

RESEARCH

Open Access



ClusterM: a scalable algorithm for computational prediction of conserved protein complexes across multiple protein interaction networks

Yijie Wang^{1†}, Hyundoo Jeong^{2†}, Byung-Jun Yoon^{3,4,5} and Xiaoning Qian^{3,4*}

From The 18th Asia Pacific Bioinformatics Conference
Seoul, Korea. 18–20 August 2020

Abstract

Background: The current computational methods on identifying conserved protein complexes across multiple Protein-Protein Interaction (PPI) networks suffer from the lack of explicit modeling of the desired topological properties within conserved protein complexes as well as their scalability.

Results: To overcome those issues, we propose a scalable algorithm—ClusterM—for identifying conserved protein complexes across multiple PPI networks through the integration of network topology and protein sequence similarity information. ClusterM overcomes the computational barrier that existed in previous methods, where the complexity escalates exponentially when handling an increasing number of PPI networks; and it is able to detect conserved protein complexes with both topological separability and cohesive protein sequence conservation. On two independent compendiums of PPI networks from *Saccharomyces cerevisiae* (Sce, yeast), *Drosophila melanogaster* (Dme, fruit fly), *Caenorhabditis elegans* (Cel, worm), and *Homo sapiens* (Hsa, human), we demonstrate that ClusterM outperforms other state-of-the-art algorithms by a significant margin and is able to identify *de novo* conserved protein complexes across four species that are missed by existing algorithms.

Conclusions: ClusterM can better capture the desired topological property of a typical conserved protein complex, which is densely connected within the complex while being well-separated from the rest of the networks. Furthermore, our experiments have shown that ClusterM is highly scalable and efficient when analyzing multiple PPI networks.

Keywords: Comparative network analysis, Multiple network alignment and clustering, Conserved module identification

*Correspondence: xqian@ece.tamu.edu

[†]Yijie Wang and Hyundoo Jeong contributed equally to this work.

³Department of Electrical and Computer Engineering, Texas A&M University, College Station 77843, TX, USA

⁴TEES-AgrLife Center for Bioinformatics and Genomic Systems Engineering (CBGSE), Texas A&M University, College Station 77843, TX, USA

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. 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 in a credit line to the data.

Background

Advanced high-throughput technologies for measuring protein interactions [1, 2] have provided researchers with rich information about protein-protein interactions (PPI) in various species [3–5]. In order to translate such information into meaningful biological knowledge about the underlying cellular functions and evolutionary mechanisms, one arising computational challenge is how to integrate these PPI data with other available data—such as sequence data—to identify conserved protein complexes that have similar cellular functions across multiple species.

Intuitively, conserved protein complexes should have the following properties with respect to their network topology and sequence homology. Within each species, proteins within a protein complex is densely connected to each other while being loosely connected to and well separated from the rest of the PPI networks [6]. Across different species, conserved protein complexes should contain many orthologs with high sequence similarity. Therefore, it appears reasonable to expect that the problem of identifying such complexes could be effectively addressed by a comparative network analysis approach.

In recent years, efforts have been made to develop computational techniques for the comparison of PPI networks across different species through global network alignment [7–12]. Most of the existing global network alignment algorithms aim to identify the one-to-one mapping with the maximum total similarity of aligned proteins across networks. However, this approach may not directly lead to accurate identification of conserved protein complexes, where an important focus is on grouping proteins that work together towards similar functionalities across species. Furthermore, proteins in conserved protein complexes often have many-to-many orthologous relationships. Local network alignment [13–18] searches for conserved subnetworks across species, which better resembles the task of identifying conserved protein complexes. However, many of them focus on the topological criteria based on conserved edges or other network motifs with specific topological structures, often motivated by conjectured evolutionary or functional models [14, 18, 19]. To the best of our knowledge, none of the existing local alignment algorithms explicitly considers the characteristic topological properties of protein complexes, in which proteins within the complex highly interact with each other but rarely interact with the rest of the network. Therefore, directly applying existing local network alignment algorithms may not result in accurate detection of conserved protein complexes with the maximum coverage of the given PPI networks. Furthermore, both local and global network alignment problems essentially reduce to the (sub)graph isomorphism problem and the alignment results tend to be sensitive to topological errors

in the PPI networks. This is certainly problematic, since currently available PPI networks may contain a significant number of false positive interactions while many true interactions are still missing. In comparison, detecting conserved protein complexes is generally more robust to similar topological errors, as we focus on interaction density within complexes and their topological separability (i.e., whether they are well separated from the rest of the PPI networks). In other words, the focus lies on groups of proteins that may potentially belong to the same complex rather than individual proteins, which makes the overall prediction less sensitive to errors in the PPI networks.

In addition to the lack of explicit modeling of the desired topological properties within conserved protein complexes, most of the existing network alignment and clustering algorithms [13–15, 17, 18] do not scale well with the number of species and the network size. For example, it is prohibitive for the network alignment algorithms in [13–15, 17, 18] to handle more than three PPI networks due to the exponential growth of the alignment graph with the number of species and the network size. Finally, to the best of our knowledge, currently, there is neither a gold standard for conserved protein complexes that contains protein complexes from multiple species nor commonly accepted metrics for assessing the performance of algorithms for predicting conserved protein complexes. As a consequence, it has been practically difficult to effectively evaluate the capability of such algorithms to unveil “true” conserved protein complexes, just based on the fact that the identified conserved protein complexes may overlap with well-known protein complexes in another species [13, 14, 17, 18].

To fill these critical gaps in a conserved protein complex identification, we propose a scalable algorithm—ClusterM—that explicitly characterizes the desired topological separability of protein complexes and also incorporates sequence similarity of proteins across the given PPI networks. ClusterM consists of three major steps. The first step is to find a set of protein spines (sets of proteins, one from each network) across different PPI networks. In the second step, well-separated subnetworks around proteins in each protein spine are identified. The final step is to look for conserved subnetworks within those identified well-separated subnetworks, which have both cohesive protein-protein interaction similarity and sequence similarity.

Last but not least, in order to evaluate and compare the performance of ClusterM and other existing algorithms, we have curated a new yeast-human reference conserved protein complex dataset based on yeast and human gold-standard complexes and propose effective evaluation metrics based on the existing measures for assessing protein complex identification algorithms for individual PPI networks. Experimental results based on

comparative analysis of yeast and human PPI networks show that ClusterM clearly outperforms the current state-of-the-art algorithms.

Results

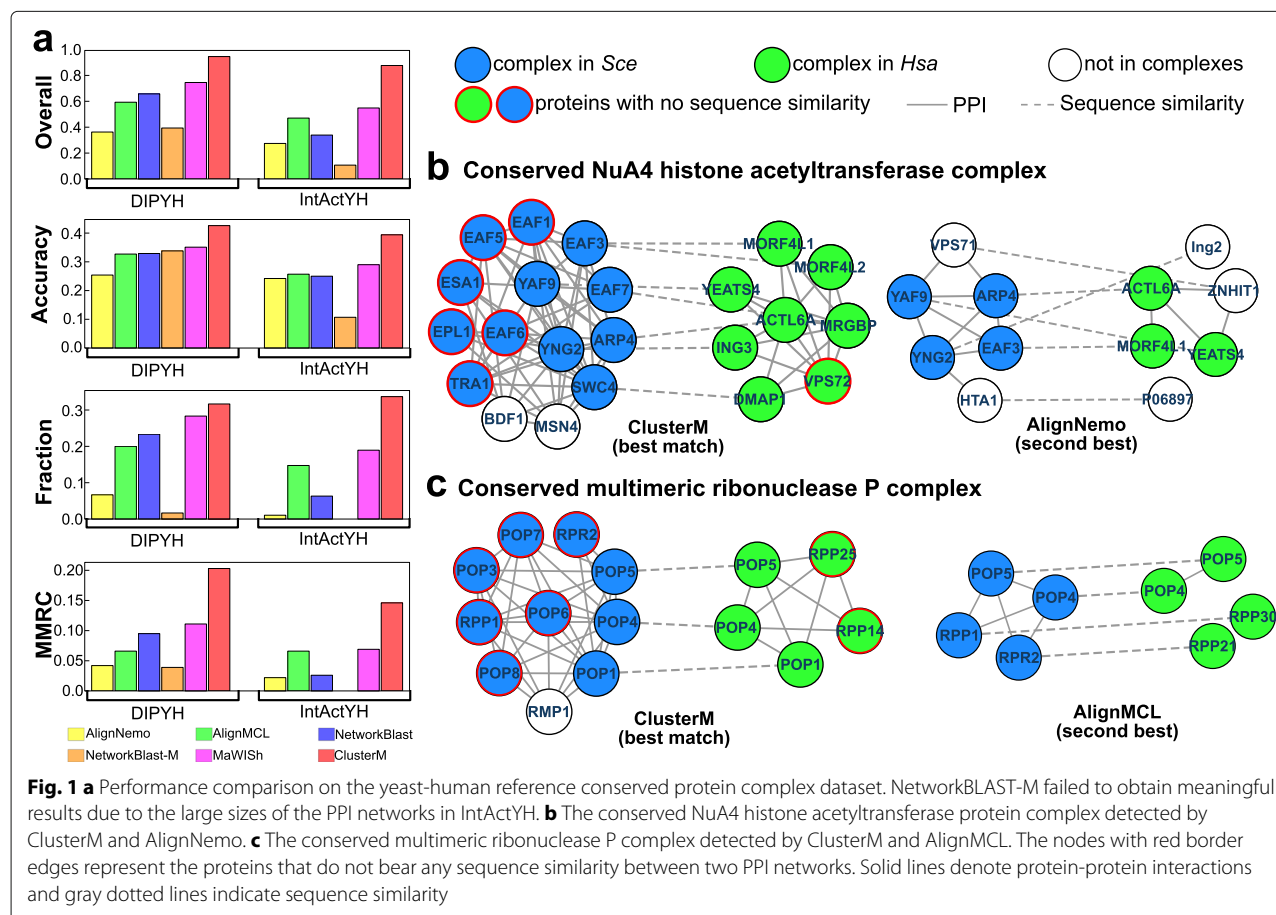
Identifying conserved complexes in yeast and human networks

We first evaluated ClusterM on the yeast and human PPI networks based on two constructed datasets: DIPYH and IntActYH. In the DIPYH dataset, yeast (*Sce*) and human (*Hsa*) PPI networks are obtained from Database of Interacting Proteins (DIP) version 20150101 [3] and the protein sequence similarity between proteins across networks are computed by BLAST [20]. The human PPI network has 4,278 proteins and 6,446 interactions, and the yeast PPI network has 5,138 proteins and 22,835 interactions. Similarly, the IntActYH dataset contains yeast and human PPI networks extracted from the IntAct database version 20150120 [4] and the corresponding protein sequence similarity across networks. The human PPI network has 23,246 proteins and 106,031 interactions, and the yeast PPI network has 6,392 proteins and 78,287 interactions. We compared ClusterM with the state-of-the-art

algorithms—AlignNemo [17], AlignMCL [21], MaWISH [14], NetworkBLAST [22], and NetworkBLAST-M [23]. The selection of parameters of different methods is discussed in the Methods section. We assessed the quality of the identified conserved protein complexes based on the yeast-human reference conserved protein complexes generated from hand-curated yeast and human protein complex data obtained from CYC2008 [24] and the Comprehensive Resource of Mammalian protein complexes (CORUM) [25]. The detailed description of the reference complexes can be found in the Methods section.

In this study, we define a composite score for performance evaluation, which consists of (i) the fraction of the reference conserved protein complexes matched by at least one identified conserved protein complex, (ii) the accuracy score [6], and (iii) the maximum matching ratio for conserved complexes (MMRC). The description of these three evaluation metrics is detailed in the Methods section.

As shown in Fig. 1a, ClusterM outperformed all competing algorithms by a large margin on every evaluation metric for both datasets. Figure 1b and c illustrate two conserved protein complexes identified by ClusterM that



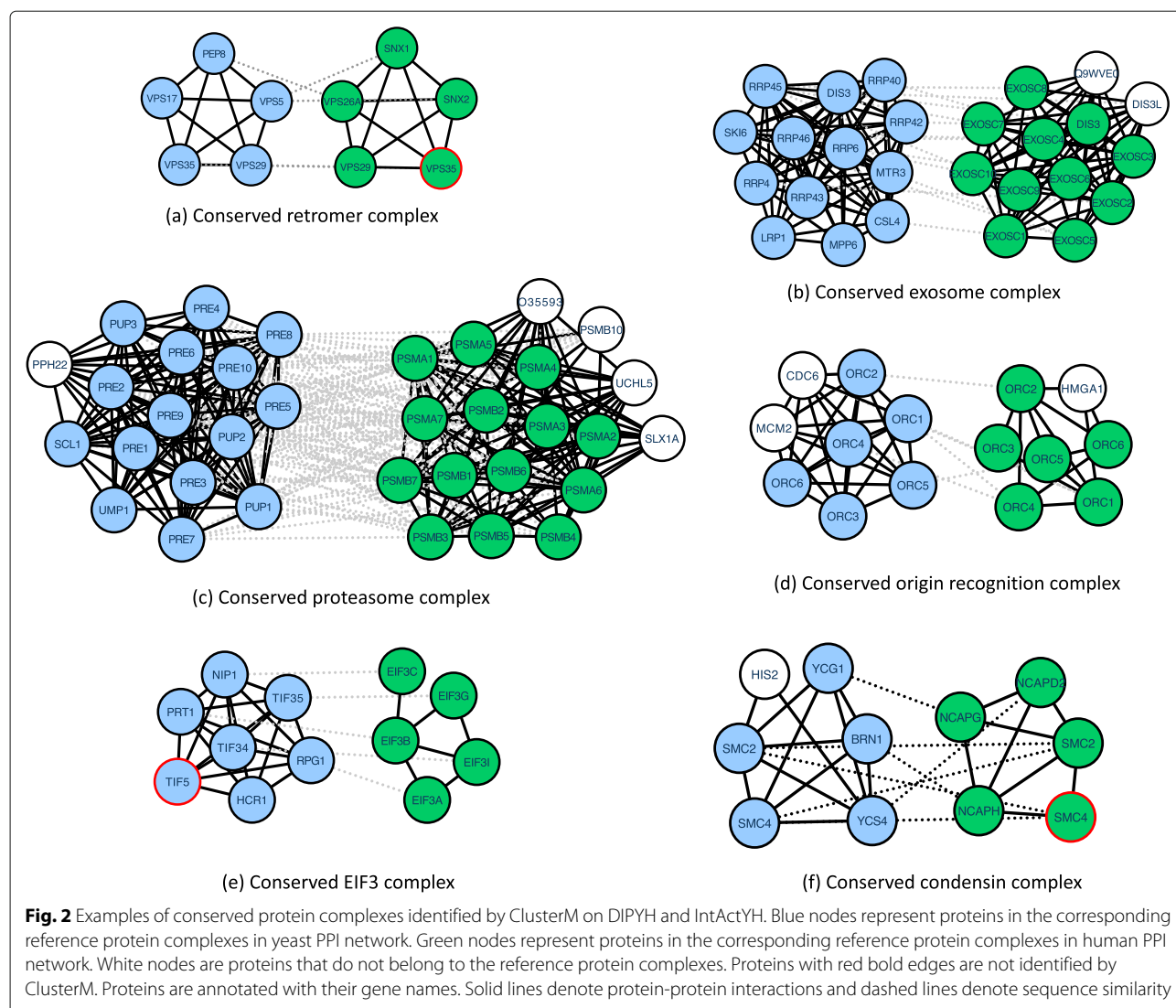
cannot be fully recovered by other methods. Clearly, ClusterM utilizes the network topology to facilitate the detection of conserved protein complexes without considering the restriction shared by most of the algorithms (AlignNemo, AlignMCL, NetworkBLAST, and MaWISH), which is only proteins with similar sequences across species can exist in the predicted conserved complexes. NetworkBLAST-M does not have such a restriction but failed to characterize the property of the conserved complexes.

We discuss two specific conserved protein complexes in Figs. 1b and c in detail to highlight the advantages of ClusterM over other competing methods. In Fig. 1b, there are a relatively larger number of unidentified proteins in the predicted conserved protein complex detected by AlignNemo. The precision of the identified complex is also small as it includes a larger number of irrelevant proteins such as Ing2, P06897, and ZNHIT1 in the conserved protein complex in human PPI network and VPS71 and HTA1 in yeast PPI network. However, ClusterM correctly predicted a larger number of proteins in the NuA4 histone acetyltransferase protein complex with a higher precision. One possible explanation for the significant difference is that ClusterM can correctly identify the proteins without a sequence similarity by effectively utilizing the sequence information and topological structure of PPI networks. On the contrary, AlignNemo failed to identify the proteins in the candidate complexes if they do not have corresponding homology with high sequence similarity. This illustration clearly shows the distinctive advantages of ClusterM over the competing methods to identify conserved protein complex across different PPI networks. In Fig. 1c, AlignMCL also failed to identify the proteins without homology information. Moreover, the predicted conserved complex in human PPI network is ill-connected. That is, only POP4 and POP5 are connected and RPP21 and RPP30 are not connected. This illustration shows that AlignMCL could have a limitation to effective utilization of the topological separability of the conserved protein complexes and AlignMCL can fail to identify the inserted (or deleted) proteins in the conserved complex if there are insufficient or noisy sequence similarity information. In Fig. 2, we show other representative examples of conserved protein complexes detected by ClusterM, which cannot be correctly identified by other algorithms. We have also applied Gene Ontology (GO) enrichment analysis to every identified conserved complexes and the results indicate that ClusterM achieves remarkably higher coverage (i.e., total number of proteins in the predicted complexes) than all competing methods and more than 90% of the predicted complexes identified by ClusterM are significantly enriched in certain GO terms (detailed results and discussions can be found in Supplementary Materials).

Robustness analysis

To investigate the degree to which the competing algorithms are sensitive to small changes in the input so that meaningful results can be derived even with noisy or incomplete data, we performed a robustness analysis of the algorithms considered in this study. For each algorithm, we optimized the parameters based on the original DIPYH dataset (yeast and human PPI from DIP database and the protein sequence similarity across species) and tested the performance on a perturbed DIPYH dataset, in which we re-wired 10% edges in both yeast and human PPI networks through the M-P procedure [26] to instill topological noise. Furthermore, we treated the protein sequence similarity between yeast and human as a bipartite network, and thereby perturbed 10% of the homologous relationships (i.e., a sequence similarity) also using the M-P procedure [26]. To remove the randomness in the M-P procedure, we generated 10 perturbed networks for each test case and reported the averaged scores.

Figure 3 shows the evaluation results—i.e., fraction, accuracy, and MMRC scores—based on four datasets: the noise-free dataset, the dataset with only topological noise, the dataset with only homology noise, and the dataset with both topological and homology noise. Each table corresponds to each dataset in Fig. 3. We first computed the composite score by summing over the fraction, accuracy and MMRC scores. Then, we computed the mean and standard deviation of the four composite scores. The mean of the composite score for ClusterM, MaWISH, NetworkBLAST (NB), AlignMCL, NetworkBLAST-M (NBM), and AlignNemo were 0.890, 0.588, 0.552, 0.486, 0.405, and 0.368, respectively. The standard deviation of the composite score for ClusterM, MaWISH, NB, AlignMCL, NBM, and AlignNemo were 0.066, 0.075, 0.073, 0.079, 0.047, and 0.053, respectively. We can see that ClusterM achieves the largest mean and the third smallest standard deviation among all compared algorithms, which implies that ClusterM is relatively less sensitive to small changes in the input data as well as the parameter selection. Although AlignNemo achieved the smallest standard deviation, it showed the smallest composite score as well and there is a clear gap to the composite score of ClusterM. MaWISH attained the second best composite score for the noise-free dataset, but for noisy datasets, its performance was unstable. This unstable behavior may be due to its dependence on the choice of the seven parameters. Even for relatively small changes in the input, those seven parameters may have to be simultaneously (and significantly) changed to obtain the best performance, which may make it practically difficult to find the optimal parameters for large datasets in a robust manner. NB and NBM showed similar performance in term of the changes of the composite scores, and they turned out to be more sensitive to noise compared to ClusterM but less sensitive when



compared to MaWISH. These results show that ClusterM is the most user-friendly algorithm among the compared methods, considering that it is robust to noise and that it does not require the joint optimization of multiple parameters to achieve good performance.

Conserved protein complex identification for multiple networks

In this section, we present and discuss a specific example that demonstrates the potential of ClusterM in identifying conserved protein complexes in multiple (more than two) PPI networks. We have constructed two PPI datasets for benchmarking protein complex identification across four PPI networks: DIPPIs and IntActPPIs, based on the corresponding PPI networks and protein amino acids sequences for human (*Hsa*), yeast (*Sce*), fly (*Dme*), and worm (*Cel*) from DIP (version 20150101) and IntAct (version 20150120). The detailed description to construct PPI

networks can be found in the Method section. Table 1 provides a basic statistical summary and the abbreviation of PPI networks. In Fig. 4a, we show the conserved proteasome core complex identified by ClusterM in IntActPPIs. This complex was completely missed by all other algorithms that were considered in this study. Proteins in light blue, green, pink, and purple represent proteins in yeast, human, fly, and worm, respectively. We further tested the statistical significance of each of the four protein complexes in the respective organisms based on the GO term annotated to proteasome core complex (GO:0005839). The *p*-values associated to the corresponding yeast, human, fly, and worm complexes are 5.58E-41, 2.80E-48, 4.88E-16, and 5.48E-9, respectively. In Fig. 4b, the spliceosomal complex (GO:0005681) is significantly enriched in the yeast, human, fly, and worm conserved complex with the *p*-values 2.37E-08, 4.16E-12, 6.03E-11, and 2.97E-08, respectively. Note that red

	Frac	Acc	MMRC		Frac	Acc	MMRC	
MaWISH	0.283	0.351	0.111		0.147	0.301	0.077	No homology noise
ClusterM	0.317	0.426	0.203		0.298	0.42	0.164	
NB	0.233	0.329	0.095		0.18	0.314	0.075	
NBM	0.017	0.338	0.039		0.007	0.330	0.038	
AlignMCL	0.200	0.327	0.066		0.115	0.303	0.064	
AlignNemo	0.067	0.254	0.042		0.058	0.296	0.04	
MaWISH	0.197	0.322	0.095		0.198	0.322	0.090	10% homology noise
ClusterM	0.33	0.426	0.188		0.28	0.416	0.14	
NB	0.183	0.322	0.082		0.132	0.30	0.062	
NBM	0.032	0.328	0.034		0.033	0.312	0.034	
AlignMCL	0.147	0.337	0.062		0.085	0.29	0.047	
AlignNemo	0.053	0.283	0.031		0.033	0.28	0.03	
No topological noise					10% topological noise			

Fig. 3 Robustness analysis on yeast and human PPI networks from DIP dataset. We display the fraction, accuracy and MMRC scores for dataset with different noise level. There are four small tables in the figure. The top-left table shows scores for noise free data. The top-right table exhibits scores for data with 10% topological noise but 0% homology noise. The bottom-left table shows scores for 10% homology noise but 0% topological noise. The bottom-right table displays scores for data with 10% topological noise but 10% homology noise

solid lines are PPIs missed in IntAct database but we can identify the interactions in STRING database [27]. The enrichment analysis of these conserved protein complexes, which could be detected solely by ClusterM, clearly shows that ClusterM can effectively mine biologically meaningful protein complexes in multiple large-scale PPI networks.

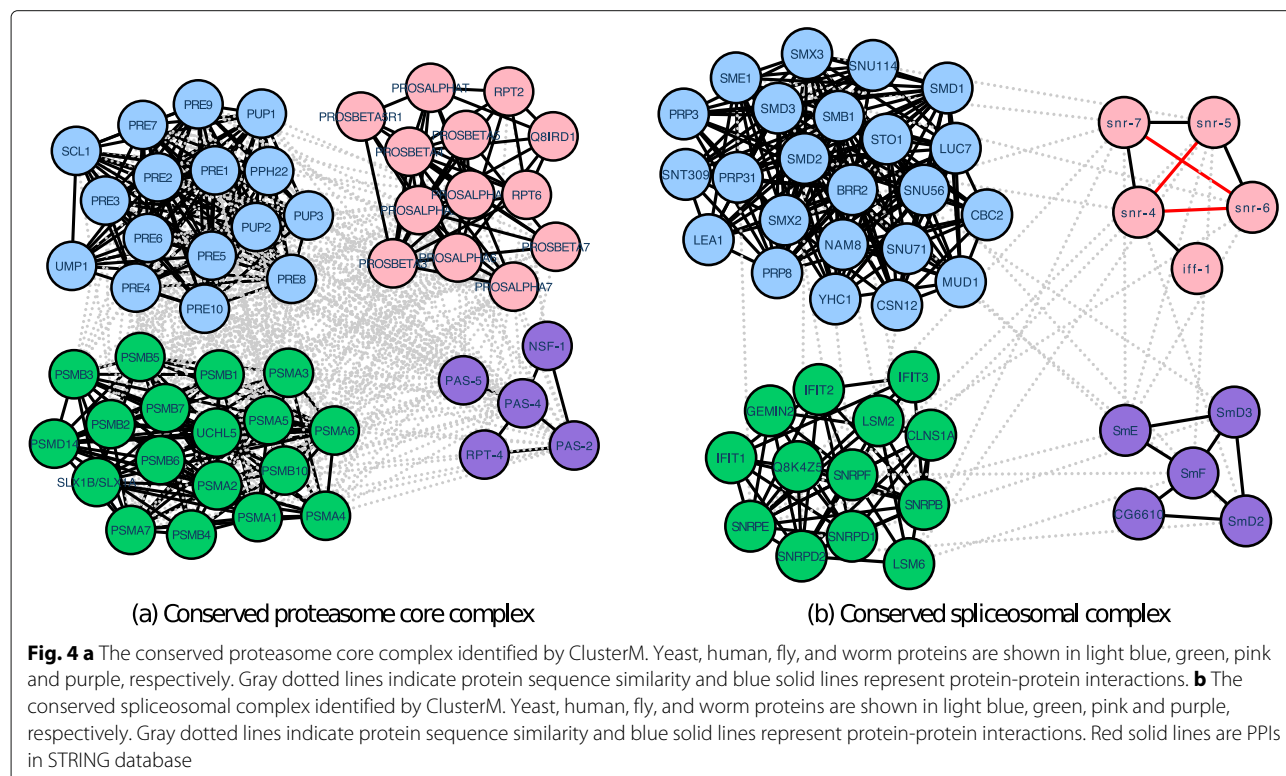
High-level GO term consistency

To examine the biological significance of the conserved protein complexes identified by ClusterM and other existing algorithms, we calculated the coverage and the mean

normalized entropy (MNE) score [7] based on high-level GO annotations for the results obtained by ClusterM, MaWISH, AlignMCL, and NetworkBLAST-M. Note that the lower MNE indicates that the proteins in the predicted conserved complex are more consistent in terms of GO annotations. We selected the three algorithms—MaWISH, AlignMCL, and NetworkBLAST-M—as they show the best performance among all compared algorithms except for ClusterM. For the blanks in Tables 2 and 3, MaWISH and AlignMCL cannot deal with more than 3 networks. We applied these four algorithms to identify conserved protein complexes for every combination of PPI

Table 1 Abbreviation and information for protein-protein interaction (PPI) networks used in the study

Dataset	Database	Version	Species	Abbreviation	#. proteins	#. interactions
DIPPIs	DIP	20150101	Yeast	SceDIP	5,138	22,835
	DIP	20150101	Human	HsaDIP	4,278	6,446
	DIP	20150101	Fly	DmeDIP	7,679	23,182
	DIP	20150101	Worm	CelDIP	2,712	4,117
IntActPPIs	IntAct	20150120	Yeast	SceIntAct	6,392	78,287
	IntAct	20150120	Human	HsaIntAct	23,246	106,031
	IntAct	20150120	Fly	DmeIntAct	11,517	41,483
	IntAct	20150120	Worm	CellIntAct	9,721	16,668

**Table 2** GO consistency and coverage comparison on DIPPIs dataset

PPI networks	measure	MaWish	AlignMCL	NB-M	CM1	CM10	CM100
SceDIP+HsaDIP	MNE	5.274	3.766	6.110	4.195	4.224	3.912
	Coverage	1013	1379	814	2533	2406	942
SceDIP+DmeDIP	MNE	3.149	2.434	3.383	2.300	2.286	2.204
	Coverage	1327	1945	1242	3496	3400	1395
SceDIP+CelDIP	MNE	3.105	2.476	3.490	2.344	2.339	2.134
	Coverage	383	809	488	1140	1099	353
HsaDIP+DmeDIP	MNE	5.823	4.976	6.252	4.753	4.734	4.757
	Coverage	1076	2909	1538	4239	3952	904
HsaDIP+CelDIP	MNE	6.972	5.359	6.483	4.937	4.978	4.724
	Coverage	260	1202	554	1656	1542	349
DmeDIP+CelDIP	MNE	3.366	2.599	3.317	2.246	2.232	2.123
	Coverage	522	1706	1044	1988	1928	403
SceDIP+HsaDIP+DmeDIP	MNE			6.168	4.470	4.444	4.614
	Coverage			1895	3628	3410	1525
SceDIP+HsaDIP+CelDIP	MNE			6.425	4.572	4.611	4.720
	Coverage			886	2415	2232	814
SceDIP+DmeDIP+CelDIP	MNE			3.746	2.554	2.563	2.414
	Coverage			1162	2736	2659	1200
HasDIP+DmeDIP+CelDIP	MNE			6.616	4.527	4.542	4.136
	Coverage			1610	2762	2516	697
SceDIP+HsaDIP+DmeDIP+CelDIP	MNE			5.981	4.305	4.339	4.149
	Coverage			1351	2049	1871	599

Abbreviations: CM1 = ClusterM($\lambda = 1$), CM10 = ClusterM($\lambda = 10$), CM100 = ClusterM($\lambda = 100$), MNE = Mean Normalized Entropy, NB-M = NetworkBLAST-M
 Bold values denote the best scores corresponding to specific criteria

Table 3 GO consistency and coverage comparison on IntAct dataset

PPI networks	measure	MaWISH	AlignMCL	NB-M	CM1	CM10	CM100
SceIntAct+HsaIntAct	MNE	4.734	3.137	4.028	3.089	3.103	3.018
	Coverage	1552	4181	478	6355	6216	3023
SceIntAct+DmeIntAct	MNE	3.679	2.705	4.208	2.307	2.310	2.446
	Coverage	1548	1290	423	2403	2391	1209
SceIntAct+CellIntAct	MNE	3.378	2.918	3.679	2.647	2.635	2.697
	Coverage	797	818	346	1578	1543	637
HsaIntAct+DmeIntAct	MNE	7.587	4.148	4.988	3.421	3.477	3.668
	Coverage	4413	5898	219	7069	6730	1941
HsaIntAct+CellIntAct	MNE	6.463	4.490	4.639	3.449	3.448	3.367
	Coverage	2263	3724	167	4077	3907	1202
DmeIntAct+CellIntAct	MNE	4.497	4.013	5.154	2.745	2.785	2.477
	Coverage	545	1120	204	1891	1833	520
SceIntAct+HsaIntAct+DmeIntAct	MNE			4.317	3.154	3.165	3.315
	Coverage			631	8640	8354	4233
SceIntAct+HsaIntAct+CellIntAct	MNE			3.848	3.294	3.307	3.352
	Coverage			408	6160	5952	2675
SceIntAct+DmeIntAct+CellIntAct	MNE			4.479	2.657	2.659	2.574
	Coverage			640	3321	3222	1282
HsaIntAct+DmeIntAct+CellIntAct	MNE			5.321	3.196	3.246	3.109
	Coverage			435	6213	5895	2168
SceIntACT+HsaIntACT+DmeIntACT+CellIntACT	MNE			4.344	3.139	3.153	3.252
	Coverage			456	5907	5668	2246

Abbreviations: CM1 = ClusterM($\lambda = 1$), CM10 = ClusterM($\lambda = 10$), CM100 = ClusterM($\lambda = 100$), NME = Mean Normalized Entropy, NB-M = NetworkBLAST-M

Note: Bold values denote the best scores corresponding to specific criteria

networks in DIPPIs, where the obtained results are presented in Table 2. The evaluation results based on IntActPPIs datasets can be found in Table 3. Both tables reveal that the conserved protein complexes identified by ClusterM consistently achieve lower MNE and higher coverage scores, where it means that ClusterM can predict a larger number of proteins with a high functional consistency.

Discussion

In this paper, we proposed a scalable algorithm—ClusterM—that can identify conserved protein complexes by integrating protein sequence information and topological structure of the PPI networks. As demonstrated by our results, ClusterM can better capture the desired topological property of a typical conserved protein complex, which is densely connected within the complex while being well-separated from the rest. Experimental results based on real-world PPI networks and protein complexes show that ClusterM significantly outperforms other state-of-the-art algorithms in the task of identifying conserved protein complexes. Additionally, the conserved protein complexes identified by

ClusterM have been shown to boast better high-level GO term consistency compared to other competing algorithms.

ClusterM is an enhanced approach that directly considers the characteristic topological structure of conserved protein complexes, which are typical densely connected within the complex and well-separated from the rest of the network. Unlike ClusterM, existing state-of-the-art algorithms (e.g., NetworkBLAST, NetworkBLAST-M, and MaWISH) focus on the interaction density of the conserved protein complex but do not explicitly consider the separability of conserved complexes from the rest of the PPI network. Furthermore, compared to algorithms such as NetworkBLAST and MaWISH, which only consider proteins with homology correspondence (determined based on protein sequence similarity) across PPI networks to reduce the overall computational complexity, ClusterM does not impose such restriction and integrates the topology and homology information in a very flexible manner. Finally, another important advantage of ClusterM is that the algorithm can easily handle multiple (more than two) PPI networks and yield biologically meaningful results. While NetworkBLAST-M can also deal with multiple PPI

networks, NetworkBLAST-M is not capable of handling large-scale networks.

Conclusions

We propose a scalable algorithm—ClusterM—that explicitly characterizes the desired topological separability of protein complexes and also incorporates sequence similarity of proteins across the given PPI networks. Thanks to the computational framework used by ClusterM, it can easily handle multiple PPI networks at the same time. We have extensively compared ClusterM with other state-of-the-art algorithms on various of PPI networks. The experiments show its out-performance over other methods. In addition, ClusterM shows its potential on analyzing four PPI networks and identifying conserved protein complexes that have not been identified before.

Materials and methods

The ClusterM algorithm

Topological separability by conductance

One of the major innovations in ClusterM as well as its core strength is that it explicitly considers topological separability when searching for conserved protein complexes. In this paper, we adopt the definition of conductance to measure the topological separability of a subnetwork. Let $G = (V, E)$ represent a PPI network, where V denotes the set of proteins in G and E is the interaction set. A is the corresponding adjacency matrix of G , where the element $A_{ij} = 1$ denotes the protein i interacts with the protein j and $A_{ij} = 0$ otherwise. The degree matrix D of G is a diagonal matrix with $D_{ii} = d_i$, where d_i is the number of interactions connecting to the protein i .

For a subnetwork S as a potential protein complex, the conductance of S in G is defined as

$$\phi(S) = \frac{|E(S, \bar{S})|}{\min\{vol(S), vol(\bar{S})\}}, S \cup \bar{S} = V, \quad (1)$$

where $E(S, \bar{S})$ denotes the set of edges between S and the rest of the network \bar{S} , and $vol(S) = \sum_{i \in S} d_i$ is the number of interactions in S . As $vol(S)$ is typically much smaller than the total number of interactions in G : $vol(S) \ll vol(V)$, indicating $vol(S) = \min\{vol(S), vol(\bar{S})\}$, we have

$$\phi(S) = \frac{|E(S, \bar{S})|}{vol(S)} = \frac{\sum_{i,j \in V_S} D_{ij} - A_{ij}}{\sum_{i \in V_S} D_{ii}}, \quad (2)$$

where V_S is the vertex set containing all vertexes in the subnetwork S .

The algorithm

ClusterM builds on the intuition that conserved protein complexes should simultaneously possess the following two properties. First, topologically, conserved protein complexes in each PPI network should be well separated

from the rest of the network and proteins within the complexes should be densely connected in order to give rise to a unique and specific biological form and function. Second, across species, there should exist many homologous proteins in the conserved complexes, which can be practically reflected by high overall protein sequence similarity. Given k PPI networks $\mathbf{G} = \{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_k\}$, where $\mathcal{G}_j(\mathcal{V}_j, \mathcal{E}_j)$ is the j th network with \mathcal{V}_j and \mathcal{E}_j denoting the corresponding proteins and interactions respectively, we use a binary adjacency matrix A_j to represent \mathcal{G}_j and use a diagonal matrix D_j to represent the degree matrix of \mathcal{G}_j with the number of interactions of each protein on its diagonal.

To identify conserved protein complexes with the aforementioned properties, our ClusterM algorithm takes three major steps:

1. Initial seeds with protein spines from multiple network alignment: In the first step, h homologous seeds can be identified as protein spines $\mathcal{U} = \{u^1, u^2, \dots, u^h\}$ by using a multiple network alignment method, where $u^i = \{(v_1^i, v_2^i, \dots, v_k^i) \mid v_1^i \in \mathcal{V}_1, v_2^i \in \mathcal{V}_2, \dots, v_k^i \in \mathcal{V}_k\}$ constitutes the i th protein spine detected through network alignment. In this work, we adopt SMETANA [28], which has been shown to be accurate and scalable. When detecting these conserved protein spines, SMETANA [28] takes both protein interaction and sequence information into account. By iterating through all protein spines in \mathcal{U} , ClusterM adopts a divide-and-conquer strategy to identify potential conserved protein complexes from each protein spine seed across the given networks.

2. Minimum-conductance set for topological separability (Task 1): First, for protein v_j^i in protein spine u^i , a minimum-conductance protein set $\hat{\mathcal{H}}_j^i$ including v_j^i , well separated from the rest of the network \mathcal{G}_j , is identified based on a novel local optimization algorithm. Initially, we acquire a set \mathcal{H}_j^i consisting of m proteins obtained in terms of the ranking of the personalized PageRank vector with respect to v_j^i in \mathcal{G}_j [29]. We further refine the results to yield $\hat{\mathcal{H}}_j^i$ based on the definition of the conductance (2) by solving the following optimization problem:

$$\begin{aligned} \min: & \frac{\mathbf{x}^T (D^{\mathcal{H}_j^i} - A^{\mathcal{H}_j^i}) \mathbf{x}}{\mathbf{x}^T D^{\mathcal{H}_j^i} \mathbf{x}} \\ \text{s.t. } & \mathbf{x}_{v_j^i} = 1, \mathbf{x}_i \in \{0, 1\}, \end{aligned} \quad (3)$$

where \mathbf{x} is a binary vector with $\mathbf{x}_i = 1$ indicating that protein i is assigned into $\hat{\mathcal{H}}_j^i$ and $\mathbf{x}_i = 0$ otherwise; $A^{\mathcal{H}_j^i}$ and $D^{\mathcal{H}_j^i}$ are adjacency and degree matrices of the induced subnetwork with respect to the protein set \mathcal{H}_j^i . The problem can be solved by transforming it into a mixed integer program (MIP) [30]. After algebraic manipulations, (3) can be transformed into the following equivalent MIP formulation:

min: z

$$\begin{aligned} \text{s.t. } z \sum_i \mathbf{x}_i d_i^{\mathcal{H}_j^i} - \sum_i \sum_j (D^{\mathcal{H}_j^i} - A^{\mathcal{H}_j^i}) \mathbf{x}_i \mathbf{x}_j &\geq 0, \\ \mathbf{x}_{v_j^i} &= 1, \mathbf{x}_i \in \{0, 1\}, \end{aligned} \quad (4)$$

where $d_i^{\mathcal{H}_j^i}$ is the i th elements on the diagonal of $D^{\mathcal{H}_j^i}$. After using standard linearization techniques [30] to linearize the terms such as $z\mathbf{x}_i$ and $\mathbf{x}_i\mathbf{x}_j$, the optimization problem can be solved exactly by existing MIP solvers. We emphasize that we here obtain the exact minimum-conductance set around the selected seed, which is critical to identify potential protein complexes with high topological separability from the rest of the networks. Because ClusterM searches for the minimum-conductance set locally around each involved protein, and since the size of each protein set \mathcal{H}_j^i is much smaller than the size of the entire PPI network in such a divide-and-conquer strategy, we can efficiently obtain the minimum-conductance set $\hat{\mathcal{H}}_j^i$ in \mathcal{H}_j^i based on (4) [30].

Here, $\hat{\mathcal{H}}_j^i$ is the subnetwork including protein v_j^i that is well separated from the other proteins in the PPI network \mathcal{G}_j . The procedure of identifying $\hat{\mathcal{H}}_j^i$ explicitly characterizes the external separability of the subnetwork $\hat{\mathcal{H}}_j^i$, which is the main advantage of ClusterM over other existing network comparative approaches [13–15, 17, 18] that allows the algorithm to effectively capture the desired topological property of conserved protein complexes.

3. Conserved protein complexes with desired topological and homologous properties (Task 2): The next task is to collect $\mathbf{H}^i = \{\hat{\mathcal{H}}_1^i, \dots, \hat{\mathcal{H}}_k^i\}$ for the corresponding protein spine u^i with $\mathbf{S}^i = \{s(p, q) \mid p \in \hat{\mathcal{H}}_m^i, q \in \hat{\mathcal{H}}_n^i, m \neq n, \forall p, q, m, n\}$ denoting potential homologous correspondence within \mathbf{H}^i . The homologous correspondence between proteins p and q can be approximated by the protein sequence similarity $s(p, q) \in [0, 1]$ [31]:

$$s(p, q) = \frac{\text{blast}(p, q)}{\sqrt{\text{blast}(p, p) \times \text{blast}(q, q)}}, \quad (5)$$

where $\text{blast}(p, q)$ is the bit score of the sequence similarity obtained by the local sequence alignment tool BLAST [20]. We note that better homologous correspondence by incorporating functional annotations may further improve the performance of conserved protein complex identification. However, we do not explore such directions to avoid biased performance evaluation due to repeated usage of functional annotations in our algorithm and evaluation metrics.

To detect conserved complexes with high interaction density as well as high sequence similarity between proteins across different species, we propose to optimize the following cost function for protein spine u^i :

$$\mathcal{J}^i = - \left(\sum_{j=1}^k \frac{\sum_{a,b \in \hat{\mathcal{H}}_j^i} A^{\hat{\mathcal{H}}_j^i}(a, b) \delta_a \delta_b}{\sum_{c \in \hat{\mathcal{H}}_j^i} \delta_c} + \lambda \frac{\sum_{j \neq l} \sum_{a \in \hat{\mathcal{H}}_j^i} \sum_{b \in \hat{\mathcal{H}}_l^i} s(a, b) \delta_a \delta_b}{\sum_{j=1}^k \sum_{c \in \hat{\mathcal{H}}_j^i} \delta_c} \right), \quad (6)$$

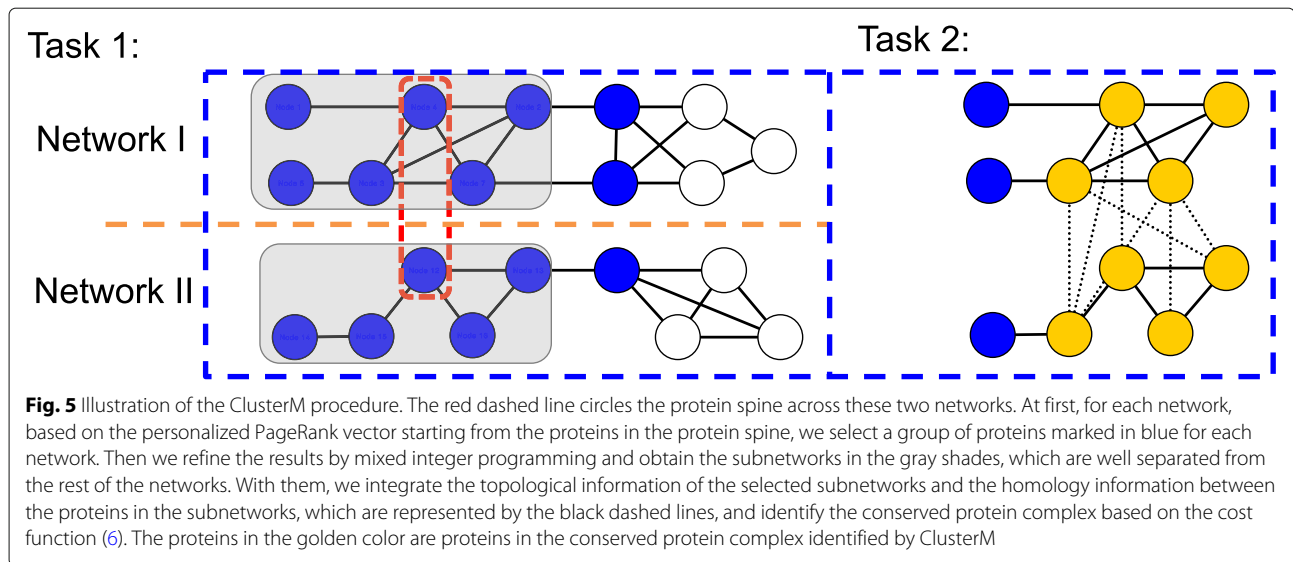
where the binary value $A^{\hat{\mathcal{H}}_j^i}(a, b)$ indicates whether there is an interaction between proteins a and b in the subnetwork induced by the protein set $\hat{\mathcal{H}}_j^i$, and δ_a is an indicator function with $\delta_a = 1$ if protein a appears in the identified conserved protein complex with respect to u^i and $\delta_a = 0$ otherwise. The first term is essentially the summation of the interaction density [32] of each subnetwork, which characterizes the topological cohesiveness of each subnetwork. The second term characterizes the homology correspondence represented by the summation of the protein sequence similarities across subnetworks divided by the total number of proteins in the identified conserved protein complex. The coefficient λ balances the contributions from network topology and protein homology information. We optimize the cost function (6) in a greedy manner by first recruiting all the proteins in \mathbf{H}^i , and greedily removing proteins based on their contributions to the cost function until further deletion does not reduce the cost function any more. The greedy algorithm outputs the conserved protein complex \mathcal{C}^i with respect to the protein spine u^i .

Figure 5 illustrates how ClusterM handles a protein spine in a pair of networks to search for conserved protein complex locally. Starting from the protein spine in the red dashed line, ClusterM first identifies the proteins in blue based on the personalized PageRank vectors. It further refines the results for separability by MIP (3) and recognize the proteins in the gray shade to be the subnetworks that are well separated from the rest of the corresponding networks. Finally, ClusterM finds the conserved protein complex in golden color based on the cost function (6).

In the post-processing step of ClusterM, we remove duplicated conserved protein complexes, and delete the i th conserved complex \mathcal{C}^i if the summation of the sequence similarity score ($\sum_{(p,q) \in \mathcal{C}^i} s(p, q)$) within \mathcal{C}^i is lower than β (set to $\beta = 0.1$ in this work).

Scalability and parallelism

The major computation in ClusterM involves SMETANA for identifying homologous seeds, approximating the PageRank vector near a protein, solving the MIP (3), and the greedy algorithm to optimize (6). The scalability of



SMETANA has been demonstrated in [28]. The computational complexity for approximating the PageRank vector near a protein is proportional to the number of involved local neighbors [29]. There is no scalability issue for solving the MIP (3) because we only consider m local proteins near the protein spine for individual PPI networks. Both the time and space complexities of the greedy algorithm based on (6) are $O(k^2m^2)$, where k is the number of species. Therefore, due to the divide-and-conquer strategy, ClusterM has better scalability for handling multiple species.

For each protein spine u^i , the identification of the well-separated subnetworks $\hat{\mathcal{H}}_1^i, \dots, \hat{\mathcal{H}}_k^i$ are independent to each other. Furthermore, the protein spines and the identification of the potential conserved protein complexes are also independent. Therefore, the computation of $\hat{\mathcal{H}}_1^i, \dots, \hat{\mathcal{H}}_k^i$ and the detection of conserved protein complexes can be easily parallelized to further reduce the total computational time for ClusterM.

Datasets

Construction of DIPPIs and IntActPIs

We download PPI networks and protein amino acids sequences for human, yeast, fly, and worm from DIP (version 20150101) and IntAct (version 20150120). To obtain a consistent protein symbol for PPI networks and conserved protein complexes, we change all protein names utilized in this study to UniProt protein symbols. The sequence similarity for each protein pair can be obtained using a local sequence alignment tool BLAST and we select the highest BLAST bit score for each protein pair as their sequence similarity score. Then, we normalize the BLAST bit score based on the equation (5) to obtain protein homologous correspondence and threshold the sequence similarity score at 0.1.

The yeast-human reference conserved protein complexes

We align yeast complexes in CYC2008 [24] and human complexes in CORUM (February 2012) based on GO terms to obtain the gold standard for yeast and human conserved protein complexes. CYC2008 provides a specific GO term associated to every collected yeast complex. For human complex in CORUM, we first download the mapping of human genes and proteins to GO terms (version 20150329) [33] and annotate proteins in CORUM with GO terms that also appear in CYC2008. If a human complex contains at least half the number of proteins annotated to a GO term in CYC2008, we align the human complex and the yeast complex with the same GO term annotation.

High-level GO terms

When analyzing biological consistency, we only consider the high-level GO terms, which suggest specific biological functions. A GO term is defined as high-level if its information content is larger than two. The definition of the information content of a GO term g is $IC = -\log(|g|/|root|)$ [34], where “root” is the corresponding root GO term (either biological process, molecular function, or cellular component) of g and the operation $|\cdot|$ counts the number of proteins annotated to a specified GO term. Additionally, we remove GO terms “inferred from electronic annotation”, “inferred from protein interactions”, and “inferred from sequence or structural similarity”, because we utilize protein-protein interactions and the protein sequence similarities in our algorithm.

Metrics for conserved protein complex prediction for pairwise PPI networks

The conserved protein complexes from pairwise PPI network alignment contain proteins from two species.

Therefore, we need to examine the correspondence between the reference and the identified conserved complexes species by species. For all competing algorithms, we remove the identified conserved protein complexes if they contain fewer than three proteins for any species.

Fraction of the matched reference conserved protein complexes

For two species, given a reference conserved complex $A = \{A_1, A_2\}$ and an identified conserved complex $B = \{B_1, B_2\}$, where A_1 and B_1 are complexes in the first species and A_2 and B_2 are complexes in the second species, we consider B matches A if the neighborhood affinity scores for complexes $w(A_1, B_1) \geq 0.25$ and $w(A_2, B_2) \geq 0.25$. The neighborhood affinity score for complexes can be calculated by

$$w(X, Y) = \frac{|X \cap Y|}{|X|} \cdot \frac{|X \cap Y|}{|Y|} = \frac{|X \cap Y|^2}{|X| |Y|}, \quad (7)$$

where X and Y are two protein sets. Therefore, we can compute the fraction of the number of matched reference conserved complexes.

Assuming the reference conserved protein complexes set is R and the identified conserved protein complexes set is P , then the fraction of the number of matched reference conserved complexes can be computed as

$$\text{frac} = \frac{|C|}{|R|}, \quad (8)$$

where $C = \{S \mid w(S_1, T_1) \geq 0.25 \text{ and } w(S_2, T_2) \geq 0.25, \forall S \in R, T \in P\}$.

Accuracy score

Suppose that we have n reference conserved complex $A^j = \{A_1^j, A_2^j\}$, $j = 1, 2, \dots, n$, and m identified complex $B^i = \{B_1^i, B_2^i\}$, $i = 1, 2, \dots, m$, and each has complexes A_1^j and B_1^i from the first species and another complexes A_2^j and B_2^i in the second species. Let t_{ij} denotes the number of proteins that exist in both the reference complex A^j and the identified complex B^i , and w_j represent the number of proteins in the j th reference complex. That is, $t_{ij} = |A_1^j \cap B_1^i| + |A_2^j \cap B_2^i|$ and $w_j = |A_1^j \cup A_2^j|$. Then, complex-wise sensitivity (Sn) and positive-predictive value (PPV) can be defined as

$$\text{Sn} = \frac{\sum_{j=1}^n \max_{i=1, \dots, m} t_{ij}}{\sum_{j=1}^n w_j}; \quad \text{PPV} = \frac{\sum_{i=1}^m \max_{j=1, \dots, n} t_{ij}}{\sum_{i=1}^m \sum_{j=1}^n t_{ij}}. \quad (9)$$

The geometric accuracy (Acc) score is the geometric mean of Sn and PPV: $\text{Acc} = \sqrt{\text{Sn} \times \text{PPV}}$.

The maximum matching ratio for conserved protein complexes (MMRC)

We can quantify the overlap between the reference and the identified conserved complexes by the maximum matching ratio for conserved complexes (MMRC). MMRC is derived based on MMR [6]. The maximum matching ratio [6] is the maximum sum of weights of edges in a bipartite graph, where the two sets of nodes represent reference complexes C and identified complexes S . The bipartite graph is represented by a weighted matrix $B_{n \times m}$, where each weight B_{ij} is the neighborhood affinity score $w(c_i, s_j)$ introduced earlier for the corresponding edge between complexes c_i and s_j . For efficiency, we only use $B_{ij} \geq 0.25$. The MMR is the solution to the following maximal matching problem.

$$\begin{aligned} \max: & \frac{1}{|C|} \sum_{i=1}^n \sum_{j=1}^m B_{ij} \sigma_{c_i, s_j} \\ \text{s.t.} & \sum_{j=1}^m \sigma_{c_i, s_j} \leq 1 \\ & \sum_{i=1}^n \sigma_{c_i, s_j} \leq 1, \end{aligned} \quad (10)$$

where σ is an indicator function with $\sigma_{c_i, s_j} = 1$ when the edge between complexes c_i and s_j is selected and $\sigma_{c_i, s_j} = 0$ otherwise.

MMRC can be obtained by modifying the overlapping weights $w(c_i, s_j)$ between reference and identified conserved complexes. The overlapping weight between reference conserved complex A and identified conserved complex B is defined as

$$o(A, B) = \frac{2w(A_1, B_1) \times w(A_2, B_2)}{w(A_1, B_1) + w(A_2, B_2)}. \quad (11)$$

Metrics for GO term consistency and coverage

We measure the functional consistency for the proteins in the conserved complexes by computing the mean normalized entropy (MNE) [7, 10] scores. We use the high-level GO term set F to annotate each protein in a conserved complex R_i . The union of GO terms used for R_i is $F_i = \{f_1, f_2, \dots, f_d\}$. The normalized entropy (NE) of R_i is computed as

$$\text{NE}(R_i) = \text{NE}(p_1, p_2, \dots, p_d) = -\frac{1}{\log(d)} \sum_j p_j \cdot \log(p_j), \quad (12)$$

where p_i is the fraction of R_i with respect to the GO term f_i . The MNE is the mean value over all $\text{NE}(R_i)$. For the coverage, we simply count the number of proteins in all conserved complexes, each of which consists of at least three proteins from each network.

Parameter selection for each method

Among the tested algorithms, ClusterM and NetworkBLAST-M have only a single tuning parameter, while NetworkBLAST and MaWISH have 5 and 7 tuning parameters, respectively. In order to obtain the best performing parameters for fair comparison of the selected algorithms, we have selected the best performing parameter(s) for each algorithm by grid search. Within at most N combinations of k parameters (p_1, p_2, \dots, p_k), we sample $n = \lfloor \sqrt[k]{N} \rfloor$ values for each parameter. The sample values are uniformly distributed between $[min_i, max_i]$ for each parameter p_i . For all the competing algorithms, we set $N = 100$ and report the results with the best performing parameters.

Implementation details

The MATLAB code and all used data are available at <http://www.ece.tamu.edu/~xqian/ClusterM/>. ClusterM is computationally efficient for handling multiple genome-scale PPI networks from different species. The memory consumption of ClusterM depends on the size of the subnetworks to be identified and the runtime of ClusterM can be significantly reduced by parallelization. It takes about 40 minutes for ClusterM to handle four PPI networks in IntActPPIs with 50,876 proteins in total on a laptop computer (16 GB memory and Intel i7 2.9 GHz cpu) with 2-core parallelization.

Abbreviations

PPI: Protein-Protein Interaction; *Sce*: *Saccharomyces cerevisiae*; *Dme*: *Drosophila melanogaster*; *Cel*: *Caenorhabditis elegans*; *Hsa*: *Homo sapiens*; MMRC: The maximum matching ratio for conserved complexes; MNE: The mean normalized entropy; GO: Gene Ontology; NB: NetworkBLAST; NBM: NetworkBLAST-M

About this supplement

This article has been published as part of *BMC Genomics Volume 21 Supplement 10, 2020: Selected articles from the 18th Asia Pacific Bioinformatics Conference (APBC 2020): genomics*. The full contents of the supplement are available online at <https://bmcbgenomics.biomedcentral.com/articles/supplements/volume-21-supplement-10>.

Authors' contributions

Conceived the method: YW, HJ, XQ, BJY. Developed the algorithm and performed the simulations: YW, HJ. Analyzed the results and wrote the paper: WY, HJ, XQ, BJY. All authors read and approved the final manuscript.

Funding

This work was supported by the National Science Foundation (NSF) Grants 1447235, 1553281, and 1812641, as well as the United States Department of Agriculture National Institute of Food and Agriculture competitive grant USDA-NIFASCR1-2017-51181-26834 through the National Center of Excellence for Melon at the Vegetable and Fruit Improvement Center of Texas A&M University. This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2019R1G1A1004803). Publication costs are funded by NSF Grant 1812641.

Availability of data and materials

The MATLAB code and all used data are available at <http://www.ece.tamu.edu/~xqian/ClusterM/>.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All the authors approve the publication of the presented work.

Competing interests

The authors declare that they have no competing interests.

Author details

¹School of Informatics, Computing and Engineering, Indiana University, Bloomington 47405, IN, USA. ²Department of Mechatronics Engineering, Incheon National University, Incheon 22012, South Korea. ³Department of Electrical and Computer Engineering, Texas A&M University, College Station 77843, TX, USA. ⁴TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering (CBGSE), Texas A&M University, College Station 77843, TX, USA. ⁵Computational Science Initiative, Brookhaven National Lab, Upton 11973, NY, USA.

Published: 18 November 2020

References

- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*. 2001;98(8):4569–74.
- WH D, M M, AC G. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics*. 2012;12(10):1576–90.
- Salwinski L, Miller C, Smith A, Pettit F, JU JB, Eisenberg D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*. 2004;32:449–51.
- Kerrien S, Aranda B, Breuza L, et al. The intact molecular interaction database in 2012. *Nucleic Acids Res*. 2012;40(D1):841–6.
- Stark C, Breitkreutz B, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res*. 2006;34:535–9.
- Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods*. 2012;9(5):471–472.
- Liao C, Lu K, Baym M, Singh R, Berger B. IsoRankN: Spectral methods for global alignment of multiple protein networks. *Bioinformatics*. 2009;25: 253–8.
- Aladag AE, Erten C. Spinal: scalable protein interaction network alignment. *Bioinformatics*. 2013;29(7):917–24.
- Hasan MM, Kahveci T. Indexing a protein-protein interaction network expedites network alignment. *BMC Bioinformatics*. 2015;16:326.
- Singh R, Xu J, Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc Natl Acad Sci USA*. 2008;105(35):12763–8.
- Clark C, Kalita J. A multiobjective memetic algorithm for ppi network alignment. *Bioinformatics*. 2015;31(12):1988–98.
- Jeong H, Yoon B-J. Accurate multiple network alignment through context-sensitive random walk. *BMC Syst Biol*. 2015;9 Suppl 1:7.
- Sharan R, Suthram S, Kelley RM, et al. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA*. 2005;102:1974–9.
- Koyuturk M, Grama A, Szpankowski W. Pairwise Local Alignment of Protein Interaction Networks Guided by Models of Evolution. In: RECOMB 2005. Berlin, Heidelberg: Springer; 2005. p. 48–65.
- Kalaev M, Bafna V, Sharan R. Fast and accurate alignment of multiple protein networks. *J Comput Biol*. 2009;16(8):989–99.
- Yoon B-J, Qian X, Sahraeian SME. Comparative analysis of biological networks: hidden markov model and markov chain-based approach. *IEEE Signal Proc Mag*. 2012;29(1):22–34.
- Ciriello G, Mina M, Guzzi PH, Cannataro M, Guerra C. Alignnemo: A local network alignment method to integrate homology and topology. *PLoS ONE*. 2012;7(6):38107.
- Hirsh E, Sharan R. Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics*. 2007;23(2):170–6.
- Berg J. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol Biol*. 2004;4(51): <https://doi.org/10.1186/1471-2148-4-51>.
- Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning dna sequences. *J Comput Biol*. 2000;7(1-2):203–14.
- Mina M, Guzzi PH. Alignmcl: Comparative analysis of protein interaction networks through markov clustering. In: *Bioinformatics and Biomedicine Workshops (BIBMW)*, 2012 IEEE International Conference On. IEEE; 2012. p. 174–81. <https://doi.org/10.1109/BIBMW.2012.6470300>.

22. Sharan R, Suthram S, Kelley R, Kuhn T, McCuine S, Uetz P, Sittler T, Karp R, Ideker T. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA*. 2005;102(6):1974–9.
23. Kalaev M, Bafna V, Sharan R. Fast and accurate alignment of multiple protein networks. In: *Proc of the 10th Annu Int Conf Res Comput Mol Bio (RECOMB 2008)*; 2008. https://doi.org/10.1007/978-3-540-78839-3_21.
24. Pu S, Wong J, Turner B, et al. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res*. 2009;37(3):825–31.
25. Ruepp A, Brauner B, Dunger-Kaltenbach I, et al. Corum: The comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*. 2008;36:646–50.
26. Maslov S, Sneppen K. Specificity and stability in topology of protein networks. *Science*. 2002;296:910–3.
27. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2014;43(D1):447–52.
28. Sahraeian SME, Yoon B-J. Smetana: Accurate and scalable algorithm for probabilistic alignment of large-scale biological networks. *PLoS ONE*. 2013;8(7):67995.
29. Andersen R, Chung F, Lang K. Local Graph Partitioning Using PageRank Vectors. In: *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*; 2006. p. 475–86. <https://doi.org/10.1109/FOCS.2006.44>.
30. Fan N, Pardalos P. Multi-way clustering and biclustering by the ratio cut and normalized cut in graphs. *J Comb Optim*. 2010;23(2):224–51.
31. Wang Y, Qian X. Joint clustering of protein interaction networks through markov random walk. *BMC Syst Biol*. 2014;8(suppl 1):9.
32. Corneil DG, Perl Y. Clustering and domination in perfect graphs. *Discret Appl Math*. 1984;9(1):27–39.
33. Ashburner M, Ball C, Blake J, et al. Gene ontology: Tool for the unification of biology. the gene ontology consortium. *Nat Genet*. 2000;25(1):25–9.
34. Shih Y-K, Parthasarathy S. Identifying functional modules in interaction networks through overlapping markov clustering. *Bioinformatics*. 2012;28(18):1473–9.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

