OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae692
https://doi.org/10.1093/bib/bbae692
Review

A comprehensive benchmark study of methods for identifying significantly perturbed subnetworks in cancer

Le Yang[1], Runpu Chen[1], Steve Goodison[2], Yijun Sun[1,3,*]

[1]Department of Microbiology and Immunology, University at Buffalo, The State University of New York, 955 Main Street, Buffalo, New York, NY 14203, United States
[2]Department of Quantitative Health Sciences, Mayo Clinic, 4500 San Pablo Rd S, Jacksonville, FL 32224, United States
[3]Department of Computer Science and Engineering, University at Buffalo, The State University of New York, 12 Capen Hall, Buffalo, New York, NY 14260, United States

*Corresponding author: yijunsun@buffalo.edu

Abstract

Network-based methods utilize protein–protein interaction information to identify significantly perturbed subnetworks in cancer and to propose key molecular pathways. Numerous methods have been developed, but to date, a rigorous benchmark analysis to compare the performance of existing approaches is lacking. In this paper, we proposed a novel benchmarking framework using synthetic data and conducted a comprehensive analysis to investigate the ability of existing methods to detect target genes and subnetworks and to control false positives, and how they perform in the presence of topological biases at both gene and subnetwork levels. Our analysis revealed insights into algorithmic performance that were previously unattainable. Based on the results of the benchmark study, we presented a practical guide for users on how to select appropriate detection methods and protein–protein interaction networks for cancer pathway identification, and provided suggestions for future algorithm development.

Keywords: perturbed subnetworks; protein–protein interaction; caner driver gene identification; cancer pathway; benchmark study

Introduction

The development of cancer within an individual is an evolutionary process driven by the accumulation of gene mutations that confer selective growth advantages to malignant cells, often by perturbing normal cellular processes [1]. Consequently, the identification of cancer driver genes and impacted molecular pathways is key to the elucidation of the mechanisms underlying cancer development. Extensive efforts to compile genomic data from large-scale tumor tissue studies, notably by The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) [2, 3], have greatly facilitated this goal. These initiatives have enabled us to identify genes that are mutated at frequencies significantly higher than those expected by random chance [4], providing insights into tumor biology that have supported numerous research endeavors [5].

Network-based approaches extend frequency-based methods by integrating frequency data with a protein–protein interaction (PPI) network [6, 7]. Since interacting proteins are often functionally related or work together, network-based approaches enable the identification of low-frequency genes that may still play significant roles in cancer. Additionally, they facilitate the detection of clustering patterns that reveal molecular pathways contributing to cancer progression. This analytical strategy has thus become a standard component in large-scale cancer studies [6, 8].

A dozen network-based methods have been previously developed [6, 7, 9–14]. However, to our knowledge, no benchmark study has yet been performed to comparatively evaluate the performance of these approaches. Lazareva *et al.* [15] conducted a benchmark study for a related problem, active module identification, where network-based approaches were utilized in gene expression analysis to identify modules of genes that display changes in expression levels under disease conditions. The study was designed to assess whether modules identified using a PPI network are biologically more meaningful than those identified using a randomly generated network. The study provided several insights, establishing that while DOMINO [9] was able to yield statistically significant results with real PPI networks, other methods identified disease-related modules based primarily on node degrees rather than the topological structures of node interactions. However, the findings of this study are not applicable to the cancer subnetwork identification problem due to the fundamental differences between gene expression and mutation data. For gene expression data, changes in gene expression levels are not necessarily directly associated with a disease due to the cascading effect [15]. Therefore, algorithms designed for active module identification tend to exclude genes that are highly differentially expressed but falsely associated with a disease. In contrast, for mutation data, there are only a few highly mutated driver

Received: May 28, 2024. **Revised:** December 2, 2024. **Accepted:** December 16, 2024

genes, and less prevalent potential drivers cannot be detected by frequency-based methods. Consequently, algorithms for cancer subnetwork identification prioritize the identification of cancer driver genes that are not highly mutated but close to highly mutated drivers in a PPI network. Another limitation of the benchmark study is that it assessed only the impact of node degrees on algorithm performance, but overlooked other critical topological attributes of PPI networks, likely due to the difficulty of generating synthetic networks with specific topological features using permutation and sampling approaches. Moreover, the study focused on active module identification, and certain methods, particularly those tailored for cancer research, were not thoroughly investigated.

In this study, we performed an in-depth benchmark analysis to evaluate and compare a set of recently developed algorithms for the identification of cancer pathways. Specifically, we generated synthetic target subnetworks, used as ground truths, by selecting gene lists from established protein complex or pathway databases. Subsequently, we simulated synthetic *P*-values and applied various computational methods to these values, evaluating their effectiveness in accurately recovering synthetic targets. The use of synthetic targets in our study enables a comprehensive comparison of different algorithms, revealing potential biases toward certain topological features of genes and subnetworks. Our analysis provides insights into algorithmic performance that were previously unattainable and presents findings that challenge established views. Firstly, contrary to the findings in a previous comparative study [15], DOMINO did not perform well in our analysis. Secondly, we found that some methods, although designed to counteract degree bias by penalizing high-degree nodes, tend to do so excessively. This over-penalization may lead to the rejection of those genes, regardless of their association with cancer. Thirdly, our analysis on eigenvector centrality revealed a shortfall in current methods in that they fail to take into consideration genes with low degree and high eigenvector centrality, thus being indirectly impacted by degree bias. Lastly, we observed that all methods under study favored community-like subnetworks. Interestingly, these subnetworks, despite their clear topological modularity, do not consistently align with disease modules described in previous studies [16]. We presented the main findings of this study as well as the comparison with the study conducted in Lazareva *et al*. [15] in Supplemental Table 1. Our findings have implications for both practitioners and developers and highlight key considerations for the future development and refinement of these algorithms.

Literature survey

We conducted a comprehensive benchmark study of 12 network-based methods, namely BioNet [17, 18], ROBUST [12], MuST [19], RegMOD [20], HotNet2 [6], hierarchical HotNet [10], DOMINO [9], DIAMOnD [21], FDRnet [7], NetMix2 [13], NetCore [11], and ClustEx [22] (Supplemental Table 2). The selection of these methods was based on three primary criteria. First, a preference was given to more recently developed methods. Secondly, we focused on methods that take as input a PPI network and *P*-values obtained from a gene-level mutation analysis (e.g. MutSig [4]) and yield a list of subnetworks as potential cancer pathways. Third, we prioritized methods that are compatible with high-performance computing platforms and require minimal configuration, due to the extensive computational demands of our experiment. Considering that the main challenge in developing network-based methods is to devise strategies to effectively integrate gene-based scores with the PPI network topology, we categorized the selected methods into five groups based on their integration strategies: neighbor-based (DIAMOnD), maximum score-based (BioNet, ROBUST, MuST), diffusion-based (RegMOD, HotNet2, hierarchical HotNet), community-based (DOMINO), and hybrid methods (FDRnet, NetMix2, NetCore, ClustEx). Next, we provide a brief discussion of the methods in each category.

Neighborhood-based approach: Neighborhood structure is a fundamental aspect of network-based analysis. DIAMOnD works by integrating neighborhood interactions in a PPI network with gene scores. It starts with a core set of genes that have high scores and iteratively expands the set. In each iteration, new genes are added to the disease module identified in the previous iteration if the new genes have a significantly higher number of connections with genes already in the module than would be expected by chance. A notable limitation of this method is that it does not take into consideration gene scores in the iterative expansion process. As a result, there is a potential for incorporating genes into a detected module that, despite their connectivity, may not be relevant to cancer.

Maximum score-based approach: The connectivity structure offers a more comprehensive view than the neighborhood structure in network analysis, as it considers the interconnections among a set of genes rather than focusing solely on individual genes. To leverage the connectivity structure, the maximum score-based approach identifies a connected subnetwork in a PPI network that maximizes the total gene score. The representative methods include BioNet, MuST, and ROBUST. Typically, gene scores are derived from *P*-values, and the task of subnetwork identification is formulated as a combinatorial optimization problem, commonly known as the Steiner tree problem [18, 23, 24]. Consequently, methods in this category can employ various techniques for efficient solution discovery. However, a key limitation is that they are unable to adequately address the scale-free nature of PPI networks. Since many genes are connected with each other through numerous high-degree nodes, these methods often end up linking a large number of cancer-related genes into a large network, thereby failing to discover subnetwork structures.

Diffusion-based approach: Diffusion processes offer a more effective way to harness complex topological structures of a network, going beyond simple connectivity analysis. Consequently, several diffusion-based methods have been developed to address the issue of the scale-free structure. Gene scores are incorporated into diffusion processes either through genes (e.g. RegMOD) or through gene interactions (e.g. HotNet2 and hierarchical HotNet). Upon completion of a diffusion process, a threshold is chosen to extract high-scoring modules. However, selecting an appropriate threshold poses a significant challenge. HotNet2 and hierarchical HotNet tackle this issue using time-intensive, permutation-based techniques, whereas RegMOD opts for a data-dependent, outlier-based approach.

Community-based approach: Community detection is an alternative way to exploit the complex topological structures of a network. In this approach, as exemplified by DOMINO, the entire PPI network is first partitioned into distinct communities using a community detection algorithm. Subsequently, sub-communities enriched with high-scoring genes are identified. While the community detection problem has been well studied and can be solved efficiently, this approach suffers from two notable drawbacks. First, the results of community detection are static and cannot be altered through post-analysis. This rigidity means that a subnetwork spanning multiple communities might not be identified. Secondly, since gene scores are not used in the

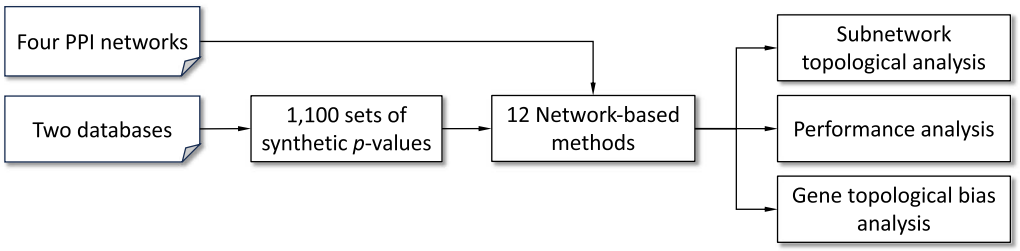

Figure 1. Overview of the benchmark study.

community detection phase, identified subnetworks often include many low-scoring genes, potentially leading to the incorporation of many genes that are not related to cancer.

Hybrid approach: The methods in this category, which include FDRnet, NetMix2, NetCore, and ClustEx, employ a combination of diffusion-based techniques and other approaches to integrate network structures with gene scores. For these methods, the initial step involves a diffusion process to exploit the network structures, and then the outcomes are utilized to guide the identification of subnetworks. Specifically, FDRnet begins by generating a local graph through diffusion and then searches within the local graph for a subnetwork that minimizes the conductance score, a metric often used in community detection. In a similar way, NetMix2 constructs a graph via diffusion and then identifies a subnetwork that maximizes the total gene score. NetCore enhances the neighborhood-based approach by incorporating diffusion-based significance scores into the process of adding genes to subnetworks. Finally, ClustEx applies traditional clustering algorithms to diffusion-based similarity metrics to identify clusters.

Experimental protocol

To evaluate the performance of network-based methods for the detection of significantly perturbed subnetworks in cancer, we devised an experimental protocol that employed extensive synthetic datasets. Figure 1 presents an overview of the benchmark study. Typically, a network-based method takes a PPI network and a set of synthetic *P*-values as input and outputs a list of subnetworks as potential cancer pathways. For PPI networks, we relied on four well-maintained and commonly used PPI networks, namely BioGRID [25], iRefIndex [26], ReactomeFI [27], and STRING [28]. For the STRING database, we retained only high-confidence interactions with confidence scores ≥ 0.9, following the guideline provided by the STRING database [28]. This strategy was also adopted in previous cancer studies [7, 8].

To generate synthetic *P*-values, we employed a two-step process. First, we selected target subnetworks in a PPI network, and then sampled *P*-values for all genes based on whether they were included in target subnetworks. To select biologically meaningful target subnetworks, we utilized two manually curated gene-set databases with distinct topological features: Reactome [27], a biological pathway database, and CORUM [29], a protein complex database where gene sets tend to be more densely connected. We started by extracting known cancer-related subnetworks from the two databases. Then, we pre-processed the subnetworks by excluding gene sets that were too small (fewer than 10 genes) and those having substantial overlaps with others (overlapping with other sets by more than 80%). We obtained only a limited number of subnetworks (12 subnetworks from CORUM, 10 from Reactome), which restricted the statistical power of our analysis. To address this issue, we expanded our selection to include target subnetworks from the full databases, including both cancer-related and general biological pathways and protein complexes. Using the same preprocessing procedure, we obtained 678 and 97 gene sets from the Reactome and CORUM databases, respectively. Then, we randomly sampled 10 gene sets as target subnetworks. The sampling process was repeated five times for each database, yielding a total of 10 sets of target subnetworks where each set contained 10 subnetworks. In Section "Performance comparison," we empirically demonstrated that the network-based methods performed similarly regardless of whether cancer-related or general biological pathways were used as target subnetworks. Hence, in the subsequent analysis to assess the potential biases of a method toward certain topological features of genes and subnetworks, we focused only on the results obtained using the general biological pathways. As we used the ReactomeFI PPI network, one concern was that using gene sets from the Reactome databases as target subnetworks might introduce bias. In Section "Performance comparison," we empirically demonstrated that this was not the case.

Given a set of target subnetworks, we employed a signal-to-noise decomposition model [7, 23, 30] to generate *P*-values for all genes in a PPI network. Specifically, we assumed that the *P*-value distribution is a mixture of two distributions: the signal distribution (where *P*-values arise from the alternative hypothesis, i.e. a gene is cancer-related) and the noise distribution (where *P*-values are derived from the null hypothesis, i.e. a gene is not cancer-related). It is well established that *P*-values from the null hypothesis follow a uniform distribution, $U(0, 1)$ [30]. Under the alternative hypothesis, the distribution of *P*-values is characterized by a high density at values close to zero, which decreases as the *P*-values increase. This distribution aligns with a specific form of the beta distribution $beta(a, 1)$. Thus, for each gene, we sampled its *P*-value from $beta(a, 1)$ if it was in a target subnetwork; otherwise, we used $U(0, 1)$. In the beta distribution, the parameter a determines the signal strength, with a smaller a corresponding to a larger signal strength. To assess the performance of a method applied to data with different signal strengths, we varied the values a from 0.01 to 0.11 with increment of 0.01. To minimize random variations, for each value of a, we repeated the sampling process 10 times. This resulted in 1100 *P*-value sets for testing. Using different combinations of 1100 sets of *P*-values and four PPI networks, in total, we conducted 4400 experiments for each method.

In addition to the simulation study, we conducted an experiment using the mutation and copy number data of nine cancers obtained from the TCGA study, including bladder cancer (BLCA), breast cancer (BRCA), colorectal adenocarcinoma (COADREAD), head and neck squamous cell carcinoma (HNSC), pan-kidney cohort (KIPAN), lung adenocarcinoma (LUAD), lung squamous cell

carcinoma (LUSC), prostate adenocarcinoma (PRAD), and uterine corpus endometrial carcinoma (UCEC). We followed the pipeline described in [7] to integrate both mutation and copy number data and calculate the *P*-value of each gene. We used the four PPI networks in the cancer experiment.

We set the parameters for each method based on the recommendations from their original papers. For methods that require a predefined list of putative cancer genes (DOMINO, DIAMOnD, NetCore, and ClustEx), we performed an empirical Bayes analysis [31] on *P*-values to estimate the probabilities of individual genes being false discoveries (i.e. not related to cancer) and selected genes with probabilities below a specified threshold. In this study, we set the threshold to 0.1, aligning with the default threshold utilized by FDRnet. When default parameters were not provided for a particular method, we manually selected a range for each parameter and conducted multiple tests to explore the influence of various parameter settings. We then used the setting with the best performance in comparison.

Performance comparison

We compared the ability of the 12 methods to detect target genes and subnetworks. Ideally, a good detection method should be able to not only identify correct target genes but also place them into correct subnetworks. Furthermore, it is crucial to control for the presence of nontarget genes (i.e. false positives) within each identified subnetwork to ensure their relevance to the task of subnetwork identification. For this purpose, we employed three metrics. First, we assessed the ability of a method to identify target genes (i.e. genes included in target subnetworks) using the *F*-score [32] computed by comparing identified and target genes. Secondly, we assessed the ability of a method to identify target subnetworks using the F_{sub} score, a natural extension of the *F*-score, that was introduced in [7] and specifically designed to evaluate identified subnetworks with respect to target ones. Lastly, to measure the ability of a method to control false positives, we calculated the false discovery rate (FDR) for identified subnetworks, which was defined as the proportion of nontarget genes in identified subnetworks.

First, we assessed the overall performance of each method evaluated on 4,400 test datasets. Figure 2(a–c) presents the *F*-scores, F_{sub} scores, and FDRs of the 12 methods as a function of the signal strength parameter a. As expected, for both *F*-scores and F_{sub} scores, the performance of all the methods dropped as the signal strength became weaker (i.e. with increasing values of a). Nevertheless, the relative ranking of the methods largely remains unchanged. Notably, in terms of *F*-scores, a group of methods (FDRnet, MuST, BioNet, ROBUST, NetMix2, DIAMOnD, NetCore) performed significantly better than all other methods. However, in terms of F_{sub} score, all these methods, except for FDRnet, performed poorly. While effective in identifying individual genes, these methods struggled to accurately determine subnetwork structures. In fact, they tended to connect identified genes into a small number of subnetworks, often resulting in one disproportionately large subnetwork (see Supplemental Table 3). Further examination revealed that while some methods consistently controlled FDRs at a certain level (FDRnet, BioNet, NetMix2, hierarchical HotNet, HotNet2), four methods (DIAMOnD, DOMINO, RegMOD, and ClustEx) appeared less effective in controlling FDR. Notably, DOMINO did not perform well across all three metrics in our evaluation. This may be due to the fact that the method first partitions a PPI network into communities and then identifies subnetworks within each detected community. While this could yield functionally cohesive groups, it may not work well in scenarios where target subnetworks do not align with pre-defined communities, potentially leading to the inclusion of many nontarget genes and the formation of suboptimal subnetworks. Based on the above results, we concluded that while many methods can effectively identify individual genes, if one considers both gene and subnetwork identification, FDRnet should be initially considered.

We performed an experiment where we applied the network-based methods to the cancer-related subnetworks extracted from the Reactome and CORUM databases and reported the *F*-scores, F_{sub} scores and FDRs (Supplemental Figure 1). We found that all the methods performed similarly regardless of whether cancer-related or general biological pathways were used as target subnetworks. Specifically, the same group of methods performed significantly better than all other methods in terms of *F*-score and FDRnet had the best performance in terms of F_{sub} score. This result suggests that general biological pathways can be used as suitable proxies for cancer-related pathways in assessing method performance. Consequently, we will use the results derived from the general biological pathways in our subsequent analyses since the large number of experiments allows for more reliable detection of true differences in method performance and leads to robust and generalizable conclusions.

We further validated our observations on the mutation and copy number data obtained from nine cancers. For BioNet, it failed to fit the distributions of *P*-values and thus it could not generate any results. For other methods, we estimated the FDRs of the identified subnetworks using the local FDR-based procedure described in [7] since there was no ground truth information to calculate exact FDRs (Supplemental Fig. 2). Based on the estimated FDRs, we excluded three methods from the performance evaluation—DIAMOnD (0.78), RegMOD (0.91), and ClustEx (0.94)—as their average FDRs exceeded 0.5, indicating an inclusion of too many irrelevant genes in their results. For the rest of the methods, we used cancer-related subnetworks from the CORUM and Reactome databases as proxy standards to indirectly evaluate the performance of these methods in terms of *F*-score and F_{sub} (Supplemental Fig. 3). We found that, although the results varied across cancers, the conclusions drawn from the synthetic data remained largely valid. For example, FDRnet, MuST, ROBUST, and NetCore achieved the highest *F*-scores in most cases, and all methods, except for FDRnet, performed poorly in terms of F_{sub} score. These findings further confirm the effectiveness of FDRnet in both synthetic and real datasets, highlighting its superiority in accurately identifying cancer-related subnetworks.

Next, we examined how the 12 methods performed when target subnetworks were extracted from two different databases. Figure 3(a and b) presents the *F*-scores and F_{sub} scores obtained by the 12 methods applied to test datasets where target subnetworks were derived from the CORUM or Reactome database ($a = 0.01$). First, we can see that the *F*-scores and F_{sub} scores obtained using the two databases differ for each method. However, the relative ranking of the methods remained unchanged. This suggested that none of the methods obtained a disproportionate advantage from using a specific type of target subnetwork. Moreover, for most methods, *F*-scores obtained using the CORUM or Reactome database did not exhibit a marked difference, suggesting that the identification of target genes is not significantly influenced by the source dataset. However, for FDRnet, which performed the best in terms of F_{sub} score, we noted that its performance on CORUM was significantly better than on Reactome (average F_{sub} score: 0.7 versus 0.55; *P*-value < 0.0001, two-tailed *t*-test). Given

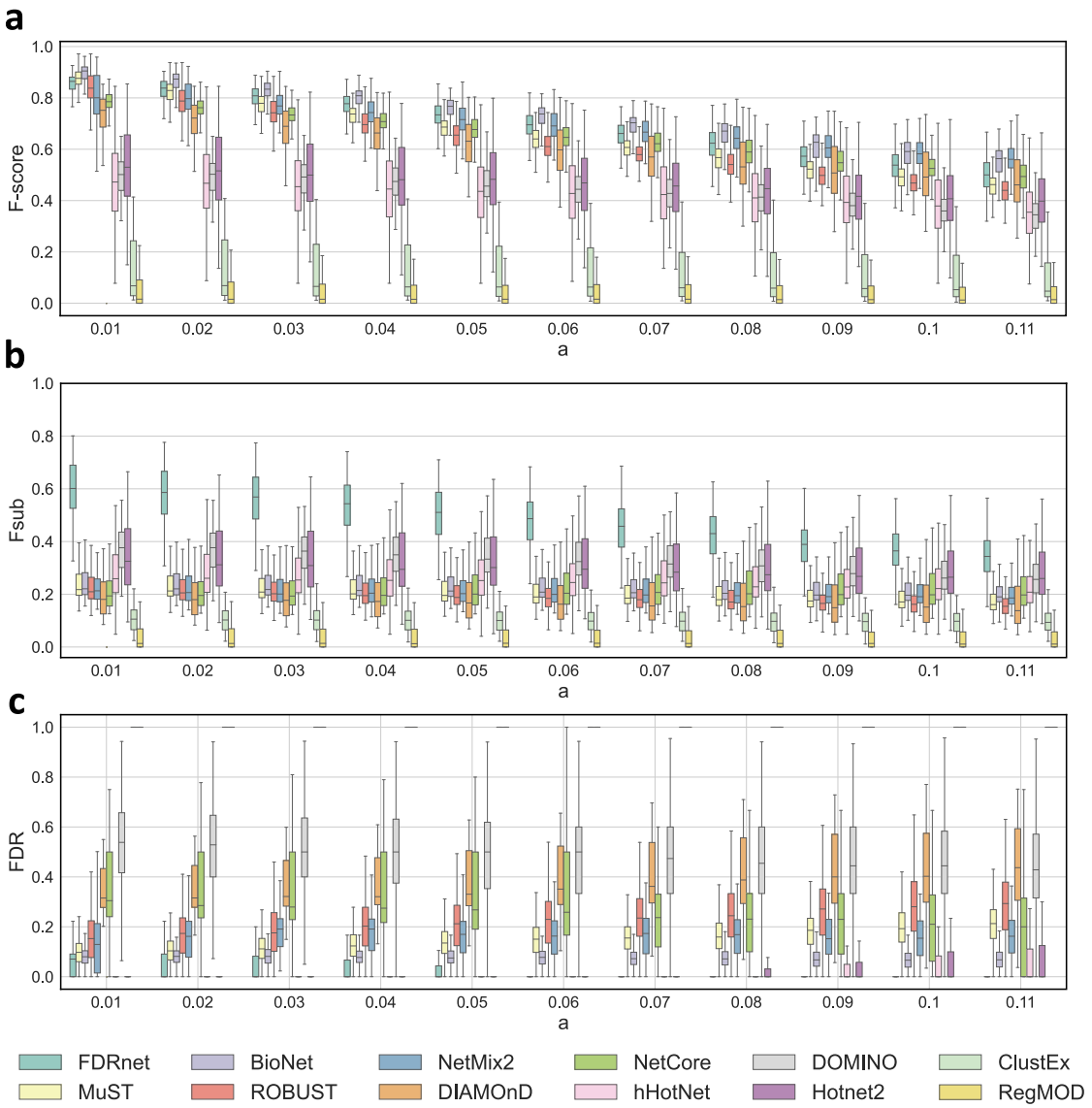

Figure 2. Comparison of 12 methods performed on test datasets generated using different signal strength parameters ranging from 0.01 to 0.11. (a) *F*-score. (b) F_{sub}. (c) FDR. In some cases, the resulting FDRs were very close to zeros (hHotNet: a = 0.01 to 0.08, HotNet2: a = 0.01 to 0.07, FDRnet: a = 0.06 to 0.11) or all ones (ClustEx and RegMOD: a = 0.01 to 0.11). hHotNet: hierarchical HotNet.

that target subnetworks from the Reactome database are typically pathways with sparser connections than those from the CORUM database, this observation highlights the need to improve existing algorithms in detecting pathway-like subnetworks. See Section "Discussion and conclusion" for a detailed discussion on possible directions for algorithm development.

We also investigated how the 12 methods performed when different PPI networks were used. Figure 3(c and d) reports the *F*-scores and F_{sub} scores of various methods applied to the four PPI networks (a = 0.01). For most methods, in terms of *F*-score, the differences in performance obtained using different PPI networks are marginal, with BioGRID and iRefIndex showing slightly better and more stable results (e.g. for FDRnet, 0.87(0.019) on BioGRID, 0.86(0.025) on iRefIndex, 0.85(0.043) on ReactomFI, 0.83(0.051) on STRING). However, for some diffusion-based methods (e.g. NetMix2 and HotNet2) an opposite trend was observed. One possible explanation is that these methods depend on densely connected structures for gene identification, which are more prevalent in ReactomeFI and STRING networks compared with the other two networks. This result is also supported by the F_{sub} scores, where we observed a better result for all the methods that used ReactomeFI and STRING and can identify multiple subnetworks (e.g. FDRnet and HotNet2). The above observations indicate that the four PPI networks have different topological structures that may affect the performance of a detection method. This motivated us to conduct an in-depth analysis on the topological bias, described in Section "Impact of topological features of subnetworks on detection rates."

Finally, we examined whether detecting target subnetworks constructed from the Reactome database in the ReactomeFI PPI network would introduce any bias. Supplemental Fig. 4 showed that this is not the case. Only for FDRnet were the results obtained using ReactomeFI superior to those obtained using other PPI networks in terms of F_{sub} score, but this difference was not statistically significant (0.62(0.08) on ReactomeFI versus 0.60(0.06) on STRING; *P*-value = 0.16, two-tailed *t*-test).

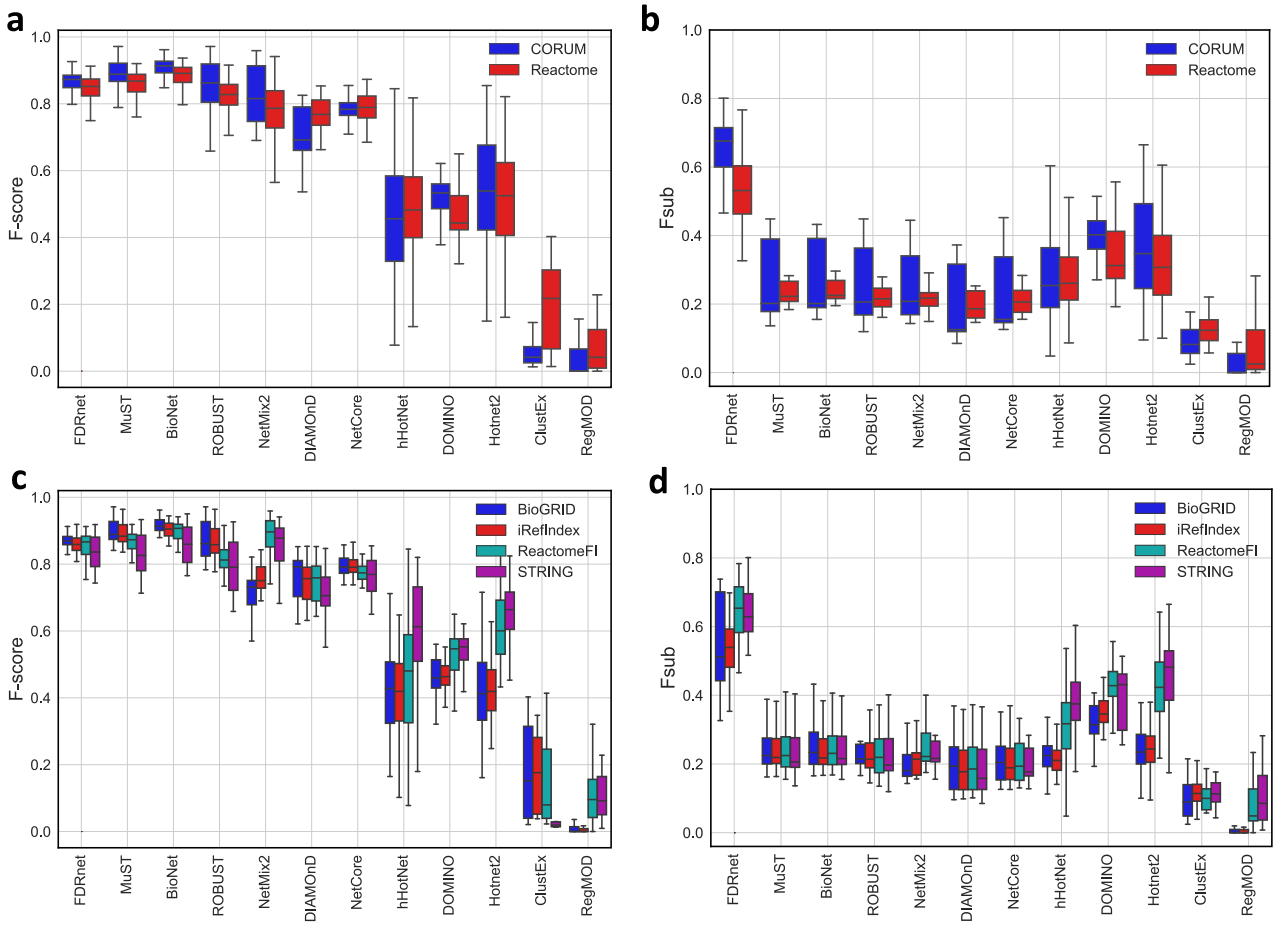

Figure 3. Comparison of 12 methods applied to detect target subnetworks extracted from two different databases in four different PPI networks. (a and b) *F*-scores and F_{sub} scores obtained when target subnetworks were derived from the CORUM or Reactome database. (c and d) *F*-scores and F_{sub} scores obtained when four different PPI networks were used. hHotNet: hierarchical HotNet.

Impact of topological features of genes on detection rates

Our benchmarking framework enables us to investigate how the topological features of a gene affect its chance of being detected. To this end, for each gene in a target subnetwork, we calculated a topological feature of the gene (e.g. degree) in a PPI network and its detection rate across experiments and performed a regression analysis to reveal the relationship between the detection rate and the topological feature. In this study, we used four widely recognized topological features, namely, degree (the number of interactions with other nodes), betweenness centrality (a measure of how often a node appears on the shortest path between each pair of nodes in a network), eigenvector centrality (a measure of the influence of a node in a network), and clustering coefficient (a measure of the degree to which nodes in a network tend to cluster together) [33]. We should point out that, except for degree, the analysis of the other features was previously infeasible, due to the fact that the past studies relied on a permutation-based procedure where it is difficult to generate networks with nodes having specific topological features. To calculate the detection rate of a gene, we counted the number of times the gene was detected and normalized it by the number of experiments where the gene was included in a target network. For the regression analysis, we employed the Lowess algorithm [34] for its efficacy and robustness.

We started with the analysis of the degree, a fundamental topological feature in network analysis. It is well documented that existing PPI networks are subject to technical bias [35] (bait proteins often exhibit more interactions) and literature bias [36] (proteins with well-characterized functions are more frequently studied). As a result, some proteins may be represented inappropriately with higher degrees [37] and have a higher probability of being detected. Thus, it has been suggested to explicitly or implicitly penalize genes with high degrees and thus reduce their chances of being detected. Figure 4 depicts the detection probability of a gene as a function of its degree in one of the four PPI networks for seven methods, including BioNet, HotNet2, DIAMOnD, DOMINO, NetCore, NetMix2, and FDRnet. BioNet was selected as a representative for the three maximum score-based methods due to their methodological similarity. Likewise, HotNet2 was chosen to represent both itself and hierarchical HotNet. We excluded ClustEx and RegMOD from the analysis because of their low mean *F*-scores (< 0.1). Visually, the observed patterns can be categorized into two classes: a steep downward trend for DOMINO and HotNet2 (class 1), and an almost flat or slight decline trend for FDRnet, BioNet, NetMix2, NetCore, and DIAMOnD (class 2). For the methods in class 1, we can see that the detection probabilities approach almost zero as the degree increases. This suggests that these methods might fail to identify important cancer genes with high degrees. Although the observed

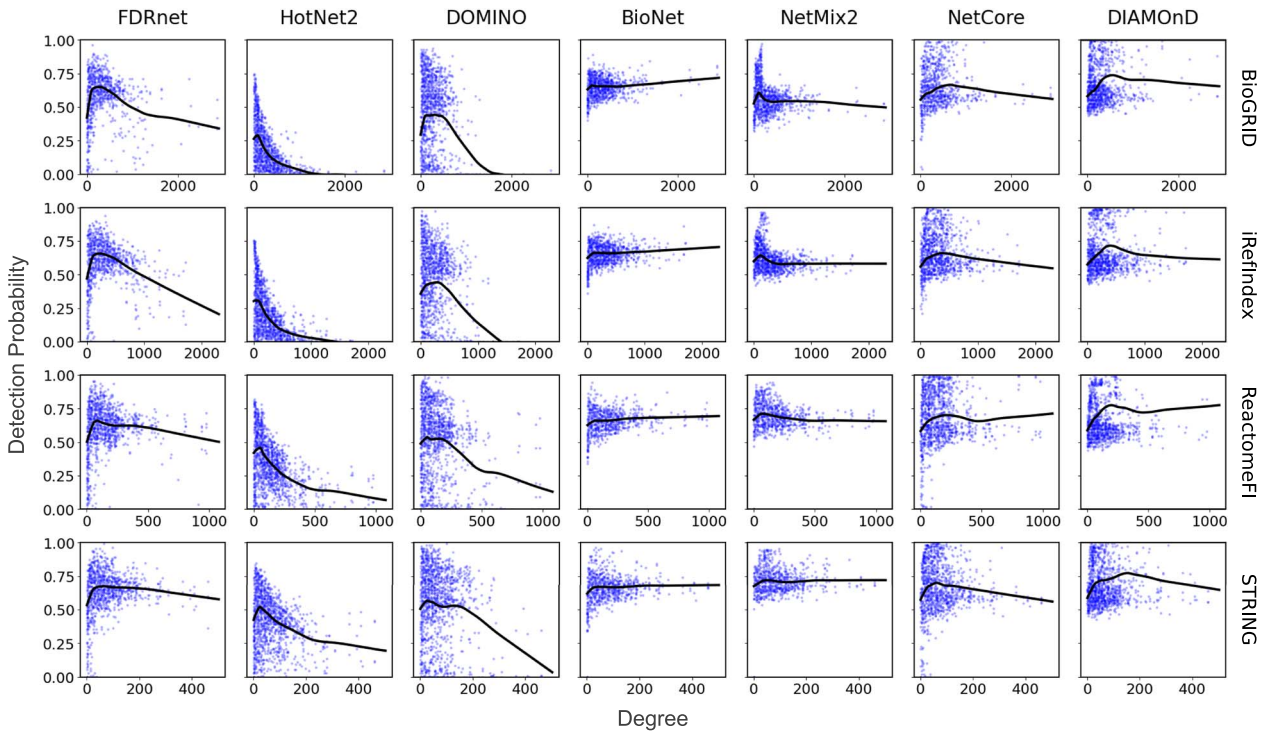

Figure 4. The detection probability of each gene for seven methods as a function of its degree in one of four PPI networks. Each dot represents a gene, and the black line was determined by the regression analysis.

patterns for the methods in class 2 are similar, a close look at detected subnetworks revealed that, except for FDRnet, all other methods tend to group all detected genes into only one or a few subnetworks through hub genes of high degree (see Supplemental Table 3), resulting in low F_{sub} scores (Fig. 2b). In contrast, by minimizing the conductance score, FDRnet implicitly penalizes high-degree genes, yet maintains a high detection probability for those genes if they are significantly mutated. This also explains why the detection probability of FDRnet did not decrease significantly as the degree increased when the ReactomeFI and STRING networks were used. As shown later in Section "Impact of topological features of subnetworks on detection rates," the same subnetwork may exhibit a more community-like structure in ReactomeFI and STRING compared with BioGRID and iRefIndex. Consequently, penalization is less stringent in ReactomeFI and STRING than in BioGRID and iRefIndex, allowing for a higher detection probability of high-degree genes.

Regarding other topological features, we anticipate similar behaviors since these measures are highly correlated with degree [38]. This is indeed the case for betweenness centrality (Supplemental Fig. 5). For the clustering coefficient, we noted that the detection probability for the methods in class 1 (i.e. DOMINO and HotNet2), which impose penalties on high-degree nodes, increases with it (Supplemental Fig. 6). By definition, the clustering coefficient of a given gene quantifies the extent to which its neighbors form a highly connected cluster [39]. The observed pattern implies a preference of these methods for genes in denser network regions, which is an expected consequence of penalizing nodes with high degrees. Thus, the pattern is also consistent with that observed for degree. However, for eigenvector centrality, while the result generally aligned with that for degree in most methods when BioGRID and iRefIndex were used as input PPI networks, notable differences were observed when ReactomeFI or STRING were used (Fig. 5). Specifically, for genes with high eigenvector centrality, the detection probability for HotNet2 does not decline with increasing centrality, and for DOMINO and DIAMOnD the detection probability even increases. To understand this phenomenon, we examined the correlation between degree and eigenvector centrality across the four PPI networks (Supplemental Fig. 7). While BioGRID and iRefIndex displayed clear linear correlations between these two measures, there are some genes in ReactomeFI and STRING that have low degrees but high eigenvector centrality. By the definition of eigenvector centrality, these low-degree genes are typically connected to some high-degree genes [40]. This connection suggests that, although indirect, their positions in a PPI network are significantly influenced by the degree bias. Therefore, researchers should take this indirect impact of the degree bias into consideration when designing algorithms to counteract it.

Impact of topological features of subnetworks on detection rates

Finally, we examined how the topological features of a subnetwork in a PPI network influence its chance of being detected by each method. Similar to the analysis for individual genes, we calculated the detection rates and topological features of target subnetworks and performed a regression analysis to determine their statistical relationships. Following a seminal study [41], we selected four subnetwork topological features: separability (measured as the internal-to-external edge ratio of a subnetwork [42]), density (the proportion of actual to possible edges in a subnetwork [43]), cohesiveness (using conductance score; calculated as the ratio of external to internal edges in a subnetwork [44]), and clustering (measured by the average clustering coefficient of the nodes in a subnetwork [39]). To determine the detection probability of a target subnetwork, we utilized the F-score to account for both

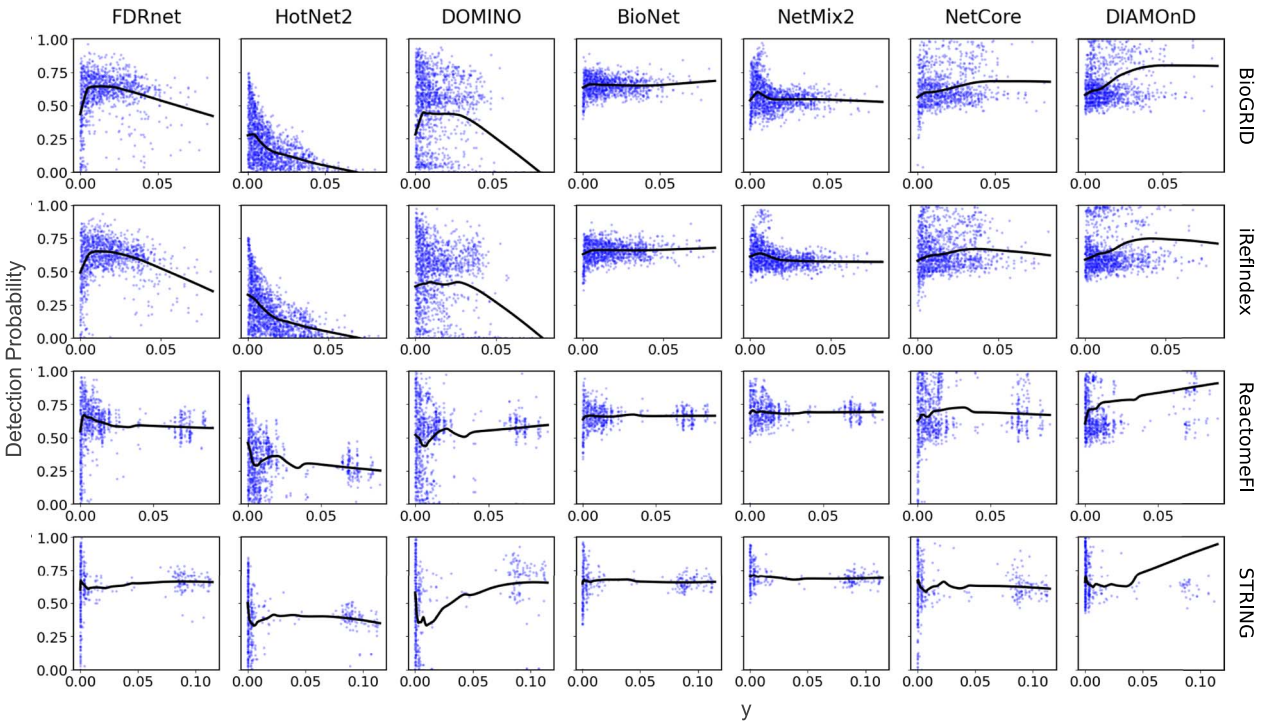

Figure 5. The detection probability of each gene for seven methods as a function of its eigenvector centrality in one of four PPI networks.

partial coverage and false positives in identified subnetworks. Specifically, we calculated the *F*-score between the target subnetwork and each identified subnetwork and used the highest score as the detection proportion. The overall detection probability for a given subnetwork was then calculated by normalizing the sum of these detection proportions across all datasets where the subnetwork was included as a target. Again, we employed the Lowess algorithm for regression analysis.

Figure 6 illustrates the regression results correlating the four topological features with the detection probability obtained by the four methods (FDRnet, HotNet2, DOMINO, hierarchical HotNet) applied to the four PPI networks. We excluded the methods that achieved low mean F_{sub} scores (ClustEx and RegMOD, See Fig. 2(b)), and those unable to identify subnetwork structures (MuST, BioNet, ROBUST, NetMix2, DIAMOnD, and NetCore, See Supplemental Table 3). The regression results show that the four methods have a preference for detecting subnetworks characterized by high separability, density, cohesiveness (low conductance), and clustering coefficient. This indicates that the four methods have a significant bias toward recognizing community-like structures, also known as topological modules. However, previous research has cautioned that although disease modules often overlap with topological modules, they are not identical [16]; disease modules are local clusters of disease-associated genes, while topological modules are local clusters of genes without regard to disease associations [16]. Therefore, it is crucial to consider this distinction in future algorithm development to enhance the accuracy in identifying disease modules.

Our analysis explains why the performance of FDRnet varied when subnetworks were derived from different databases, as shown in Fig. 3(b). To this end, we examined the distributions of the four topological metrics of the subnetworks derived from the Reactome and CORUM databases (Supplemental Fig. 8). Notably, the distributions of the density for subnetworks from the CORUM database center around 0.5, while the distributions for the Reactome database are markedly skewed toward the lower end. Thus, as FDRnet has a preference for subnetworks with higher densities (Fig. 6), the better performance on CORUM are expected.

Our analysis also explains why the existing methods performed differently when different input PPI networks were used, as shown in Fig. 3(c and d). A key observation in Supplemental Fig. 8 is that, in most cases, the topological features of the same sets of target subnetworks exhibited flatter distributions in ReactomeFI and STRING, compared with those in BioGRID and iRefIndex. This suggests that subnetworks in ReactomeFI and STRING are more inclined to form community-like structures compared with those in BioGRID and iRefIndex. Since we have shown that community-like structures have more chances to be detected, it is reasonable to conclude that using ReactomeFI or STRING as an input PPI network is more likely to lead to improved F_{sub} scores.

Discussion and conclusion

Our study demonstrated the complexity of identifying cancer genes and subnetworks, a task influenced by multiple factors. These factors include not only the selection of a detection algorithm, but also the input PPI network and the structural features of target genes and subnetworks. For researchers who aim to identify cancer genes and subnetworks, our findings support several recommendations. First, the choice of a detection algorithm is of paramount importance. Our data showed that the performance can vary significantly among different methods. For those who focus solely on gene identification, methods based on maximum scores have proven to be particularly effective due to their high accuracy in discerning individual genes. However, since they do not penalize high-degree genes, there is a risk that the results are affected by the construction bias of PPI networks. For tasks that aim to identify both genes and subnetworks, FDRnet consistently emerged as the best method across various settings.

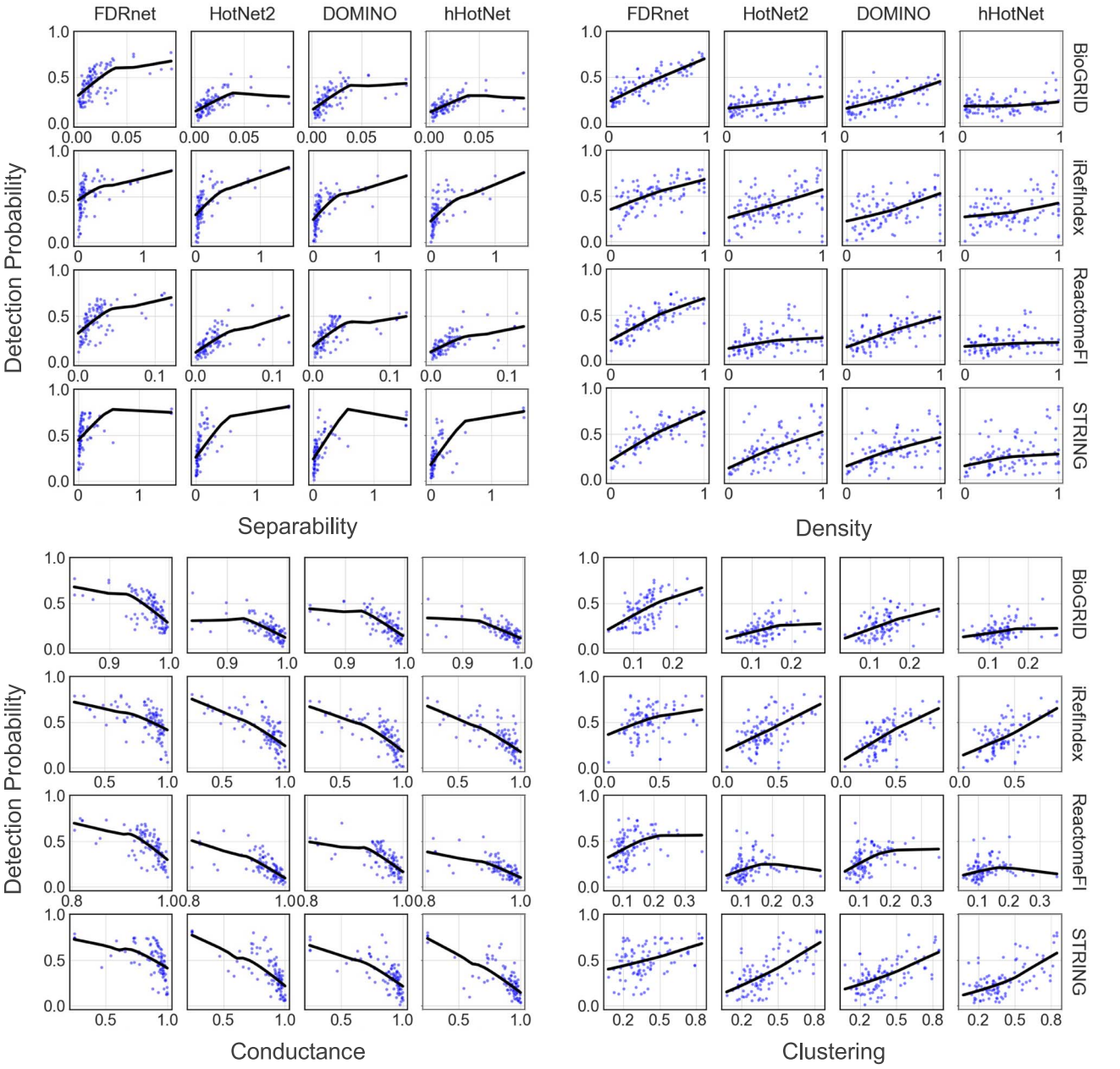

Figure 6. The detection probability of a subnetwork for four methods as a function of its topological features including separability, density, conductance, and clustering metrics. Each dot represents a subnetwork, and the black line was determined by the Lowess regression analysis. hHotNet: hierarchical HotNet.

Secondly, the choice of the input PPI network is not to be overlooked. Different PPI networks, constructed from distinct data sources and based on different principles, have unique topological structures. This is especially crucial for subnetwork identification, given that all methods that we tested tend to prefer community-like subnetworks. In this context, we recommend using the ReactomeFI and STRING networks, as subnetworks mapped to these networks are more likely to exhibit community-like structures. Finally, having some prior knowledge or hypotheses about genes or subnetworks that one aims to detect can be advantageous. For example, if target subnetworks to be detected are more like signaling pathways than protein complexes, performance often degrades. This is particularly relevant given that FDRnet—the best-performing method on both types of data—achieved a higher F_{sub} score for protein complexes compared with pathways. Additionally, if it is anticipated that some genes with high degrees may play important roles in a disease, methods such as HotNet2 and DOMINO may not be suitable, as they tend to exclude such genes even if they show significant connection with the disease in a gene-based analysis. In contrast, FDRnet imposes a soft penalty on high-degree genes, thereby alleviating the above issue. Looking ahead, there are several exciting avenues for the development of algorithms in this domain. First, our results clearly show that existing methods do not perform well on noncommunity-like subnetworks, such as those found in the Reactome pathway database. Our analysis of the topological bias indicates that this limitation arises because all existing methods favor community-like structures. Given that disease modules and topological modules are not identical, there is a pressing need to develop algorithms

capable of better identifying noncommunity-like subnetworks. Exploring topological features other than community-like ones may yield more insightful characterizations of target subnetworks. For example, pathway structure may play an important role in mitigating disruptive effects of gene mutations, thereby preserving the robustness of biological systems. Secondly, our findings highlight that overcoming the degree bias remains a persistent challenge. This underscores the need for strategies that can effectively penalize high-degree genes rather than unfairly excluding them. In addition to degree, eigenvector centrality, which provides a more intricate view of the influence of a gene on its connectivity pattern, also merits attention. Future algorithm development should incorporate this metric to ensure that the degree bias is appropriately addressed.

In summary, we have presented a comprehensive benchmarking study of subnetwork identification methods, utilizing a ground truth-based approach. We anticipate that the results will guide the selection of proper methods for cancer pathway identification and inspire the development of new algorithms. While our ground truth-based strategy offers a practical benchmarking framework, it is important to recognize that it introduces an inherent bias, arising from the selection of target subnetworks from general databases and from incomplete knowledge about cancer-related subnetworks. Nevertheless, our approach provides a valid reference point, and the databases that we have carefully chosen are likely to be representative of subnetworks that may have roles in cancer progression. In future work, we will extend this benchmarking study to include additional methods and various types of networks.

Key Points

- We proposed a novel benchmarking framework using synthetic data for the comparison of existing methods for the identification of cancer pathways.
- We conducted an in-depth analysis to investigate the ability of existing methods to detect target genes and subnetworks and to control false positives, and how they perform in the presence of topological biases at both gene and subnetwork levels.
- Our analysis provided insights into algorithmic performance that were previously unattainable and presents findings that challenge established views.
- We presented a practical guide for users on how to select appropriate detection methods and protein–protein interaction networks for cancer pathway identification and provided suggestions for future algorithm development.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions.

Author contributions

Le Yang (Conceptualization, Investigation, Formal analysis, Writingreview & editing), Runpu Chen (Formal analysis, Writingreview & editing), Steve Goodison (Formal analysis, Writingreview & editing), and Yijun Sun (Conceptualization, Formal analysis, Writingreview & editing)

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This research is supported in part by National Institutes of Health (R01CA241123 to Y.S. and S.G.), National Institutes of Health (R01CA266113 to S.G. and Y.S.), and National Institutes of Health (R01CA269075 to Y.S. and S.G.)

Data availability

All the datasets used in this study are publicly available and accessible. The BioGRID 4.4.212 PPI network, the iRefIndex 18.0 PPI network, and the STRING 11.5 network were downloaded from https://downloads.thebiogrid.org/BioGRID, https://irefindex.vib.be/wiki/index.php/iRefIndex, and https://string-db.org/, respectively. The Reactome pathway database and the ReactomeFI network were downloaded from https://reactome.org/download-data. The CORUM database was downloaded from https://mips.helmholtz-muenchen.de/corum/. All the TCGA datasets were downloaded from the TCGA firehose website (https://gdac.broadinstitute.org/).

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell* 2011;**144**:646–74. https://doi.org/10.1016/j.cell.2011.02.013.
2. Weinstein JN, Collisson EA, Mills GB. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20. https://doi.org/10.1038/ng.2764.
3. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* 2020;**578**:82–93. https://doi.org/10.1038/s41586-020-1969-6.
4. Lawrence MS, Stojanov P, Polak P. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;**499**:214–8. https://doi.org/10.1038/nature12213.
5. Tomczak K, Czerwińska P, Wiznerowicz M. Review the cancer genome atlas (TCGA): An immeasurable source of knowledge. *Contemp Oncol* 2015;**2015**:68–77.
6. Leiserson MDM, Vandin F, Hsin-Ta W. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* 2015;**47**:106–14. https://doi.org/10.1038/ng.3168.
7. Yang L, Chen R, Goodison S. *et al.* An efficient and effective method to identify significantly perturbed subnetworks in cancer. *Nat Comput Sci* 2021;**1**:79–88. https://doi.org/10.1038/s43588-020-00009-4.
8. Reyna MA, Haan D, Paczkowska M. *et al.* Pathway and network analysis of more than 2500 whole cancer genomes. *Nat Commun* 2020;**11**:1–17.
9. Levi H, Elkon R, Shamir R. DOMINO: A network-based active module identification algorithm with reduced rate of false calls. *Mol Syst Biol* 2021;**17**:e9593. https://doi.org/10.15252/msb.20209593.
10. Reyna MA, Leiserson MDM, Raphael BJ. Hierarchical HotNet: Identifying hierarchies of altered subnetworks. *Bioinformatics* 2018;**34**:i972–80. https://doi.org/10.1093/bioinformatics/bty613.

11. Barel G, Herwig R. NetCore: A network propagation approach using node coreness. *Nucleic Acids Res* 2020;**48**:e98–8. https://doi.org/10.1093/nar/gkaa639.
12. Bernett J, Krupke D, Sadegh S. *et al.* Robust disease module mining via enumeration of diverse prize-collecting Steiner trees. *Bioinformatics* 2022;**38**:1600–6. https://doi.org/10.1093/bioinformatics/btab876.
13. Chitra U, Park TY, Raphael BJ. NetMix2: A principled network propagation algorithm for identifying altered subnetworks. *J Comput Biol* 2022;**29**:1305–23. https://doi.org/10.1089/cmb.2022.0336.
14. Yang L, Chen R, Melendy T. *et al.* Identifying significantly perturbed subnetworks in cancer using multiple protein–protein interaction networks. *Cancer* 2023;**15**:4090. https://doi.org/10.3390/cancers15164090.
15. Lazareva O, Baumbach J, List M. *et al.* On the limits of active module identification. *Brief Bioinform* 2021;**22**:bbab066. https://doi.org/10.1093/bib/bbab066.
16. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. *Nat Rev Genet* 2011;**12**:56–68. https://doi.org/10.1038/nrg2918.
17. Ideker T, Ozier O, Schwikowski B. *et al.* Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002;**18**:S233–40. https://doi.org/10.1093/bioinformatics/18.suppl_1.S233.
18. Beisser D, Klau GW, Dandekar T. *et al.* BioNet: An R-package for the functional analysis of biological networks. *Bioinformatics* 2010;**26**:1129–30. https://doi.org/10.1093/bioinformatics/btq089.
19. Sadegh S, Matschinske J, Blumenthal DB. *et al.* Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nat Commun* 2020;**11**:3518. https://doi.org/10.1038/s41467-020-17189-2.
20. Qiu Y-Q, Zhang S, Zhang X-S. *et al.* Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinformatics* 2010;**11**:26. https://doi.org/10.1186/1471-2105-11-26.
21. Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol* 2015;**11**:e1004120. https://doi.org/10.1371/journal.pcbi.1004120.
22. Ding Z, Guo W, Jin G. ClustEx2: Gene module identification using density-based network hierarchical clustering. In: *2018 Chinese Automation Congress*, pp. 2407–12. Xi'an, China: IEEE, 2018.
23. Dittrich MT, Klau GW, Rosenwald A. *et al.* Identifying functional modules in protein–protein interaction networks: An integrated exact approach. *Bioinformatics* 2008;**24**:i223–31. https://doi.org/10.1093/bioinformatics/btn161.
24. Hwang FK, Richards DS. Steiner tree problems. *Networks* 1992;**22**:55–89. https://doi.org/10.1002/net.3230220105.
25. Oughtred R, Stark C, Breitkreutz B-J. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 2019;**47**:D529–41. https://doi.org/10.1093/nar/gky1079.
26. Razick S, Magklaras G, Donaldson IM. iRefIndex: A consolidated protein interaction database with provenance. *BMC Bioinformatics* 2008;**9**:405. https://doi.org/10.1186/1471-2105-9-405.
27. Fabregat A, Jupe S, Matthews L. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* 2018;**46**:D649–55. https://doi.org/10.1093/nar/gkx1132.
28. Szklarczyk D, Gable AL, Lyon D. *et al.* STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**:D607–13. https://doi.org/10.1093/nar/gky1131.
29. Giurgiu M, Reinhard J, Brauner B. *et al.* CORUM: The comprehensive resource of mammalian protein complexes–2019. *Nucleic Acids Res* 2018;**47**:D559–63.
30. Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values. *Bioinformatics* 2003;**19**:1236–42. https://doi.org/10.1093/bioinformatics/btg148.
31. Efron B, Tibshirani R, Storey JD. *et al.* Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 2001;**96**:1151–60. https://doi.org/10.1198/016214501753382129.
32. van Rijsbergen CJ. *Information Retrieval*. United Kingdom: Butterworth, 1979.
33. Newman M. *Networks*. United Kingdom: Oxford University Press, 2018.
34. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 1979;**74**:829–36. https://doi.org/10.1080/01621459.1979.10481038.
35. Stibius KB, Sneppen K. Modeling the two-hybrid detector: Experimental bias on protein interaction networks. *Biophys J* 2007;**93**:2562–6. https://doi.org/10.1529/biophysj.106.098236.
36. Schaefer MH, Serrano L, Andrade-Navarro MA. Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front Genet* 2015;**6**:137790. https://doi.org/10.3389/fgene.2015.00260.
37. Rolland T, Taşan M, Charloteaux B. *et al.* A proteome-scale map of the human interactome network. *Cell* 2014;**159**:1212–26. https://doi.org/10.1016/j.cell.2014.10.050.
38. Valente TW, Coronges K, Lakon C. *et al.* How correlated are network centrality measures? *Connections* 2008;**28**:16–26.
39. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature* 1998;**393**:440–2. https://doi.org/10.1038/30918.
40. Bonacich P. Some unique properties of eigenvector centrality. *Social Networks* 2007;**29**:555–64. https://doi.org/10.1016/j.socnet.2007.04.002.
41. Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth. *Knowl Inf Syst* 2015;**42**:181–213. https://doi.org/10.1007/s10115-013-0693-z.
42. Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2000;**22**:888–905. https://doi.org/10.1109/34.868688.
43. Fortunato S. Community detection in graphs. *Phys Rep* 2010;**486**:75–174. https://doi.org/10.1016/j.physrep.2009.11.002.
44. Leskovec J, Lang KJ, Mahoney M. Empirical comparison of algorithms for network community detection. In: *Proceedings of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA*, pp. 631–40. Association for Computing Machinery, 2010.

Briefings in Bioinformatics, 2025, **26(1)**, bbae692
https://doi.org/10.1093/bib/bbae692
Review