the set of GO slim attributes of a protein vertex. Each protein vertex \( pi \) in \( V \) has a set of GO slim attributes \( F_v(pi) = \{G_{S1}, G_{S2},..., G_{Sn}\} \), where \( m = |F_v(pi)| \) and \( F_v(pi) \subseteq A_v \). Likewise, \( A_e = \{T_1, T_2,..., T_r\} \) is the set of type attributes for PPIs, and \( F_e \) is a function that returns the set of type attributes of a PPI. Each PPI \( e_i \) in \( E \) has a set of type attributes \( F_e(e_i) = \{T_1, T_2,..., T_r\} \), where \( r = |F_e(e_i)| \), \( F_e(e_i) \subseteq A_e \) and \( Fe(e_i) \subseteq A_e \).

In this study, the type attributes of PPIs included high-throughput type and biomedical literature type (\( A_e = \{T_1, T_2\} \)).

Figure 3a shows an example of an attributed PPI network. The GO slim attributes of protein vertices and type attributes of PPI data are given in Figure 3b. It can be seen that each protein vertex has a GO slim attribute set and each edge has a type attribute set. For instance, \( P_2 \) has two GO slim attributes (\( GS_1 \) and \( GS_2 \)), and \( e_3 \) has two type attributes (\( T_1 \), high-throughput type; \( T_2 \) biomedical literature type). Given the set of GO slim attributes \( A_v \), we define an attribute set \( S \) as a subset of \( Av \). Moreover, we denote by \( V(S) \subseteq V \) the vertex set induced by \( S \) (i.e., \( V(S) = \{P \in V | S \subseteq F_v(P)\} \)) and by \( E(S) \subseteq E \) as the edge set induced by \( S \) (i.e., \( E(S) = \{(P, P) \in E | P \in V(S)\} \)). The subgraph \( G(S) \), induced by \( S \), is the pair \( (V(S), E(S)) \).

Figure 3c,d depict the subgraphs induced by the attribute set \( \{GS_1\} \) and \( \{GS_1, GS_2\} \), respectively.

Ontology correlated clique score

Definition 1 - Ontology correlated clique. Given a protein vertex set \( C \) and an edge set \( E_c \), the induced subgraph \( G(S) (C \subseteq V(S), E_c = \{(P, P) \in E(S) | P \in C) \}) \), an ontology correlated clique is a pair (\( (C, E_c) \), such that for each protein vertex \( Pi \) in \( C \), the degree of \( Pi \) is \( |C| - 1 \). \( S \) is the common ontology attribute set of \( C \).

In general, we can mine many Ontology correlated cliques with different common ontology attribute sets in an attributed PPI network. Figure 4 shows three ontology correlated cliques of the attributed PPI network in Figure 3.

Definition 2 - Structural correlated function \( \eta \). Given an ontology slim attribute set \( S \), the structural correlation of \( S \), \( \eta(S) \), is given as:

\[
\eta(S) = \frac{|K_s|}{|V(S)|}
\]

where \( K_s \) is the set of vertices in ontology correlated cliques in \( G(S) \). In this study, we only considered cliques of three or more.

A correlation function can be used to measure the dependence between ontology attribute set \( S \) and the density of the associated vertices [21]. This indicates how likely \( S \) is to be part of a clique. The larger the structural correlation function \( \eta(S) \), the more valuable the ontology attribute set \( S \). In Figure 3, \( K_{GS_1} = \{P_2, P_3, P_4, P_5, P_6\} \), \( K_{GS_2} = \{P_2, P_3, P_4, P_5\} \), \( K_{GS_3} = \{\} \) and \( K_{GS_1, GS_2} = \{P_2, P_3, P_4, P_5\} \). Thus the corresponding values of \( \eta(GS_1) \), \( \eta(GS_2) \) and \( \eta(GS_3) \) are 0, 0.625, and 0 and 1, respectively. Therefore, the protein vertices annotated by...
ontology attribute set $\{GS_1, GS_2\}$ are more likely to be part of a protein complex than those annotated by $\{GS_1\}$ or $\{GS_3\}$.

Definition 3 - Attributed network density. Given an attributed network $G = (V, E, A_v, A_e, F_v, F_e)$, the density of $G$, $\text{Density}(G)$, is given as:

$$\text{Weight}(e_i) = \sum_{T_i \in F_i(e_i)} w_j$$

$$\text{Density}(G) = \frac{2 \cdot \sum_{e_i \in E} \text{Weight}(e_i)}{|V| \cdot |V - 1|}$$

Since high-throughput PPI data and biomedical literature PPI data may differ in importance for protein complex prediction, we assign a weight to each type attribute to model their relative contributions. In equation (2), $w_j$ denotes the weight of the contribution of type attribute
of the PPIs, and weight\((e_i)\) is the weight of \(e_i\). Moreover, the edges of attributed networks may have different weights due to their different type attributes. For example, if we set \(w_1 = 0.6\) and \(w_2 = 0.4\), weight\((e_1)\), weight\((e_2)\) and weight\((e_3)\) are 0.6, 0.4 and 1, respectively (Figure 3). This indicates that \(e_3\) makes a more important contribution, or is known with higher confidence, than \(e_1\) and \(e_2\).

The ontology correlated clique score \(((C, E_c), S)\) is calculated as follows:

\[
\text{Clique Score}((C, E_c), S) = \eta(S) \cdot |C| \cdot |S| \cdot \text{Density}(C, E_c)
\]

where \(S\) is the common ontology attribute set of \(C\). Based on equation (4), we can evaluate these ontology correlate cliques based on both topology structure and the similarity of the ontology attributes. Table 1 shows the statistics of the ontology correlated cliques in Figure 4.

**Protein complex prediction from attributed PPI networks**

Our method for predicting protein complexes from attributed PPI networks involves two phases. In the first phase, we use high-throughput PPI data, and biomedical literature PPI data with GO, to construct two attributed PPI networks. The relative contributions of the high-throughput and literature PPI data is weighted automatically for each network. In the second phase, high-throughput PPI data, and biomedical literature PPI data with GO, are used to construct two whole attributed PPI networks. Based on the relative contributions of the PPI data, we predict the protein complexes from the whole attributed PPI networks.

In the first phase, we construct two attributed PPI networks with GO slims annotations and each type PPI data in turn: one for high-throughput PPI data and GO slims annotations, and the other for biomedical literature PPI data and GO slims annotations. We set the initial contribution weight of PPI data as \(w_1 = w_2 = 0.5\), and used the cliques mining algorithm [22] to enumerate all maximal cliques of three or more from the two attributed PPI networks in turn, and calculated the ontology attribute set for each maximal clique. The candidate clique set \(Candidate\) is comprised of all maximal ontology correlated cliques, which are generally overlapped. Maximal ontology correlated cliques in \(Candidate\) are ranked in descending order of clique score, denoted as \(((C_1, E_{c1}), S_1), ((C_2, E_{c2}), S_2),...,((C_m, E_{cm}), S_m))\). The top ranked clique \(((C_1, E_{c1}), S_1)\)
is then deleted from Candidate and inserted into the seed clique set Seed. To ensure the seed cliques non-overlapping, we use the same method [8] to remove or prune overlapping candidate cliques until the candidate clique set Candidate is empty. Two seed clique sets Seed1 and Seed2 are generated from the attributed PPI networks constructed using high-throughput PPI data and biomedical PPI data, respectively. The quality of seed clique set Seedi, generated from the attributed PPI networks constructed using $T_i$ type PPI data, indicated the value of $T_i$ type PPI data for protein complex prediction. Therefore, the average clique score can be used to evaluate the relative contributions of high-throughput experimental and literature PPI data. Based on equations (5) and (6), the contribution weight of $T_i$ type PPI data can be automatically calculated.

$$C_{\text{Degree}}(T_i) = \frac{\sum_{(C, E, S) \in \text{Seed}_i} \text{Clique Score}((C, E), S) \cdot |E_j|}{|E_i|}$$

(5)

$$w_i = \frac{C_{\text{Degree}}(T_i)}{\sum_{i=1}^{2} C_{\text{Degree}}(T_i)}$$

(6)

where Seed$_{T_i}$ denotes the seed clique set generated from the attributed PPI networks constructed using $T_i$ type PPIs, and $|E_{T_i}|$ is the total number of $T_i$ type PPI data.

In the second phase, we firstly construct the whole attributed PPI networks using high-throughput PPI data, literature PPI data and GO slims annotations. Secondly, we use the contribution weight $w_1$ and $w_2$ computed in the first phase to weight the whole attributed PPI networks. Thirdly, we generate the seed clique set Seed from the whole attributed PPI networks using the method described above for the first phase. Finally, we augment the seed cliques by adding each close neighbor protein vertex one by one. The closeness score is used

| Clique | S     | Density | $\eta(S)$ | Clique Score |
|--------|-------|---------|-----------|-------------|
| (P2, P5, P6) | (GS1) | 0.667   | 0.625     | 1.251       |
| (P2, P3, P5) | (GS1, GS2) | 0.867   | 1         | 5.202       |
| (P2, P3, P4, P5) | (GS1, GS2) | 0.833   | 1         | 6.664       |

($w_1 = 0.6$ and $w_2 = 0.4$).
to measure how closely a protein vertex \( P_k \) with ontology attribute set \( S_k \) is connected to a seed clique \(( (C_j, E_{cj}), S_j) \), where \( P_k \notin C_j \). The closeness score of \( P_k \) with respect to \(( (C_j, E_{cj}), S_j) \) is defined as follows:

\[
\text{Close} \_ \text{Score}(P_k, ((C_j, E_{cj}), S_j)) = \frac{|S_k \cap S_j|}{|S_j| + 1} + \frac{\sum_{e \in E_cj} \sum_{T_c \in F_c(e)} |w_j|}{|C_j|}
\]  
(7)

where \( E_cj \) is the set of edges between \( P_k \) and \(( (C_j, E_{cj}), S_j) \). In equation (7), \( |S_k \cap S_j| / |S_j| + 1 \) gives the GO annotation attribute similarity and \( \frac{\sum_{e \in E_cj} \sum_{T_c \in F_c(e)} |w_j|}{|C_j|} \) calculates the topology connectivity between the protein vertex \( P_k \) and the seed clique \(( (C_j, E_{cj}), S_j) \). Therefore, the closeness score provides a reasonable combination of both annotation attribute similarity and topology connectivity. If the \( \text{Close} \_ \text{Score}(P_k, ((C_j, E_{cj}), S_j)) \geq \text{extend} \_ \text{thres} \), then \( P_k \) was added to the seed clique \(( (C_j, E_{cj}), S_j) \). Therefore the final predicted protein complexes are generated by adding the close neighbor proteins to the seed complexes. Here, \( \text{extend} \_ \text{thres} \) is a predefined threshold and the value of \( \text{extend} \_ \text{thres} \) ranges from 0 to 1. The smaller the value of \( \text{extend} \_ \text{thres} \), the more neighbor proteins are added to the seed clique. If the value of \( \text{extend} \_ \text{thres} \) is close to 1, only the closest neighbors in both topology structure and biology attributes are added. The optimal value of \( \text{extend} \_ \text{thres} \) can usually be determined in preliminary experiments.

Results and discussion

In this section, the datasets and evaluation metrics used in the experiments are described. The value of biomedical literature PPI data for protein complex prediction is then considered. The impact of the \( \text{extend} \_ \text{thres} \) parameter is assessed, along with the relative contributions of high-throughput and literature PPI data. Finally, our method is compared with current state-of-the-art protein complex prediction methods.

Datasets and evaluation metrics

The two high-throughput PPI datasets used in our experiment are the Gavin dataset [3] and the Krogan dataset [23]. The Gavin dataset contains 1430 proteins and 6531 interactions, and the Krogan dataset contains 2675 proteins and 7080 interactions. The Biomedical literature data is a corpus of MEDLINE abstracts downloaded from PubMed. GO slim data is downloaded from http://www.yeastgenome.org. The benchmark protein complex dataset CYC2008 [24] includes 408 manually curated heterometric protein complexes for which experimental evidence has been reported.

To assess the quality of predicted protein complexes, we match generated complexes with the benchmark complex set CYC2008. Let \( P(V_p, E_p) \) be an predicted complex and \( B(V_{BP}, E_{BP}) \) be a known complex. We define the neighborhood affinity score \( \text{NA}(P,B) \) between \( P(V_p, E_p) \) and \( B(V_{BP}, E_{BP}) \) as follows:

\[
\text{NA}(P,B) = \frac{|V_p \cap V_{BP}|^2}{|V_p| \times |V_{BP}|}
\]

In this experiment, we consider \( P(V_p, E_p) \) and \( B(V_{BP}, E_{BP}) \) to match each other if \( \text{NA}(P,B) \) is larger than 0.2, which is the same as most methods for protein complex prediction [12].

Precision, recall and \( F\)-score has been used to evaluate the performance of most previous protein complex prediction methods. Precision measures the fidelity of the predicted protein complex set. Recall quantifies the extent to which a predicted complex set captures the known complexes in the benchmark set. \( F\)-score provides a reasonable combination of both precision and recall, and can be used to evaluate the overall performance. Recently, sensitivity (Sn), positive predictive value (PPV) and accuracy (Acc) have also been used to evaluate protein complex prediction tools. Acc represents a tradeoff between Sn and PPV. The advantage of the geometric mean is that it yields a low score when either Sn or PPV are low. A high degree of accuracy thus requires a high performance for both criteria. These definitions have been described in detail by Li et al. [12]. To keep in line with most previous studies, we calculate precision, recall and \( F\)-score in this study, and also report Sn, PPV and Acc.

Most protein complex prediction studies reported to date have involved \textit{Saccharomyces cerevisiae}. In this study, we download all \textit{S. cerevisiae}-related abstracts from 1990-2012 from MEDLINE using PubMed. We use the user-defined kernel interface of the SVMLight package http://svmlight.joachims.org/ to implement the HSP kernel.

Almed, BiolInfer, IEPA, HPRD50 and LLL are the five annotated PPI corpora that are most commonly used for PPI data extraction, and were used in this study to construct training datasets [19]. The results of the five training datasets (Table 2) include Annotated PPIs (the number of PPI data extracted from the annotated corpora) and Extracted PPIs (the number of PPI data extracted from the literature). Extracted PPI data found in only one of the Gavin or Krogan datasets is shown. BiolInfer is the largest corpora among the five annotated PPI corpora, however only 1196 PPIs are extracted with the BiolInfer training set. Almed extracts the most PPIs (2957), and LLL extracts 1871 PPIs despite only containing 300 annotated PPIs.

Firstly, we apply COACH [9], CMC [8], Cluster ONE [11] and COAN [14] to the PPI data extracted from literature. \( F\)-score results on five literature PPI data are...
listed in Table 3. From Table 3, it can be seen that literature PPI data extracted by Aimed corpora achieves the highest F-score on four protein complex prediction methods. We next apply the four complex prediction methods to the hybrid PPI data that comprised high-throughput experimental and literature-extracted PPI data. The F-score on the Gavin dataset are listed in Table 4. “Gavin+AImed” denotes this data comprised Gavin PPI data and PPI data extracted from the literature using the Aimed corpora training dataset. From Table 4 it can be seen that the F-scores of hybrid PPI data generally outperform those of the Gavin PPI data. For example, COACH and CMC are improved the F-score by 0.014 (from 0.406 to 0.42), and 0.038 (from 0.321 to 0.359) on the Gavin+AImed PPI data, respectively. In addition, we randomly add 1000, 2000 or 3000 interactions to the Gavin dataset. For each simulation, we perform 10 times randomization experiments and calculate the mean value and standard deviation. Table 4 shows that the F-scores achieved on the Gavin + random PPI data are inferior to on the Gavin PPI data.

The F-scores for the Krogan dataset are listed in Table 5. Again, all four approaches achieve better performance on the hybrid PPI data. This suggests that integrating literature PPI data and high-throughput PPI data can effectively improve the performance of protein complex prediction.

The effect of extend_thres
We construct two attributed PPI networks to integrate high-throughput PPI data, biomedical literature PPI data and GO as described in the Section 2.2. Attributed PPI network I is constructed using GO, the Gavin dataset, and extracted PPI data using the Aimed corpora as the training dataset. Attributed PPI network II is constructed using GO, the Krogan dataset, and extracted PPI data using the Aimed corpora as the training dataset. To study the effect of the extend_thres parameter, we first evaluate our method on Attributed PPI network I. Our method proves sensitive to extend_thres between 0.05 and 0.6 (Table 6). The F-score performance ranges from 0.386 to 0.447. Precision, recall and F-score are 0.506, 0.314 and 0.387, respectively, when extend_thres = 0.05. This indicates that too many proteins are added to the seed cliques to construct complexes during the seed cliques augment phase, suggesting the value of extend_thres is too small. As extend_thres is increased, the number of proteins added decreased sharply. When extend_thres = 0.1, precision, recall and F-score improve significantly to 0.589, 0.36 and 0.447, respectively. When extend_thres is increased from 0.1 to 0.6, precision, recall and F-score all decrease. The closeness score calculated using equation (7) provides a reasonable combination of both annotation attribute similarity and topology connectivity. In order to maintain a closeness score larger than 0.6, the candidate proteins must have highly similar ontology attribute set and topology connectivity to seed cliques during the seed cliques augment phase. However, there are few such candidate proteins in the attributed PPI networks. Therefore,

### Table 2 PPI extraction results for five training corpora

| Training dataset | Annotated PPIs | Extracted PPIs | Different from Gavin | Different from Krogan |
|------------------|---------------|---------------|----------------------|----------------------|
| Aimed            | 5834          | 2957          | 2729                 | 2659                 |
| BioInfer         | 9666          | 1196          | 1100                 | 1073                 |
| IEPA             | 817           | 2223          | 2072                 | 2039                 |
| HPRD50           | 433           | 2573          | 2390                 | 2362                 |
| LLL              | 300           | 1871          | 1756                 | 1722                 |

### Table 3 Results of PPI data extracted from biomedical literature

|        | COACH | CMC  | ClusterONE | COAN  |
|--------|-------|------|------------|-------|
| Aimed  | 0.1938| 0.1395| 0.1873     | 0.153 |
| BioInfer| 0.1233| 0.1072| 0.1578     | 0.1231|
| IEPA   | 0.1325| 0.1015| 0.1597     | 0.1422|
| HPRD50 | 0.1566| 0.1175| 0.1462     | 0.1396|
| LLL    | 0.1345| 0.1103| 0.1403     | 0.1325|

### Table 4 Results of Gavin PPI data and biomedical literature PPI data

|                | COACH | CMC  | ClusterONE | COAN  |
|----------------|-------|------|------------|-------|
| Gavin dataset  | 0.406 | -    | -          | 0.404 |
| Gavin + random I| 0.402 | 0.003| 0.311      | 0.009 | 0.408 | 0.012 | 0.402 | 0.004 |
| Gavin + random II| 0.398 | 0.005| 0.298      | 0.005 | 0.389 | 0.013 | 0.401 | 0.005 |
| Gavin + random III| 0.395 | 0.005| 0.283      | 0.012 | 0.366 | 0.013 | 0.393 | 0.008 |
| Gavin + Aimed   | 0.42  | -    | 0.359      | -     | 0.429 | -     | 0.428 | -     |
| Gavin + BioInfer| 0.414 | -    | 0.329      | -     | 0.415 | -     | 0.413 | -     |
| Gavin + IEPA    | 0.406 | -    | 0.342      | -     | 0.427 | -     | 0.409 | -     |
| Gavin + HPRD50  | 0.417 | -    | 0.32       | -     | 0.423 | -     | 0.409 | -     |
| Gavin + LLL     | 0.41  | -    | 0.337      | -     | 0.411 | -     | 0.419 | -     |

Gavin + random I, Gavin + random II, and Gavin + random III show the results of randomly adding 1000, 2000 and 3000 interactions to the Gavin dataset, respectively. The highest F-score of each approach is shown in bold.
Table 5 The results of Krogan PPI data and biomedical literature PPI data

| Extend_thres | COACH | CMC | ClusterONE | COAN |
|--------------|-------|-----|------------|------|
|              | F     | σF | F     | σF | F     | σF | F     | σF |
| Krogan dataset | 0.441 | 0.358 | 0.401 | 0.451 | 0.441 | 0.358 | 0.401 | 0.451 |
| Krogan + random I | 0.439 | 0.002 | 0.353 | 0.004 | 0.379 | 0.014 | 0.445 | 0.002 |
| Krogan + random II | 0.436 | 0.004 | 0.349 | 0.006 | 0.354 | 0.021 | 0.449 | 0.006 |
| Krogan + random III | 0.433 | 0.004 | 0.347 | 0.006 | 0.34 | 0.018 | 0.444 | 0.007 |
| Krogan + Aimed | 0.457 | 0.411 | 0.417 | 0.464 | 0.457 | 0.411 | 0.417 | 0.464 |
| Krogan + Bioinfer | 0.453 | 0.366 | 0.405 | 0.458 | 0.453 | 0.366 | 0.405 | 0.458 |
| Krogan + IEPA | 0.444 | 0.398 | 0.393 | 0.453 | 0.444 | 0.398 | 0.393 | 0.453 |
| Krogan + HPRD20 | 0.445 | 0.384 | 0.389 | 0.463 | 0.445 | 0.384 | 0.389 | 0.463 |
| Krogan + LLL | 0.454 | 0.393 | 0.404 | 0.453 | 0.454 | 0.393 | 0.404 | 0.453 |

Krogan + random I, Krogan + random II, and Krogan + random III show the results of randomly adding 1000, 2000 and 3000 PPIs to the Krogan dataset, respectively. The highest F-score of each approach is shown in bold.

The performance precision, recall and F-score are relatively unaffected when extend_thres varied between 0.6 and 1.0.

The Sn, PPV and Acc metrics are then evaluated. When extend_thres is changed from 0.05 to 0.6, PPV increased whereas Sn decreases. When extend_thres ranges between 0.6 and 1.0, Sn, PPV and Acc do not change appreciably (0.331, 0.701 and 0.482, respectively). This is due to more neighbor proteins being added to the seed cliques when extend_thres is low, resulting in the predicted complexes having better coverage of the benchmark dataset complexes, and improving the Sn metric. In contrast, only the closest neighbor proteins are added to the seed clique when extend_thres is high. This increases the likelihood of predicted complexes being true positives, and improves the PPV metric. Acc is defined as the geometric mean of Sn and PPV, which is potentially more comprehensive for evaluating performance. Similar to F-score, Acc is maximized (0.531) when extend_thres = 0.1.

We also evaluate the effects of extend_thres on Attributed PPI network II (Table 7; compare with Table 6). Again, the highest F-score (0.477) and Acc (0.551) are achieved when extend_thres = 0.1.

The relative contributions of experimental and literature-extracted PPI data

We evaluate the relative contributions of high-throughput experimental and literature PPI data to protein complex prediction performance. Another strength of our method is that it automatically computes contribution weights. The statistics of the contributions of literature data are listed in Table 8. In Attributed PPI network I, high-throughput and literature PPI data contribute weights of 0.59 and 0.41, respectively. In Attributed PPI network II, these are 0.55 and 0.45, respectively. Furthermore, we evaluate the effect of weight mechanism on these networks (Table 9). When no weight mechanism is used, equal weight is given to high-throughput and literature PPI data. The weight mechanism improves F-score by 0.025 and 0.014 on Attributed PPI networks I and II, respectively.

Table 6 The effect of extend_thres on protein complex prediction performance using Attributed PPI network I

| Extend_thres | P     | R     | F     | Sn     | PPV     | Acc     |
|--------------|-------|-------|-------|--------|---------|---------|
| 0.05         | 0.506 | 0.314 | 0.387 | 0.569  | 0.389   | 0.471   |
| 0.1          | 0.589 | 0.36  | 0.447 | 0.521  | 0.541   | 0.531   |
| 0.2          | 0.55  | 0.341 | 0.421 | 0.451  | 0.611   | 0.525   |
| 0.3          | 0.524 | 0.321 | 0.398 | 0.39   | 0.635   | 0.505   |
| 0.4          | 0.515 | 0.311 | 0.388 | 0.349  | 0.677   | 0.486   |
| 0.5          | 0.506 | 0.314 | 0.387 | 0.332  | 0.699   | 0.482   |
| 0.6          | 0.502 | 0.314 | 0.386 | 0.331  | 0.701   | 0.482   |
| 0.7          | 0.502 | 0.314 | 0.386 | 0.331  | 0.701   | 0.482   |
| 0.8          | 0.502 | 0.314 | 0.386 | 0.331  | 0.701   | 0.482   |
| 0.9          | 0.502 | 0.314 | 0.386 | 0.331  | 0.701   | 0.482   |
| 1.0          | 0.502 | 0.314 | 0.386 | 0.331  | 0.701   | 0.482   |

F: F-score, P: precision, R: recall. The highest F-score of each approach is shown in bold.

Table 7 The effect of extend_thres on protein complex prediction performance using Attributed PPI network II

| Extend_thres | P     | R     | F     | Sn     | PPV     | Acc     |
|--------------|-------|-------|-------|--------|---------|---------|
| 0.05         | 0.571 | 0.316 | 0.407 | 0.581  | 0.413   | 0.49    |
| 0.1          | 0.636 | 0.382 | 0.477 | 0.525  | 0.576   | 0.551   |
| 0.2          | 0.599 | 0.367 | 0.457 | 0.447  | 0.647   | 0.538   |
| 0.3          | 0.567 | 0.365 | 0.444 | 0.389  | 0.702   | 0.523   |
| 0.4          | 0.559 | 0.355 | 0.434 | 0.348  | 0.72    | 0.501   |
| 0.5          | 0.551 | 0.35  | 0.428 | 0.339  | 0.752   | 0.498   |
| 0.6          | 0.551 | 0.348 | 0.426 | 0.336  | 0.734   | 0.497   |
| 0.7          | 0.551 | 0.348 | 0.426 | 0.336  | 0.734   | 0.497   |
| 0.8          | 0.551 | 0.348 | 0.426 | 0.336  | 0.734   | 0.497   |
| 0.9          | 0.551 | 0.348 | 0.426 | 0.336  | 0.734   | 0.497   |
| 1.0          | 0.551 | 0.348 | 0.426 | 0.336  | 0.734   | 0.497   |

F: F-score, P: precision, R: recall. The highest F-score of each approach is shown in bold.

Table 8 The contribution weight of high-throughput experimental and literature PPI data using Attributed PPI networks for protein complex prediction

| Extend_thres | Attributed networks | PPIs | Weight |
|--------------|---------------------|------|--------|
|              | I                   | 0.59 | 0.41   |
|              | II                  | 0.635 | 0.297 |

High-throughput Literature

| Extend_thres | Attributed networks | PPIs | Weight |
|--------------|---------------------|------|--------|
|              | I                   | 0.55 | 0.45   |

Extend_thres = 0.1.
Comparison with other protein complex prediction methods

We compare our method with the following established leading protein complex prediction methods: Cluster ONE [11], COACH [9], CMC [8], HUNTER [25], MCL [5] and MCODE [4] (Table 10).

GO provides GO terms or slims to describe gene product characteristics in three different aspects, including Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). In the GO data, some of CC attributes are directly pertinent to protein complex. Firstly, we evaluate the effect of CC attributes of GO data to our method. In the Table 10, “Our method (BP, MF, CC)” and “Our method (BP, MF)” denote our method performed on whole GO slim data, and the GO slim data which removes CC attributes set, respectively. From Table 10, it can be seen that the F-score reduces 0.011 and 0.017 on Attributed PPI network I and II, respectively, when CC attributes set is removed from GO slim data.

Secondly, we compare our method using Attributed PPI network I with Cluster ONE, COACH, CMC, HUNTER, MCL and MCODE using the Gavin PPI network. Our method achieves an F-score of 0.447, which is significantly superior to the other methods (Table 10). Cluster ONE achieves the highest Acc of 0.534. It is worth noting that COACH predicts 326 complexes, which is much more than other methods. In contrast, HUNTER only predicts 69 complexes, albeit with the highest precision of 0.87 and a low recall of 0.206. MCL predicts 103 complexes, and achieves the highest Sn of 0.53.

Finally, we compare our method using Attributed PPI network II with the other methods using the Krogan PPI network. From Table 10, it can be seen that our method also achieves the highest F-score of 0.477. Cluster ONE achieves the highest Acc of 0.585. HUNTER and MCL achieve the highest precision (0.865) and Sn (0.57), respectively.

In summary, our method can integrate high-throughput experimental PPI data, biomedical literature PPI data, and GO by constructing attributed PPI networks. This approach outperforms existing protein complex prediction tools.

Conclusions

We exploit the natural language processing technique to extract PPI data from the biomedical literature and integrate this data with high-throughput experimental PPI data and GO to construct attributed PPI networks. Using these networks, we develop a novel method for protein complex prediction that automatically calculate the relative contributions of experimental and literature data. This approach outperforms established leading protein complex prediction tools. In the future, we intend to

Table 9 Performance comparison of the weight mechanism

| Weight | P   | R   | F   | Sn  | PPV  | Acc  |
|--------|-----|-----|-----|-----|------|------|
| Weight I | 0.589 | 0.36 | 0.447 | 0.521 | 0.541 | 0.531 |
| No weight I | 0.556 | 0.341 | 0.422 | 0.52 | 0.526 | 0.523 |
| Weight II | 0.636 | 0.382 | 0.477 | 0.525 | 0.576 | 0.551 |
| No weight II | 0.623 | 0.368 | 0.463 | 0.523 | 0.567 | 0.544 |

Weight I and Weight II denote the performance using the weight mechanism on attributed PPI networks I and II, respectively. No weight I and No weight II denote the performance without the weight mechanism F: F-score, P: precision, R: recall.

Table 10 Performance comparison with other protein complex prediction methods

| PPIN    | Methods               | #Complexes | P     | R     | F     | Sn    | PPV   | Acc   |
|---------|-----------------------|------------|-------|-------|-------|-------|-------|-------|
| Attr. PPIN I | Our method (BP,MF,CC) | 231        | 0.589 | 0.36  | 0.447 | 0.521 | 0.541 | 0.531 |
|          | Our method (BP,MF)    | 182        | 0.659 | 0.326 | 0.436 | 0.471 | 0.571 | 0.518 |
|          | ClusterONE            | 199        | 0.568 | 0.331 | 0.418 | 0.468 | 0.609 | 0.534 |
|          | COACH                 | 326        | 0.525 | 0.333 | 0.406 | 0.44  | 0.547 | 0.49  |
|          | CMC                   | 120        | 0.608 | 0.218 | 0.321 | 0.371 | 0.606 | 0.474 |
| Gavin PPIN | HUNTER               | 69         | 0.87  | 0.206 | 0.333 | 0.386 | 0.508 | 0.443 |
|          | MCL                   | 103        | 0.718 | 0.245 | 0.366 | 0.53  | 0.489 | 0.509 |
|          | MCODE                 | 70         | 0.739 | 0.154 | 0.255 | 0.283 | 0.519 | 0.384 |
| Attr. PPIN II | Our method (BP,MF,CC) | 247        | 0.636 | 0.382 | 0.477 | 0.525 | 0.576 | 0.551 |
|          | Our method (BP,MF)    | 206        | 0.679 | 0.348 | 0.46  | 0.477 | 0.578 | 0.525 |
|          | ClusterONE            | 464        | 0.375 | 0.431 | 0.401 | 0.523 | 0.655 | 0.585 |
|          | COACH                 | 345        | 0.617 | 0.343 | 0.441 | 0.432 | 0.544 | 0.485 |
|          | CMC                   | 111        | 0.748 | 0.235 | 0.358 | 0.381 | 0.589 | 0.474 |
|          | HUNTER                | 74         | 0.865 | 0.199 | 0.323 | 0.374 | 0.569 | 0.462 |
| Krogan PPIN | MCL                  | 309        | 0.291 | 0.245 | 0.266 | 0.57  | 0.396 | 0.475 |
|          | MCODE                 | 72         | 0.75  | 0.159 | 0.263 | 0.27  | 0.552 | 0.386 |

#Complexes refers to the number of predicted complexes. F: F-score, P: precision, R: recall. The highest F-score of each approach is shown in bold.
incorporate a post-processing phase and make even better use of literature data extraction to further improve protein complex prediction performance.

Competing interests
The authors declare that they have no competing interests.

Declarations
This work is supported by grant from the Natural Science Foundation of China (No. 61303088, 6134020 and 61272373), the Fundamental Research Funds for the Central Universities (No. DUT14CQ04).

This article has been published as part of BMC Genomics Volume 16 Supplement 2, 2015: Selected articles from the Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/16/S2

Published: 21 January 2015

References
1. Li M, Chen J-e, Wang J-x, Hu B, Chen G: Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. BMC bioinformatics 2008, 9(1):398.
2. Rajagopala SV, Sikorski P, Kumar A, Mosca R: The binary protein-protein interaction landscape of Escherichia coli. 2014, 32(3):2851-290.
3. Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon A-M, Cruciat C-M: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002, 415(6868):141-147.
4. Bader GD, Hogue CW: An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics 2003, 4(1):2.
5. Srihari S, Ning K, Leong HW: MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure. BMC bioinformatics 2010, 11(1):504.
6. Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T: CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics 2006, 22(8):1021-1023.
7. Palla G, Derényi I, Farkas U, Vicsek T: Uncovering the overlapping community structure of complex networks in nature and society. Nature 2005, 435(7043):814-818.
8. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP: Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 2006, 440(7084):637-643.
9. Tomita T, Takahashi H: The worst-case time complexity for generating all maximal cliques and computational experiments. Theoretical Computer Science 2006, 363(1):28-42.
10. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP: Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 2006, 440(7084):637-643.
11. Chu H, Chen S-H, Ho C-W, Ko M-T, Lin C-Y: A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles. BMC bioinformatics 2010, 11(Suppl 1):S25.

Cite this article as: Zhang et al.: Integrating experimental and literature protein-protein interaction data for protein complex prediction. BMC Genomics 2013, 16(Suppl 2):S4.

www.biomedcentral.com/submit

Submit your manuscript at www.biomedcentral.com/submit

— BioMed Central

Submit your next manuscript to BioMed Central and take full advantage of:

• Convenient online submission
• Thorough peer review
• No space constraints or color figure charges
• Immediate publication on acceptance
• Inclusion in PubMed, CAS, Scopus and Google Scholar
• Research which is freely available for redistribution

doi:10.1186/1471-2164-16-S2-S4