BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Integrating data and knowledge to identify functional modules of genes: a multilayer approach

Lifan Liang[1], Vicky Chen[1,2], Kunju Zhu[1,4], Xiaonan Fan[1,3], Xinghua Lu[1] and Songjian Lu[1*]

Abstract

Background: Characterizing the modular structure of cellular network is an important way to identify novel genes for targeted therapeutics. This is made possible by the rising of high-throughput technology. Unfortunately, computational methods to identify functional modules were limited by the data quality issues of high-throughput techniques. This study aims to integrate knowledge extracted from literature to further improve the accuracy of functional module identification.

Results: Our new model and algorithm were applied to both yeast and human interactomes. Predicted functional modules have covered over 90% of the proteins in both organisms, while maintaining a comparable overall accuracy. We found that the combination of both mRNA expression information and biomedical knowledge greatly improved the performance of functional module identification, which is better than those only using protein interaction network weighted with transcriptomic data, literature knowledge, or simply unweighted protein interaction network. Our new algorithm also achieved better performance when comparing with some other well-known methods, especially in terms of the positive predictive value (PPV), which indicated the confidence of novel discovery.

Conclusion: Higher PPV with the multiplex approach suggested that information from both sources has been effectively integrated to reduce false positive. With protein coverage higher than 90%, our algorithm is able to generate more novel biological hypothesis with higher confidence.

Keywords: Protein-protein interaction, Graph clustering, Random walk, Multiplex, Topic modeling, Gene expression, Functional module, Protein complex

Background

Understanding the mechanisms of pathway perturbations underlying complex human diseases remains a difficult problem, hindering the development of targeted therapeutics. Complex diseases involve many genes and molecules that interact within context-specific cellular networks, such as signaling networks, physical interaction networks, and co-expression networks [1]. For example, cancer was often viewed as the disruption of cellular signaling networks. Such complex networks are inherently modular [2], meaning that genes usually perform certain biological function in separate groups. Therefore, to investigate complicated cellular mechanism, it is necessary to characterize the modular structure of cellular networks.

A functional module is defined as a group of genes or their products which are related by one or more genetic or cellular interactions, e.g. co-regulation, co-expression or membership of a protein complex, of a metabolic or signaling pathway or of a cellular aggregate (e.g. chaperone, ribosome, protein transport facilitator) [3]. Since physical protein-protein interactions directly indicate the cooperation of gene products to drive a biological process, a variety of clustering methods were developed to identify functional modules from protein-protein interaction networks [4]. Zinman, et al. [5] have found that functional interactions that are part of functional modules

* Correspondence: songjian@pitt.edu
[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
Full list of author information is available at the end of the article

are conserved at a much higher rates, further supporting the advantage of using protein interaction networks.

In the past decade, a vast amount of methods has been developed to identify functional modules in protein-protein interaction networks. As summarized by previous reviews [4], a majority of these algorithms can be categorized into: (1) density-based [6, 7], identifying densely connected groups of proteins; (2) partition-based [8], separate all sparsely connected nodes; (3) flow simulation-based [8–11], simulating a biological or functional flow; (4) core attachment-based [12–14], exploiting the core-attachment structure of protein relations. Recently, evolutionary algorithm [15, 16] has been adopted to avoid poor local minimum; and node embedding [17] have been used to transform the graph clustering problem into a conventional clustering problem. In addition, algorithms [18, 19] combining two or more approaches described above has emerged. Unfortunately, the computational methods for functional module identification are clearly limited by the poor quality of the underlying PPI data, which is noisy with high rates of false positive and false negative [20, 21]. However, the various approaches to identify functional modules have served as the foundation to inspire further improvement. In this study, we followed the flow simulation-based approach to capture the dynamics of multiplex networks.

Another popular approach is to identify functional modules from co-expression network. Unlike protein interaction networks, edges in co-expression networks indicate differential expression of two genes within the same sample or condition. It assumes that tightly interacting and functionally dependent proteins are co-expressed across most conditions. This assumption is a reliable heuristic for functional module identification, despite that co-expression is not direct evidence for functional relation. Studies had successfully identified stable functional modules from co-expression networks across species [22]. Therefore, expression status of co-expression functional modules should be highly related to activities or behavior of cells. Many biological studies have identified active functional modules related to certain diseases from co-expression networks [23, 24].

However, in the case of co-expression network, identifying functional modules at the appropriate granularity is a big challenge. As each experimental condition usually has perturbed multiple signaling pathways, differentially expressed genes in each condition usually correspond to multiple dysregulated biological processes [20]. This could result in predicted functional modules being a superset of several real functional modules.

In addition, high-throughput expression data also has its own data quality issues. For example, RNAseq data still suffered from technical issues, such as batch effects and contamination. Recent studies have developed different methods to improve accuracy of module identification by integrating co-expression network and protein interaction networks [3, 20, 25–30] or other heterogeneous data sources [31–33], while others integrated homogeneous data sources to improve confidence [34, 35]. However, data quality issues common in high-throughput data, especially the experimental aspects, remain unresolved.

Besides high-throughput data, decades of research efforts have obtained and validated vast amounts of biological knowledge through wet-lab experiments, which are valuable resources for further research. Such knowledge should contain much less errors compared to high-throughput data. A few studies have attempted to utilized the literature for functional module identification [36–39]. However, relying on literature alone may lead to findings biased towards well studied genes, providing less novel insights [40, 41].

Since high-throughput data is less biased towards well-known genes and literature has fewer data quality issues, integrating these two information sources seems promising [42]. Methods [17, 43, 44] have been proposed to integrate prior knowledge. However, some of the methods suffer two major issues: (1) identification is restricted in the scope of knowledge, resulting in knowledge bias unresolved; (2) adoption of strong prior knowledge, such as Gene Ontology, that may not be independent from the gold standard, leads to overly optimistic evaluation results.

To address the issues above, this study has followed a multilayer network approach for data integration. Multiplex is a natural way to represent interactions in a complex system from multiple perspectives [45]. By treating prior knowledge separately in one layer, the identification process is not confined by knowledge, but enhanced by knowledge. In addition, although it is common practice to aggregate multiplex into a single weighted network [46–48], research on multiplex suggested that important information can be lost during aggregation [49]. Thus, this study seeks to capture multiplex dynamics with random walk / diffusion theoretic.

Random walks on multiplex can induce congestion even when each single layer remains decongested [50]. Also, the fraction of nodes a random walker can travel has increased, owing to their resilience to uniformly random failures [51]. Thus, the dynamics of diffusion is able to capture the additional information brought by multiplex formation. In this study, we first computed the first k step visit probability of the nodes in the multiplex, which can be viewed as the uncompressed, exact solution for node embedding [52]. Then we identified modules on the probability matrix with an objective function, named isolation, that promotes both module density and minimum cut in terms of k-step connectivity.

Two major hypothesis were tested in this study: (1) gene-topic associations extracted from literature is able to reveal functional relations of genes and provide information complementary to high-throughput data; (2) integration of multiple information sources with multiplex approach can improve the accuracy of functional module identification.

Result

We first identified differentially expressed genes from RNA expression data. Then we calculated topic-gene association from Pubmed titles and abstracts. These two types of data were used to calculate functional similarity among genes used as edge weights for protein interaction networks respectively. The two weighted PPI networks were further connected with the multiplex approach. Finally, we developed a clustering algorithm to identify functional modules with locally maximum isolation from the two-layer protein interaction network.

Our clustering algorithm on multiplex was compared with itself on single layer network to show the effectiveness to information integration. To further demonstrate its performance, a network integration algorithm named Similarity Network Fusion (SNF) [48] was also compared. Then the proposed algorithm was compared against other methods in terms of protein coverage and accuracy.

Descriptive statistics

BioGrid curation of PPI for *Saccharomyces cerevisiae* contained 32,353 interactions among 4518 gene products. The transcriptomic profile of yeast perturbation experiments contained expression values of 5980 genes under 1525 knockout conditions. The topic-gene association matrix contained 216 topics and 5348 genes.

After network construction, the yeast interactome based on topic modeling had 4187 genes and 30,989 interactions; the yeast interactome based on transcriptomic profiles contained 4179 genes and 30,887 interactions; the interactome based on the combination of the transcriptomic interactome and the topic-gene associations contained 8302 genes and 65,793 interactions.

The protein interaction network contained 10,945 nodes and 56,471 edges. The transcriptomic profile of breast cancer patients in TCGA contained 1218 samples and 20,252 genes. The topic-gene association matrix contained 209 topics and 16,712 genes.

After network construction, the human interactome based on transcriptomic profiles contained 10,029 genes and 49,909 edges. The human interactome based on topic modeling contained 10,368 genes and 48,806 edges. The combined interactome contained 19,266 genes and 212,292 edges.

Single-layer versus multiplex

We first checked if a method using both knowledge and expression data can obtain better performance than those using only protein interaction networks or combined with topic associaion. As shown in Fig. 1, 2, 3, 4, after being weighted by topic association ("human_topic" and "yeast_topic" in the legend), sensitivity, PPV and accuracy have been improved improved across different datasets and different gold standards. It was shown that topic-association data provided additional information about functional relations among genes.

After integrating the interactomes weighted by topic association and gene co-expression ("human_two_layers" and "yeast_two_layers" in Figs. 1, 2, 3, 4), PPV was further improved while sensitivity decreased slightly. This suggests our algorithm tends to identify clusters with less false positives, at the cost of inducing a few false negatives. Overall, accuracy increased with the multiplex approach.

The performance of the network fusion approach ("human_snf" and "yeast_snf" in Figs. 1, 2, 3, 4) seems to differ in different datasets. In the case of the human interactome, SNF has increased PPV and decreased sensitivity, which is similar with our method, though the

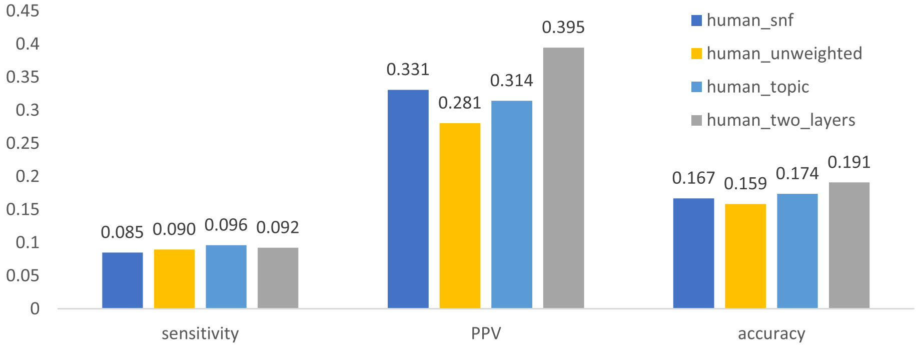

Fig. 1 Performance of isolation clustering on three different human interactomes, using Gene Ontology as gold standard

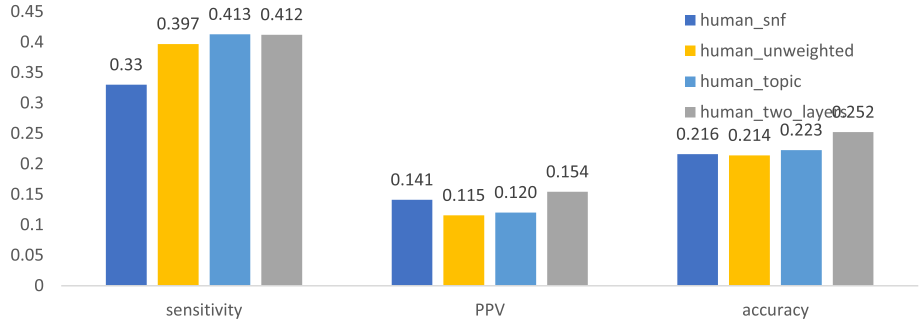

Fig. 2 Performance of isolation clustering on three different human interactomes, using CORUM as gold standard

overall performance gain is not obvious. For the yeast interctome, SNF yielded a performance worse than the single layer clustering in terms of sensitivity, PPV and accuracy. The reason could be that the iterative matrix computation procedure of SNF is more likely to return an almost uniform distribution of edge weights if the network density is high.

Comparison with other methods

We then compared our clustering method with some other well-known methods in terms of solution sizes, protein coverage, and accuracy. All the clusters with less than 3 proteins or larger than 200 proteins were removed. As shown in Tables 1 and 2, the distribution of cluster size for our method (isolation) is more skewed towards size 3–10. For the species of yeast, CYC2008 has over 83.3% of proteins with size less than 10, while the percentage of MCL, Infomap, Isolation was 73.8, 64.4, and 92.3% respectively. For the species of human, 89.5% of proteins complexes in CORUM contain less than or equal to 10 gene products, while 88.9% of functional modules generated by Isolation has such small size. Assuming that this distribution of CORUM and CYC2008 represents the true distribution of protein complexes, it indicated that the modular structure characterized by Isolation clustering was similar with that within real cells.

Protein coverage rates

As shown in Fig. 5, clusters generated by ClusterOne, MCODE, and Walktrap can only cover around half of the interactome. MCL, Infomap, and Isolation had covered over 90% of the interactome. Significantly higher coverages indicated that clustering methods based on random walks (i.e. MCL, Infomap, and Isolation) may provide more information about novel proteins so as to generate more biological insights. In the next section, only MCL, Infomap, and Isolation were compared against each other to in terms of accuracy.

Geometric accuracy

As shown in Figs. 6 and 7, Isolation has outperformed MCL and Infomap in yeast interactome in terms of geometric accuracy. The accuracy by our method is slightly higher than other methods. However, in the case of human interactomes, these three methods yielded very similar performance in every aspect.

Examples of clusters

Our clustering results have found many overlaps with known complexes. Two of them were perfect matches (Fig. 8). For some genes misclassified to a complex, we are able to identify close functional relations from literature. For example, our methods had grouped PINX1 with

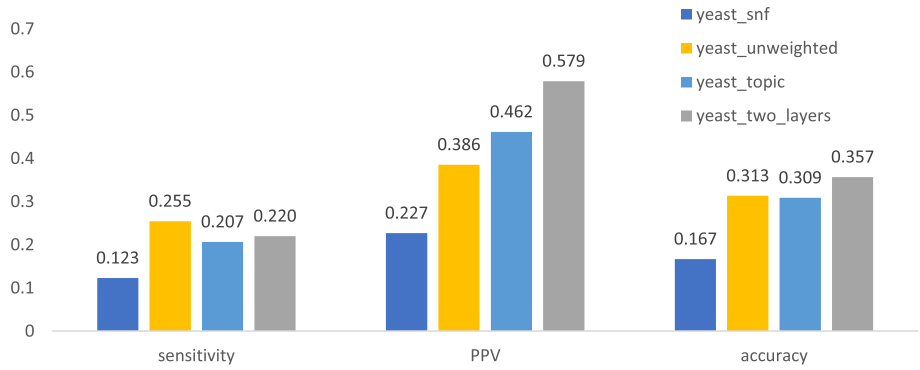

Fig 3. Performance of isolation clustering on three different yeast interactomes, using Gene Ontology as gold standard

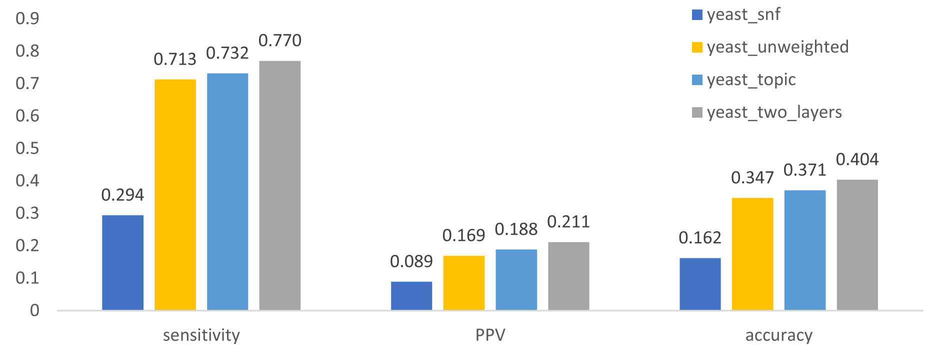

Fig. 4 Performance of isolation clustering on three different human interactomes, using CYC2008 as gold standard

TRF-Rap1 complex I (Fig. 9). Although PINX1 is not part of the complex, it is well studied that PINX1 can mediate TRF1 (or TERF1) and TERT accumulation in nucleus and enhances TERF1 binding to telomeres [53, 54], thus affecting the function of the complex.

Furthermore, "misclassified" genes without direct evidence supports may be more interesting since they could provide new insights for current knowledge. For example, C18orf21 was grouped with Rnase/Mrp complex by our method (Fig. 10). Several studies have found genetic association between variants in C18orf21 and phenotypes of human. Besides the high-throughput data (BioPlex [55]) used in this study, no further experiments have been conducted to investigate the functions of C18orf21. Our results suggested that C18orf21 could function through regulating Rnase/Mrp complex. Another example was shown in Fig. 11, where PNMA6A, DRAP1, PTCD3, AURKAIP1, and DDX55 were grouped with the 28S ribosomal subunit. Through literature we found that these misclassified genes, except PNMA6A, have significant impact on mitochondrial ribosome though detailed mechanisms are not clear [56–58].

Discussion

As illustrated in the result section, isolation clustering tends to identify isolated regions supported by both layers in the multiplex. Such tendency reduces false positives while inducing more false negatives. As shown in results, our new clustering algorithm, Isolation, has achieved better, or at least comparable, performance with other well-known clustering algorithms based on random walk. Particularly, subnetworks with locally maximal isolation exhibited higher confidence of being true positive when compared with MCL and Infomap. When compared with clustering algorithms such as ClusterOne, our algorithm has covered over 90% of proteins while density-based clustering can only cover around 50%. This leads to higher PPV from density-based clustering algorithms.

In addition, PPV is more important than sensitivity in terms of the confidence of true discovery. PPV of 1 indicates that the predicted module is a subset of a certain functional group in the gold standard, which means that every gene within the predicted module is related. On the other hand, sensitivity of 1 means a certain real functional module is a subset of a predicted module, which doesn't validate other functional relationships among the predicted module. Thus when end users try to identify genes functionally relevant with a specific gene, it is natural to focus more on positive predictive value or precision rather than composite scores or sensitivity used by most methodological evaluations. From this perspective, our integrative approach provides more practical values.

However, since the method is trading sensitivity for PPV, it could be problematic when data is more prevalent with false negative than false positive. In most cases, this means our algorithm is more suitable for dense networks. Users of our method should analyze the sparsity of the network before conducting the algorithm.

Table 1 The distribution of cluster size by different methods on yeast interactomes. The rightmost column is the gold standard used in this study

Size	MCL	Walktrap	Infomap	MCODE	ClusterOne	NCMine	Isolation	CYC2008
3–10	342	158	275	135	426	1096	995	198
11–50	107	44	140	21	86	612	82	36
51–100	13	5	7	11	10	23	1	2
100–200	0	5	5	2	4	0	0	0
> 200	1	0	0	0	0	0	0	0

Table 2 The distribution of cluster size by different methods on human interactomes. The rightmost column is the gold standard used in this study

Size	MCL	Walktrap	Infomap	MCODE	ClusterOne	NCMine	Isolation	CORUM
3–10	1008	323	506	319	1322	1943	2131	1562
11–50	353	83	241	47	108	253	260	176
51–100	15	13	16	8	0	0	4	5
100–200	0	3	3	1	0	0	1	2
> 200	0	0	0	3	0	0	0	0

Selected examples in the result section have shown that false positive genes could be functionally related in a way other that protein complexes. This illustrated one fundamental limitation for functional module identification and its evaluation. Biological experiments should be conducted to further verify the predicted modules.

This study also demonstrated that topic modeling of biomedical literature is an effective complementary source of information. Knowledge validated and curated in the form of literature are generally more reliable than high-throughput data. By integrating knowledge into the functional module identification process, false positives caused by data quality issues can be reduced. Thus, functional modules are identified with higher confidence. However, topic modeling of biomedical literature is not an easy task. The nHDP model used in this study took roughly 7 days to generate the topic mixtures. Future research may consider alternative information sources for integration.

Conclusion

In this paper, we have proposed a multiplex approach to integrate high-throughput data and literature for functional module identification and developed a clustering method that can utilize the topology based on random walk. Results showed that our algorithm is able to generate more novel biological hypothesis with higher confidence.

Methods

Topic modeling of genes

The title and abstract information of biomedical articles were downloaded from Pubmed. First, by treating each gene as a document, tf-idf scores were calculated to identify words most pertinent to a certain gene. To filter the documents, words with tf-idf scores lower than 167 were removed; and the vocabulary was restricted to 13,000. Second, a word vector was then created for each gene by going through its list of 200 words with the highest tf-idf scores and including only the ones that occur in the vocabulary. For each cancer sample, word vectors for its differentially expressed genes were combined. nHDP [59] was used to identify the latent topics in the set of combined word vectors.

Topic-document associations and topic-word associations generated from nHDP were further utilized to calculate the gene-topic association scores used in this study. Association strength between a certain gene g and a certain topic t was calculated by the total sum of products of: (1) a specific word w's count in g's word vector, (2) t's probability in document d, (3) the word w's

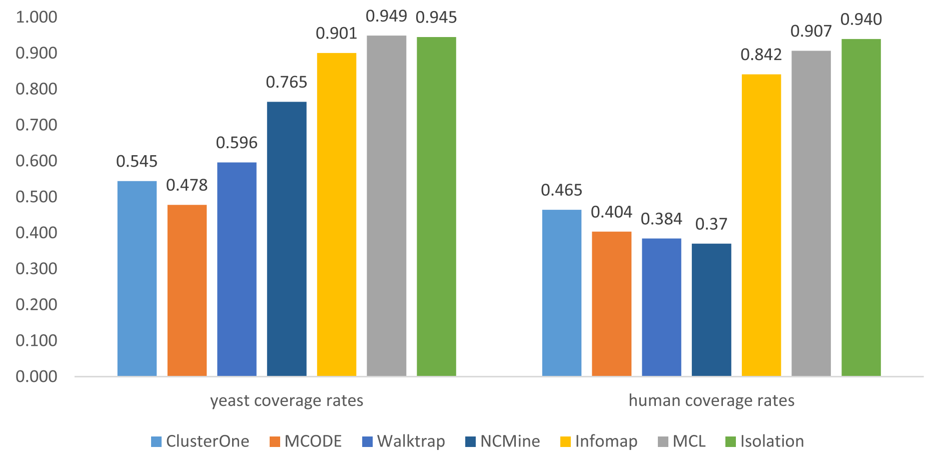

Fig. 5 In clustering for both yeast and human interactomes, clustering based on random walks has covered most proteins, while density-based clustering discarded around half the proteins

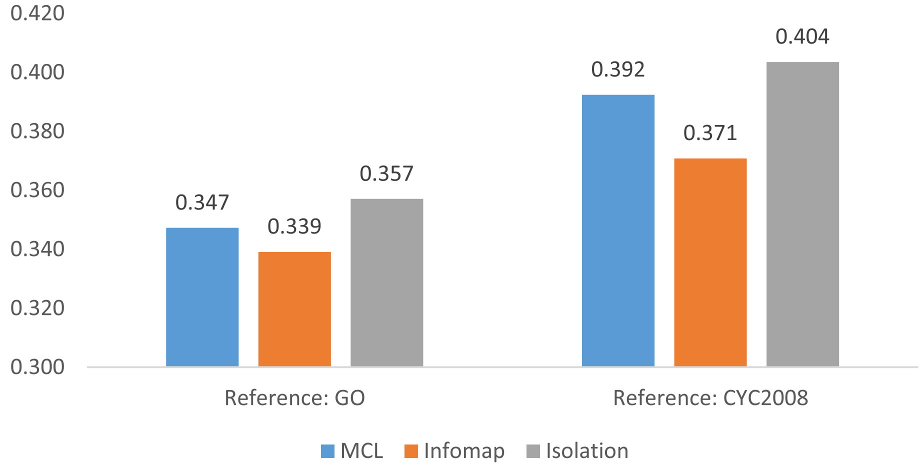

Fig. 6 Comparison of geometric accuracy of MCL, Infomap, and isolation on yeast interactomes

probability in t. For detailed description of this section, please see steps A-E in [37, 60].

Similarity measure

Functional similarity among genes was calculated with topic-gene association matrix and transcriptomic profiles respectively.

For the topic-gene association matrix, association scores less than one were set zero. Measure of similarity was computed based on Simrank [61]:

$$T_i = c_1\left(S^T G_i S\right) \tag{1}$$

$$G_i = c_2\left(S^T T_{i-1} S\right) \tag{2}$$

where S was a *g by n* matrix containing the association score between n topics and g genes, G_i was the g by g matrix containing the similarity among genes in the *i*th iteration, T_i was the n by n matrix containing the similarity among topics in the *i*th iteration, and c_1 and c_2 were the hyper-parameters controlling the impact of later iterations. In this study, both c_1 and c_2 were set to 0.8. The eq. (1) and (2) were iterated until T and G reach convergence. Note that only the similarity matrix G was used in the next section.

For the transcriptomic profile data, expression values were dichotomized. Genes expressions higher or lower than 95% interval of the distribution was encoded as one, otherwise zero. Cosine similarity was used to compute the similarity among genes, which is:

$$sim_{ij} = \frac{exp_i \cdot exp_j}{\sqrt{\| exp_i \| \cdot \| exp_j \|}} \tag{3}$$

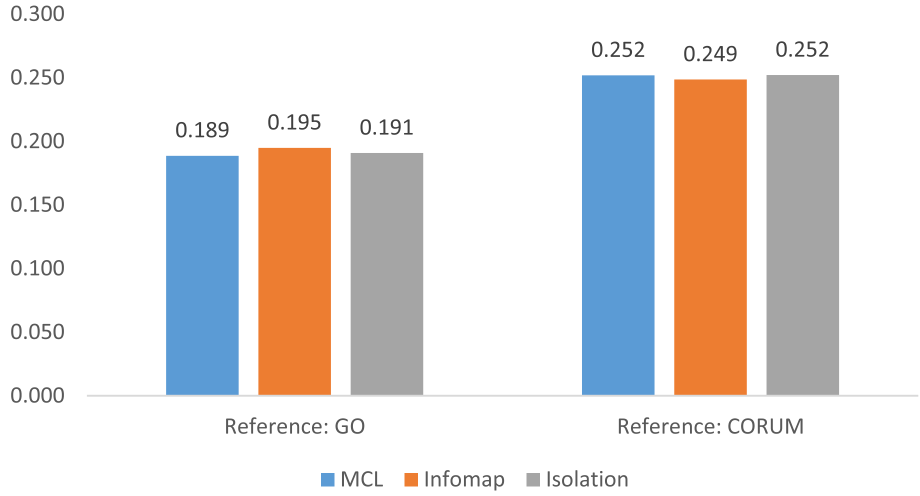

Fig. 7 Comparison of geometric accuracy of MCL, Infomap, and isolation on human interactomes

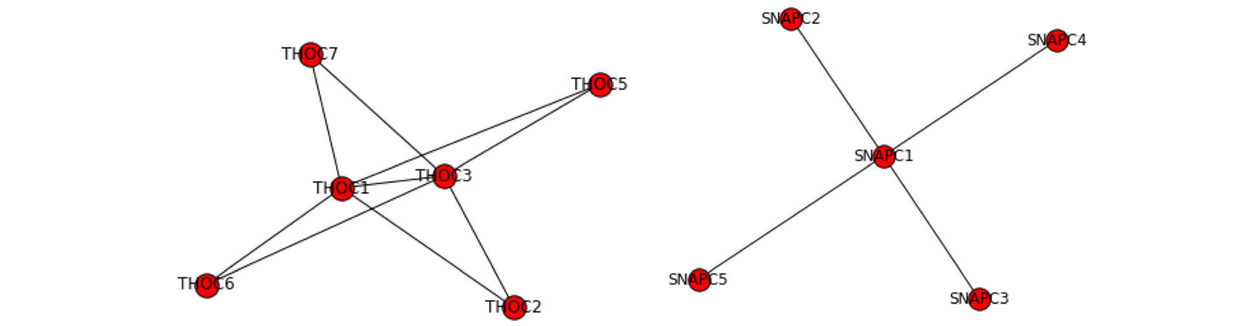

Fig. 8 The two predicted complexes perfectly matched to CORUM complexes. On the left is matched to hTREX84 complex. On the right is matched to SNAPc complex

where exp_i was the vector of expression values of the ith gene across all the experiments, $\|exp_i\|$ is the L2 norm of that vector.

Network construction

Computation of similarity matrix

Protein-protein interaction (PPI) networks were used as the base network. The similarity measures computed in the last section were used as the edge weights for these PPI networks. Thus, the topic-based interactome consisted of the topology of a PPI network with edge weights from topic-gene association matrix; and the expression-based interactome consisted of the topology of a PPI network with edge weights from transcriptomic profile data. For PPI curated in BioGrid for yeast, we only selected interactions supported by at least two studies.

These two interactomes were further combined into one network by treating each interactome as a layer and connecting the same gene across different layers, as demonstrated in Fig. 12.

Network integration

For all the networks described above, self-loops were removed. Edges with zero similarity and nodes with zero weighted degree were removed. The combined network is represented by supra-adjacency matrix [62]:

$$A = \begin{bmatrix} A_1 & I_N \\ I_N & A_2 \end{bmatrix}$$

where A_i is the adjacency matrix for the ith layer, I_N is an N by N identify matrix, N is the number of nodes in a single layer.

Clustering algorithm

The algorithm developed in this study consists of two steps: (1) transform the adjacency matrix into a matrix

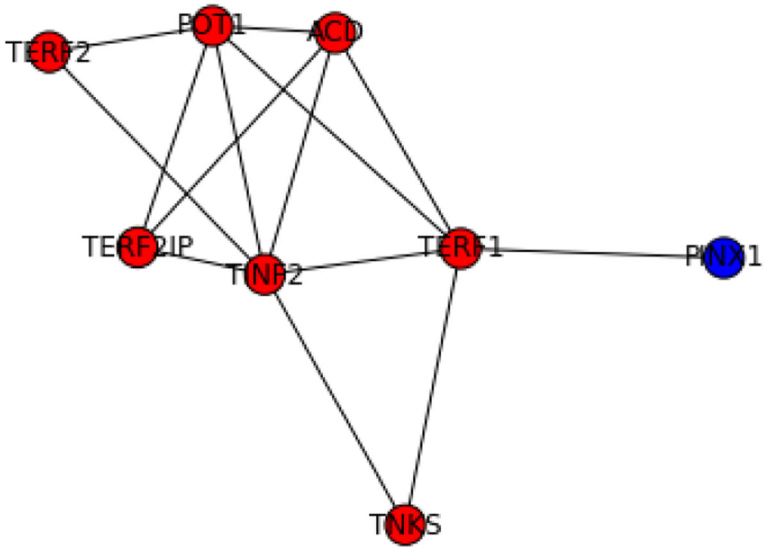

Fig. 9 Predicted complex matched to telomere-associated protein complex and TRF-Rap1 complex I, 2MD. Blue nodes were genes predicted but absent in the gold standard

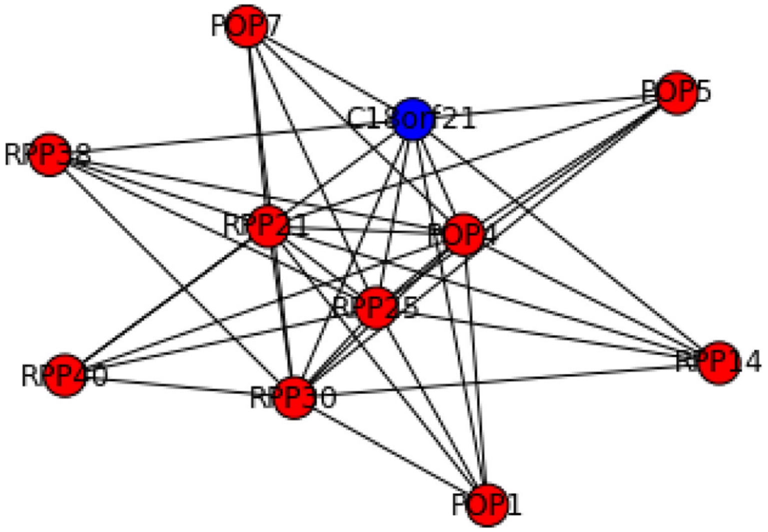

Fig. 10 Predicted complex matched to Rnase/Mrp complex. Blue nodes were genes predicted but absent in the gold standard

representing k-step walks visiting probability; (2) enumerate each node to identify clusters with locally optimal isolation.

Network transformation

With the network constructed from previous steps, the Markov transition matrix, M, should be computed next, which is:

$$M_{ij} = A_{ij}/A_{i.} \tag{4}$$

where $A_{i.}$ is the sum of the ith row of A.

From M, we further computed a matrix C, where C_{ij} is the probability that node j is visited if a walk of K steps start from node i. In this study, K is always set to 10. Since C_{ij} is complementary to the probability that node j never show up in the path, it can be computed as:

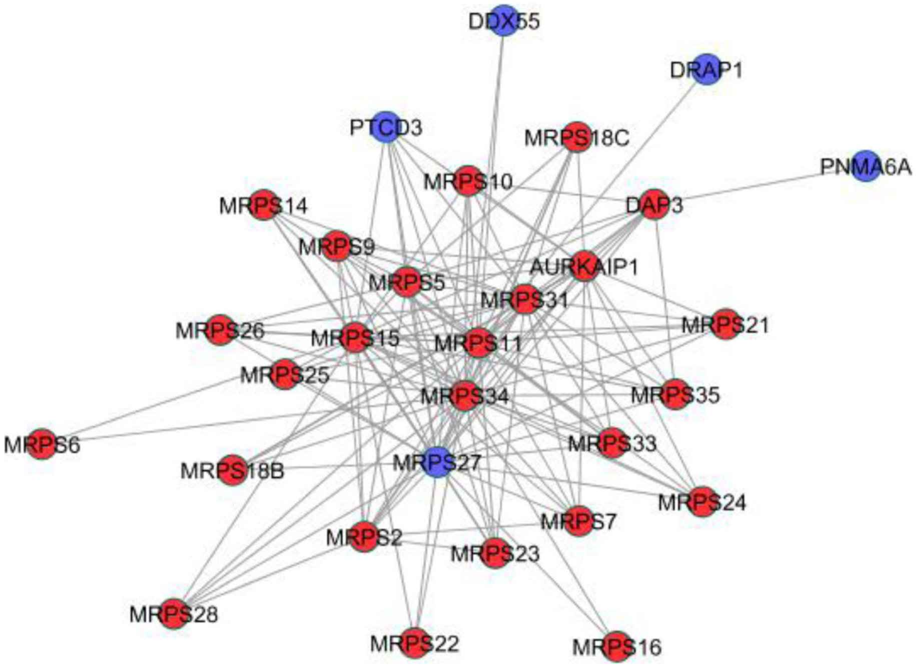

Fig. 11 Predicted complex matched to 39S ribosomal subunit, mitochondrial. Blue nodes were genes predicted but absent in the gold standard

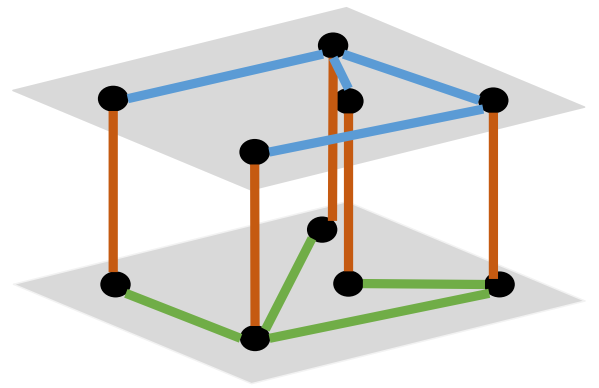

Fig 12. Illustration of the combined interactome, brown edges were artificial edges added to connect these two layers

$$C_{ij} = 1 - \mathbf{1}_i{}^{\boldsymbol{T}} \left(MI_{-j}\right)^{K} \mathbf{1} \tag{5}$$

where $\mathbf{1}_i$ is the vector with only the ith element as one, others zero, I_{-j} is an identity matrix with the jth diagonal value zero, $\mathbf{1}$ is the vector of 1.

As the vectorization of the operation above, the matrix C can be computed by the procedure below:

Box 1 Algorithm for computing the matrix C

```
$C_1 = A \cdot (\mathbf{1} - I)$
for i in (2: K):
    $diag\ (C_{i-1}) = \mathbf{0}$
    $C_i = A \cdot C_{i-1}$
$C = \mathbf{1} - C_i$
```

Objective function

Let us denote t_{ij} as the number of times node j is present in the path started from node i, then t_{ij} is sampled from a Bernoulli distribution with probability C_{ij}. Thus, C_{ij} can also be viewed as the expected number of times node j is present if a k-step walk is started from node i, which is:

$$C_{ij} = \ Pr\left(t_{ij} = 1\right) = E\left(t_{ij}\right) \tag{6}$$

We further denote R as a subset of nodes, and t_{iR} as the total number of nodes of R present in the walk:

$$t_{iR} = \sum\nolimits_{j \in R} t_{ij} \tag{7}$$

According to linearity of expectation, we can derive that:

$$\mathrm{E}(t_{iR}) = \sum\nolimits_{j \in R} C_{ij} \tag{8}$$

We can further generalize the equation by denoting t_{QR} as the total number of nodes in R present in a walk started from a node in Q. A walk is started from a node i in R for W_i times. From the law of total expectation, we can derive that:

$$E\left(t_{QR}\right) = \sum\nolimits_{i \in R} \sum\nolimits_{j \in Q} W_j C_{ij} \tag{9}$$

Assuming W_j =1 for every j, we developed two measures to capture the degree of isolation of a subset R. One is retention:

$$retention = \frac{E(t_{RR})}{E(t_{RG})} = \frac{\sum_{i \in R} \sum_{j \in R} C_{ij}}{\sum_{i \in R} \sum_{j \in G} C_{ij}} \tag{10}$$

where G is the subset for all the nodes within the graph, t_{RR} is the expected number of nodes of R visited in the k-step walks started from each node in R once, t_{RG} is the expected total number of nodes of G visited in the k-step walks started from each node in R once. The higher retention, the more likely walkers started in R will stay in R.

The other is:

$$exclusivity = \frac{E(t_{RR})}{E(t_{GR})} = \frac{\sum_{i \in R} \sum_{j \in R} C_{ij}}{\sum_{i \in G} \sum_{j \in R} C_{ij}} \tag{11}$$

where t_{GR} is the expected total number of nodes of R visited in the k-step walks started from all the nodes in G once. The higher exclusivity, the less likely walkers outside R will get in.

Combining these two measures, the objective function, named isolation in this study, is (Fig. 13):

$$\begin{aligned} isolation_{RR} &= \frac{E(t_{RR})}{E(t_{RG}) + E(t_{GR})} \\ &= \frac{\sum_{i \in R} \sum_{j \in R} C_{ij}}{\sum_{i \in R} \sum_{j \in G} C_{ij} + \sum_{i \in G} \sum_{j \in R} C_{ij}} \end{aligned} \tag{12}$$

Optimization procedures

To identify clusters with maximal isolation, we adopted a greedy approach iterating between two phases.

One is expansion. In the expansion phase, isolation is calculated for each individual node outside the cluster:

$$isolation_{iR} = \frac{\sum_{j \in R} C_{ij} + \sum_{j \in R} C_{ji}}{\sum_{j \in G} C_{ij} + \sum_{j \in G} C_{ji}} \tag{13}$$

Top 10 nodes with $isolation_{iR}$ higher than the original cluster are added into the cluster.

The other is shrinking. In this phase, isolation is calculated for each individual node within the cluster. All the nodes with $isolation_{iR}$ lower than original cluster are removed from the cluster.

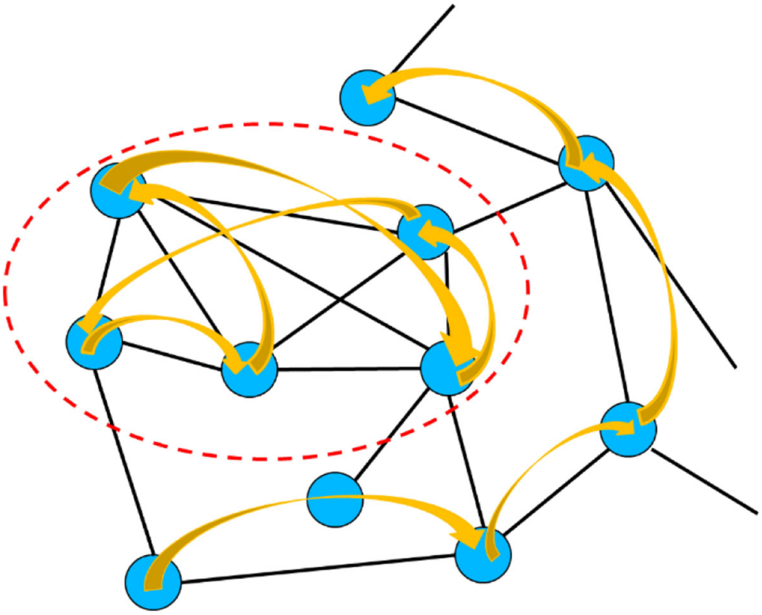

Fig. 13 Illustration of the intuition of the objective function. Nodes within the red dotted circle would be a region with high isolation since walkers inside are likely to stay within and walkers outside are unlikely to get in

The algorithm keeps iterating between expansion and shrinking until there is no more qualified nodes for expansion.

Proof of convergence

For expansion, let us denote the set of qualified nodes as X and the resulting cluster as R'. For each node i within X, $isolation_{iR} > isolation_{RR}$. In other words:

$$\frac{\sum_{i\in X}\sum_{j\in R}C_{ij}+\sum_{i\in X}\sum_{j\in R}C_{ji}}{\sum_{i\in X}\sum_{j\in G}C_{ij}+\sum_{i\in G}\sum_{j\in X}C_{ij}} > \frac{E(t_{RR})}{E(t_{RG})+E(t_{GR})}$$

Thus:

$$\frac{E(t_{RR})+\sum_{i\in X}\sum_{j\in R}C_{ij}+\sum_{i\in X}\sum_{j\in R}C_{ji}}{E(t_{RG})+E(t_{GR})+\sum_{i\in X}\sum_{j\in G}C_{ij}+\sum_{i\in G}\sum_{j\in X}C_{ij}} > \frac{E(t_{RR})}{E(t_{RG})+E(t_{GR})}$$

On the other hand,

$$isolation_{R'R'} = \frac{E(t_{RR})+\sum_{i\in X}\sum_{j\in R}C_{ij}+\sum_{i\in X}\sum_{j\in R}C_{ji}+E(t_{XX})}{E(t_{RG})+E(t_{GR})+\sum_{i\in X}\sum_{j\in G}C_{ij}+\sum_{i\in G}\sum_{j\in X}C_{ij}}$$
$$> \frac{E(t_{RR})+\sum_{i\in X}\sum_{j\in R}C_{ij}+\sum_{i\in X}\sum_{j\in R}C_{ji}}{E(t_{RG})+E(t_{GR})+\sum_{i\in X}\sum_{j\in G}C_{ij}+\sum_{i\in G}\sum_{j\in X}C_{ij}}$$

Hence $isolation_{R'R'} > isolation_{RR}$ after expansion.

Similarly, increase of isolation after shrinking can be proved. Thus, our objective function, isolation, is always increasing during iterations and convergence is guaranteed.

Box 2 Clustering Algorithm

```
function IsolationOptimization(C);
Input: The matrix C
Output: The list of tuples of index
let R be an empty list
let S be a set of indexes of all the nodes in C
Sort S in the descending order of RowSum(C)
while S != Null do
  region = S.pop()
  candidates = expand(C, region)
  while candidates != Null do
    region = region.add(candidates)
    region = shrink(C, region)
    candidates = expand(C,region)
  end while
  R.append(region)
  S = SetDifference(S, region)
end while
return (R)
end
```

Merging overlapped clusters

Highly overlapping clusters are likely to exist for this method. Additionally, for integrated networks, duplicate gene IDs in the same cluster need to be removed. Therefore, overlapping among clusters were evaluated by Jaccard coefficients:

$$overlap\left(C_i, C_j\right) = \frac{\left|C_i \cap C_j\right|}{\left|C_i \cup C_j\right|} \tag{14}$$

where C_i was the ith cluster, $|C_i|$ was the number of genes in C_i. $C_i \cap C_j$ was the intersection of C_i and C_j, and $C_i \cap C_j$ is the union of C_i and C_j. A graph with clusters as nodes was constructed. There is an edge between cluster i and j if $overlap(C_i, C_j) > 0.8$. Collections of highly overlapping clusters is identified as a connected component of the graph. Union and intersection of all the highly overlapping clusters is computed and added into the collection as new clusters. For each collection, the cluster with highest isolation will remain while all the others will be removed.

Evaluation

Metrics

Three sets of measures were adopted in this study to evaluate the clustering performance: (1) Geometric accuracy [21], PPV [21], sensitivity [21]; (2) F measure [63], precision [63], and recall [63]; (3) protein coverage rates, which is the number of unique proteins included by the clustering methods over the total number of proteins of the interactome.

Precision, recall and F-measure

Let P denote the sets of complexes predicted by a computational method and B the real ones in the gold standard. The neighborhood affinity score NA(p, b) between the element p in P, V_p, and the element b in B is defined as:

$$NA(p, b) = \frac{\left|V_p \cap V_b\right|^2}{\mid V_p \mid \times \mid V_b \mid}$$

where V_p is the set of vertexes in the predicted subnetwork p. When NA(p,b) > 0.2, we say p matches b, or vice versa. Thus precision and recall can be defined as:

$$Precision = \frac{N_{cp}}{\mid P \mid}$$

$$Recall = \frac{N_{cb}}{|B|}$$

where N_{cp} is the number of elements in P that matches at least one element in B and N_{cb} the number of elements in B that matches at least one element in P. F-measure, or F1, can then be defined as:

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

Sensitivity, positive predictive value (PPV) and geometric accuracy

With the notations above, let's denote the overlap between p and b as:

$$T_{pb} = \left|V_p \cap V_b\right|$$

Sensitivity and PPV are then defined as:

$$Sensitivity = \frac{\sum_b^B \max_p \left\{T_{pb}\right\}}{\sum_b^B \mid V_b \mid}$$

$$PPV = \frac{\sum_p^P \max_b \left\{T_{pb}\right\}}{\sum_p^P \mid V_p \mid}$$

As a summary metric, geometric accuracy, or simply accuracy, is defined as:

$$Accuracy = \sqrt{Sensitivity \times PPV}$$

Protein coverage

Denote V_P as:

$$V_P = Union\left(\left\{V_p | p \in P\right\}\right)$$

Then protein coverage can be defined as:

$$Coverage = \frac{\mid V_P \mid}{N}$$

where N is the total number of proteins in the interactome.

Experiment datasets

Three types of data were collected, including protein-protein interactions, transcriptomic profiles, and research paper titles and abstracts from NCBI. BioPlex 2.0 [55] was used as the protein interaction network. RNA-seq data about 1097 breast cancer patients stored in TCGA [64] was used as the transcriptomic profile data. For the application on yeast interactome, protein interaction data was collected from BioGrid (version 3.4.162). Microarray datasets about systematic perturbation in yeast [25, 65] was also used in this study. We download over one million paper titles and abstracts from NCBI on March 23rd, 2016 for yeast and on September 25th, 2017 for human.

Gold standard

Gene Ontology (GO) and manually curated database of protein complexes were used as gold standard. The GO annotation for *Homo sapiens* was downloaded from Gene Ontology consortium. This file annotated 19,473 gene products and was submitted on September 26, 2017. The GO annotation file for *Saccharomyces cerevisiae* was downloaded from Saccharomyces Genome Database (SGD). This file annotated 6357 gene products with 4393 GO terms, 2104 of which are in the category of biological process. The list of 408 yeast protein complex was derived from CYC2008 [66], covering 1627 unique genes. Human protein complexes were downloaded from CORUM [67], with 2693 complexes and 4413 unique genes.

Control methods

We used Infomap [68], Markov Clustering algorithm (MCL), MCODE, Walktrap [69], NCMine [13] and ClusterOne [70] as the methods for comparison. These algorithms have been widely used to identify functional modules or protein complexes in protein interaction networks. We implemented MCL based on the work of Enright [9], used igraph to run Infomap and Walktrap, and used Cytoscape to run ClusterOne, NCMine and MCODE.

Since multiplex is applicable to most clustering methods based on random walk. Performance of MCL, Infomap, and Walktrap were evaluated based on their results on multiplex. For density-based clustering methods (i.e. MCODE, ClusterOne) and core-attachment approaches (i.e. NCMine), clustering results of the single layer with better performance were shown and compared.

In addition, we compared the multiplex approach with Similarity Network Fusion (SNF). The aggregated matrix generated by SNF is fed to the Isolation clustering algorithm. Our preliminary results showed that the network transformation step in the clustering procedure yields a uniform distribution of edges weights for most nodes. Therefore, the performance of SNF shown in the result section was generated by conducting merely the optimization step in the isolation algorithm.

Abbreviations

GO: Gene Ontology; NCBI: National Center for Biotechnology Information; nHDP: Nested hierarchical dirichlet process; PPI: Protein-protein interactions; PPV: Positive predictive value; SNF: Similarity network fusion; TCGA: The Cancer Genome Atlas

Acknowledgements

Not applicable.

Funding

This work was supported by U.S. National Library of Medicine (Grant R00LM011673 to Dr. Songjian Lu) and National Human Genome Research Institute (Grant U54HG008540 to Xinghua Lu). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

Source codes and the data generated by our methods in different stages were available through https://github.com/LifanLiang/Isolaton-clustering. As for the gold standard used in this study, CYC2008 can be obtained through http://wodaklab.org/cyc2008/downloads; CORUM through http://mips.helmholtz-muenchen.de/corum/#download; GO annotation file for yeast and human through http://www.geneontology.org/page/download-go-annotations.

Authors' contributions

LL developed and implemented the isolation clustering algorithm; SL conceived, supervised this study, and conducted differential expression analysis; VC and XL provided the topic modeling of genes; ZJ provided biological interpretation of clustering results; XF assisted in developing the multiplex framework. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA. [2]Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc, Frederick, USA. [3]Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an 710072, Shanxi, China. [4]Clinical Medicine Research Institute, Jinan University, Guangzhou 51063, Guangdong, China.

Received: 11 December 2018 Accepted: 9 April 2019
Published online: 02 May 2019

References

1. Choobdar S, Ahsen ME, Crawford J, Tomasoni M, Lamparter D, Lin J, et al. Open community challenge reveals molecular Network modules with key roles in diseases. BioRxiv. 2018. https://doi.org/10.1101/265553.
2. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402(6761 Suppl):C47–52. https://doi.org/10.1038/35011540.
3. Tornow S, Mewes HW. Functional modules by relating protein interaction networks and gene expression. Nucleic Acids Res. 2003;31:6283–9. https://doi.org/10.1093/nar/gkg838.
4. Ji J, Zhang A, Liu C, Quan X, Liu Z. Survey: functional module detection from protein-protein interaction networks. IEEE Trans Knowl Data Eng. 2014;26:261–77. https://doi.org/10.1109/TKDE.2012.225.
5. Zinman GE, Zhong S, Bar-Joseph Z. Biological interaction networks are conserved at the module level. BMC Syst Biol. 2011;5:134. https://doi.org/10.1186/1752-0509-5-134.
6. Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003;4:2.
7. Natarajan M, Lin K-M, Hsueh RC, Sternweis PC, Ranganathan R. A global analysis of cross-talk in a mammalian cellular signalling network. Nat Cell Biol. 2006;8:571–80. https://doi.org/10.1038/ncb1418.
8. King AD, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering. Bioinformatics. 2004;20:3013–20. https://doi.org/10.1093/bioinformatics/bth351.
9. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30:1575–84. https://doi.org/10.1093/nar/30.7.1575.
10. Maruyama O, Chihara A. NWE: node-weighted expansion for protein complex prediction using random walk distances. Proteome Sci. 2011;9(Suppl 1):S14. https://doi.org/10.1186/1477-5956-9-S1-S14.

11. Macropol K, Can T, Singh AK. RRW: repeated random walks on genome-scale protein networks for local cluster discovery. BMC Bioinformatics. 2009; 10:283. https://doi.org/10.1186/1471-2105-10-283.
12. Pellegrini M, Baglioni M, Geraci F. Protein complex prediction for large protein protein interaction networks with the Core&Peel method. BMC Bioinformatics. 2016;17(Suppl 12):372. https://doi.org/10.1186/s12859-016-1191-6.
13. Tadaka S, Kinoshita K. NCMine: Core-peripheral based functional module detection using near-clique mining. Bioinformatics. 2016;32:3454–60. https://doi.org/10.1093/bioinformatics/btw488.
14. Wu M, Li X, Kwoh C-K, Ng S-K. A core-attachment based method to detect protein complexes in PPI networks. BMC Bioinformatics. 2009;10:169. https://doi.org/10.1186/1471-2105-10-169.
15. He T, Chan KCC. Evolutionary graph clustering for protein complex identification. IEEE/ACM Trans Comput Biol Bioinform. 2018;15:892–904. https://doi.org/10.1109/TCBB.2016.2642107.
16. Ramadan E, Naef A, Ahmed M. Protein complexes predictions within protein interaction networks using genetic algorithms. BMC Bioinformatics. 2016;17(Suppl 7):269. https://doi.org/10.1186/s12859-016-1096-4.
17. Xu B, Li K, Zheng W, Liu X, Zhang Y, Zhao Z, et al. Protein complexes identification based on go attributed network embedding. BMC Bioinformatics. 2018;19:535. https://doi.org/10.1186/s12859-018-2555-x.
18. Wang Y, Qian X. Finding low-conductance sets with dense interactions (FLCD) for better protein complex prediction. BMC Syst Biol. 2017;11(Suppl 3):22. https://doi.org/10.1186/s12918-017-0405-5.
19. Wang R, Liu G, Wang C, Su L, Sun L. Predicting overlapping protein complexes based on core-attachment and a local modularity structure. BMC Bioinformatics. 2018;19:305. https://doi.org/10.1186/s12859-018-2309-9.
20. Bader JS, Chaudhuri A, Rothberg JM, Chant J. Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol. 2004;22:78–85. https://doi.org/10.1038/nbt924.
21. Li X, Wu M, Kwoh C-K, Ng S-K. Computational approaches for detecting protein complexes from protein interaction networks: a survey. BMC Genomics. 2010;11(Suppl 1):S3. https://doi.org/10.1186/1471-2164-11-S1-S3.
22. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003;302:249–55. https://doi.org/10.1126/science.1087447.
23. Shi B, Wang X, Han X, Liu P, Wei W, Li Y. Functional modules analysis based on coexpression network in pancreatic ductal adenocarcinoma. Pathol Oncol Res. 2014;20:293–9. https://doi.org/10.1007/s12253-013-9694-1.
24. You Q, Zhang L, Yi X, Zhang K, Yao D, Zhang X, et al. Co-expression network analyses identify functional modules associated with development and stress response in Gossypium arboreum. Sci Rep. 2016;6:38436. https://doi.org/10.1038/srep38436.
25. Huang S-SC, Fraenkel E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. Sci Signal. 2009;2:ra40. https://doi.org/10.1126/scisignal.2000350.
26. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Biol. 2010;6: e1000662. https://doi.org/10.1371/journal.pcbi.1000662.
27. Dey KK, Hsiao CJ, Stephens M. Visualizing the structure of RNA-seq expression data using grade of membership models. PLoS Genet. 2017;13: e1006599. https://doi.org/10.1371/journal.pgen.1006599.
28. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics. 2008;24:i223–31. https://doi.org/10.1093/bioinformatics/btn161.
29. Keretsu S, Sarmah R. Weighted edge based clustering to identify protein complexes in protein-protein interaction networks incorporating gene expression profile. Comput Biol Chem. 2016;65:69–79. https://doi.org/10.1016/j.compbiolchem.2016.10.001.
30. Zhang Z, Song J, Tang J, Xu X, Guo F. Detecting complexes from edge-weighted PPI networks via genes expression analysis. BMC Syst Biol. 2018; 12(Suppl 4):40. https://doi.org/10.1186/s12918-018-0565-y.
31. Cheng L, Liu P, Wang D, Leung K-S. Exploiting locational and topological overlap model to identify modules in protein interaction networks. BMC Bioinformatics. 2019;20:23. https://doi.org/10.1186/s12859-019-2598-7.
32. Vella D, Marini S, Vitali F, Di Silvestre D, Mauri G, Bellazzi R. MTGO: PPI network analysis via topological and functional module identification. Sci Rep. 2018;8:5499. https://doi.org/10.1038/s41598-018-23672-0.
33. Ou-Yang L, Yan H, Zhang X-F. A multi-network clustering method for detecting protein complexes from multiple heterogeneous networks. BMC Bioinformatics. 2017;18(Suppl 13):463. https://doi.org/10.1186/s12859-017-1877-4.
34. Taghipour S, Zarrineh P, Ganjtabesh M, Nowzari-Dalini A. Improving protein complex prediction by reconstructing a high-confidence protein-protein interaction network of Escherichia coli from different physical interaction data sources. BMC Bioinformatics. 2017;18:10. https://doi.org/10.1186/s12859-016-1422-x.
35. Ma C-Y, Chen Y-PP, Berger B, Liao C-S. Identification of protein complexes by integrating multiple alignment of protein interaction networks. Bioinformatics. 2017;33:1681–8. https://doi.org/10.1093/bioinformatics/btx043.
36. Liu Y, Liang Y, Wishart D. PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic Acids Res. 2015;43:W535–42. https://doi.org/10.1093/nar/gkv383.
37. Chen V, Paisley J, Lu X. Revealing common disease mechanisms shared by tumors of different tissues of origin through semantic representation of genomic alterations and topic modeling. BMC Genomics. 2017;18(Suppl 2): 105. https://doi.org/10.1186/s12864-017-3494-z.
38. Kim J, Kim J-J, Lee H. An analysis of disease-gene relationship from Medline abstracts by DigSee. Sci Rep. 2017;7:40154. https://doi.org/10.1038/srep40154.
39. Yang Z, Yu F, Lin H, Wang J. Integrating PPI datasets with the PPI data from biomedical literature for protein complex detection. BMC Med Genet. 2014; 7(Suppl 2):S3. https://doi.org/10.1186/1755-8794-7-S2-S3.
40. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. Sci Rep. 2018;8:1362. https://doi.org/10.1038/s41598-018-19333-x.
41. Gaudet P, Dessimoz C. Gene ontology: pitfalls, biases, and remedies. Methods Mol Biol. 2017;1446:189–205. https://doi.org/10.1007/978-1-4939-3743-1_14.
42. Ferranti D, Krane D, Craft D. The value of prior knowledge in machine learning of complex network systems. Bioinformatics. 2017;33:3610–8. https://doi.org/10.1093/bioinformatics/btx438.
43. Xu Y, Zhou J, Zhou S, Guan J. CPredictor3.0: detecting protein complexes from PPI networks with expression data and functional annotations. BMC Syst Biol. 2017;11(Suppl 7):135. https://doi.org/10.1186/s12918-017-0504-3.
44. Chen W, Liu J, He S. Prior knowledge guided active modules identification: an integrated multi-objective approach. BMC Syst Biol. 2017;11 Suppl 2:8. https://doi.org/10.1186/s12918-017-0388-2.
45. Networks - Mark Newman - Oxford University Press. https://global.oup.com/academic/product/networks-9780198805090. Accessed 10 Apr 2018.
46. Gligorijevic V, Barot M, Bonneau R. deepNF: deep network fusion for protein function prediction. Bioinformatics. 2018;34:3873–81. https://doi.org/10.1093/bioinformatics/bty440.
47. Ou-Yang L, Wu M, Zhang X-F, Dai D-Q, Li X-L, Yan H. A two-layer integration framework for protein complex detection. BMC Bioinformatics. 2016;17:100. https://doi.org/10.1186/s12859-016-0939-3.
48. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11:333–7. https://doi.org/10.1038/nmeth.2810.
49. Cozzo E, Kivelä M, Domenico MD, Solé-Ribalta A, Arenas A, Gómez S, et al. Structure of triadic relations in multiplex networks. New J Phys. 2015;17: 073029. https://doi.org/10.1088/1367-2630/17/7/073029.
50. Solé-Ribalta A, Gómez S, Arenas A. Congestion induced by the structure of multiplex networks. Phys Rev Lett. 2016;116:108701. https://doi.org/10.1103/PhysRevLett.116.108701.
51. De Domenico M, et al. The physics of spreading processes in multilayer networks. Nat Phys. 2016;12(10):901. https://doi.org/10.1038/nphys3865.
52. Perozzi B, Al-Rfou R, DeepWalk SS. Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining - KDD' ' '14. New York: ACM Press; 2014. p. 701–10. https://doi.org/10.1145/2623330.2623732.
53. Zhou XZ, Lu KP. The Pin2/TRF1-interacting protein PinX1 is a potent telomerase inhibitor. Cell. 2001;107:347–59.
54. Yonekawa T, Yang S, Counter CM. PinX1 localizes to telomeres and stabilizes TRF1 at mitosis. Mol Cell Biol. 2012;32:1387–95. https://doi.org/10.1128/MCB.05641-11.
55. Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, et al. Architecture of the human interactome defines protein communities and disease networks. Nature. 2017;545:505–9. https://doi.org/10.1038/nature22366.

56. Koc EC, Cimen H, Kumcuoglu B, Abu N, Akpinar G, Haque ME, et al. Identification and characterization of CHCHD1, AURKAIP1, and CRIF1 as new members of the mammalian mitochondrial ribosome. Front Physiol. 2013;4:183. https://doi.org/10.3389/fphys.2013.00183.
57. Davies SMK, Rackham O, Shearwood A-MJ, Hamilton KL, Narsai R, Whelan J, et al. Pentatricopeptide repeat domain protein 3 associates with the mitochondrial small ribosomal subunit and regulates translation. FEBS Lett. 2009;583:1853–8. https://doi.org/10.1016/j.febslet.2009.04.048.
58. Schmid SR, Linder P. D-E-A-D protein family of putative RNA helicases. Mol Microbiol. 1992;6:283–91. https://doi.org/10.1111/j.1365-2958.1992.tb01470.x.
59. Paisley J, Wang C, Blei DM, Jordan MI. Nested hierarchical dirichlet processes. IEEE Trans Pattern Anal Mach Intell. 2015;37:256–70. https://doi.org/10.1109/TPAMI.2014.2318728.
60. Identifying Patterns of Cancer Disease Mechanisms by Mining Alternative Representations of Genomic Alterations - D-Scholarship@Pitt. http://d-scholarship.pitt.edu/30319/. Accessed 3 Oct 2018.
61. Jeh G, Widom J. SimRank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining' ' - KDD '02. New York: ACM Press; 2002. p. 538. https://doi.org/10.1145/775047.775126.
62. Boccaletti S, Bianconi G, Criado R, del Genio CI, Gómez-Gardeñes J, Romance M, et al. The structure and dynamics of multilayer networks. Phys Rep. 2014;544:1–122. https://doi.org/10.1016/j.physrep.2014.07.001.
63. Hu AL, Chan KCC. Utilizing both topological and attribute information for protein complex identification in PPI networks. IEEE/ACM Trans Comput Biol Bioinform. 2013;10:780–92. https://doi.org/10.1109/TCBB.2013.37.
64. Network CGAR, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer genome atlas pan-Cancer analysis project. Nat Genet. 2013;45:1113–20. https://doi.org/10.1038/ng.2764.
65. Yeger-Lotem E, Riva L, Su LJ, Gitler AD, Cashikar AG, King OD, et al. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. Nat Genet. 2009;41:316–23. https://doi.org/10.1038/ng.337.
66. Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res. 2009;37:825–31. https://doi.org/10.1093/nar/gkn1005.
67. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res. 2010;38 Database issue:D497–501. https://doi.org/10.1093/nar/gkp914.
68. Rosvall M, Axelsson D, Bergstrom CT. The map equation. Eur Phys J Spec Top. 2009;178:13–23. https://doi.org/10.1140/epjst/e2010-01179-1.
69. Pons P, Latapy M. Computing communities in large networks using random walks. In: pInar Y, Güngör T, Gürgen F, Özturan C, editors. Computer and information sciences - ISCIS 2005. Berlin: Springer Berlin Heidelberg; 2005. p. 284–93. https://doi.org/10.1007/11569596_31.
70. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods. 2012;9:471–2. https://doi.org/10.1038/nmeth.1938.