BB
xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

MEXCOWalk: Mutual Exclusion and Coverage Based Random Walk to Identify Cancer Modules

Rafsan Ahmed [1], Ilyes Baali [1], Cesim Erten [2], Evis Hoxha [2] and Hilal Kazan [2,*]

[1] Electrical and Computer Engineering Graduate Program, Antalya Bilim University, Antalya, 07190, Turkey and
[2] Department of Computer Engineering, Antalya Bilim University, Antalya, 07190, Turkey.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Genomic analyses from large cancer cohorts have revealed the mutational heterogeneity problem which hinders the identification of driver genes based only on mutation profiles. One way to tackle this problem is to incorporate the fact that genes act together in functional modules. The connectivity knowledge present in existing protein-protein interaction networks together with mutation frequencies of genes and the mutual exclusivity of cancer mutations can be utilized to increase the accuracy of identifying cancer driver modules.
Results: We present a novel edge-weighted random walk-based approach that incorporates connectivity information in the form of protein-protein interactions, mutual exclusion, and coverage to identify cancer driver modules. MEXCOWalk outperforms several state-of-the-art computational methods on TCGA pan-cancer data in terms of recovering known cancer genes, providing modules that are capable of classifying normal and tumor samples, and that are enriched for mutations in specific cancer types. Furthermore, the risk scores determined with output modules can stratify patients into low-risk and high-risk groups in multiple cancer types. MEXCOwalk identifies modules containing both well-known cancer genes and putative cancer genes that are rarely mutated in the pan-cancer data. The data, the source code, and useful scripts are available at: https://github.com/abu-compbio/MEXCOwalk.
Contact: hilal.kazan@antalya.edu.tr

1 Introduction

Recent advances in high-throughput DNA sequencing technology have allowed several projects such as the TCGA (Weinstein *et al.*, 2013) to construct and release genomic data from thousands of tumors. This further gave rise to the design of several computational approaches for the systematic detection of cancer-related somatic mutations.

Combining the somatic mutations data with additional information in the form of interaction networks or gene expression data, several computational approaches focus on prioritizing independent genes to provide hypothesized *candidate driver genes*, those defined as being causally linked to oncogenesis (Erten *et al.*, 2011; Lawrence *et al.*, 2013; Yang *et al.*, 2017b; Dopazo and Erten, 2017). Although such gene rankings provide valuable insight regarding potential genes of interest, in many cases mutations at different loci could lead to the same disease (Vanunu *et al.*, 2010). This genetic heterogeneity may reflect an underlying molecular mechanism in which the cancer-causing genes form some kind of functional pathways or *candidate driver modules*. Several computational methods have been suggested for the identification of candidate modules; see Dimitrakopoulos and Beerenwinkel (2017); Deng *et al.* (2017); Zhang and Zhang (2018) for recent surveys.

The module identification approaches as applied to cancer can be viewed in two broad categories based on the types of input data they employ. The *de novo* methods rely only on genetic data to discover novel genetic interactions, as well as cancer-related functional modules (Miller *et al.*, 2011; Vandin *et al.*, 2011b; Leiserson *et al.*, 2013; Liu *et al.*, 2017). Due to the large solution space such methods usually apply a prefiltering based on alteration frequency to reduce the inherent computational complexity which may reduce sensitivity by overlooking modules involving rare alterations (Deng *et al.*, 2017).

On the other hand, knowledge-based methods, in addition to genomic data, incorporate prior knowledge in the form of pathways, networks and functional phenotypes to identify driver modules. Such methods can be subcategorized based on the optimization goals set within the computational problem formulations they employ in defining the biologically motivated cancer driver module identification problem.

Since a driver pathway tends to be perturbed in a relatively large number of patients, in the first subcategory of methods including Hotnet (Vandin *et al.*, 2011a), Hotnet2 (Leiserson *et al.*, 2014), Hierarchical Hotnet (Reyna

© The Author 2015.

et al., 2018), the *coverage* of the modules as identified by the mutation frequencies of the comprising genes over a cohort of samples constitutes an informal optimization goal. Heat-diffusion over an interaction network that diffuses the mutation frequencies throughout the network is a common attribute in these methods. The resulting diffusion values are then employed to extract modules exhibiting a large degree of connectedness as formulated with an appropriate graph-theoretical connectivity definition, usually the *strong connectivity*.

The second subcategory of knowledge-based module identification methods incorporate an appropriate definition of an important concept, *mutual exclusion*, in their computational problem formulations. It refers to the phenomenon that for a group of genes which exhibit evidence of shared functional pathway, simultaneous mutations in the same patients are less frequent than is expected by chance (Yeang *et al.*, 2008). Several cancer module identification methods incorporate this observation in the employed combinatorial optimization problem definitions. In MEMo, maximal clique extraction in a similarity graph derived from an interaction network or functional relation graph is used and the maximal cliques are postprocessed taking into account the mutual exclusivity results (Ciriello *et al.*, 2012). In Babur *et al.* (2015) a method based on seed-and-growth on a network, where the growth strategy is determined with respect to a suitably defined mutual exclusion score is proposed to identify pan-cancer modules using TCGA data. BeWith proposes an ILP formulation that combines interaction density within a module and several mutual exclusion definitions as a maximization goal (Dao *et al.*, 2017). The ILP formulation incorporates constraints in the form of desired number of modules and maximum number of genes per module. MEMCover combines pairwise mutual exclusion scorses with confidence values of interactions in the network (Kim *et al.*, 2015). To maximize high-confidence interactions, mutual exclusion, and coverage simultaneously; heavy subnetworks covering every disease case at least k times are found following a greedy iterative seed-and-growth heuristic.

We propose MEXCOWalk, a knowledge-based method that incorporates protein-protein interaction (PPI) network data and mutation profiles, and employs a random walk-based approach to extract driver modules for cancer. We first provide a novel optimization problem definition for identifying driver modules, that takes into account network connectivity, mutual exclusion, and coverage. This definition, in accordance with the intuitive interpretations of mutual exclusivity and coverage of a module, normalizes their contributions by suitable functions of module size. Computational intractability of the provided optimization problem is shown for completeness. The MEXCOWalk algorithm tackles this computationally hard problem using a heuristic based on random walk on an appropriately vertex-weighted and edge-weighted graph. The former weights incorporate information regarding coverages, whereas the latter reflect a novel combination of the coverages and the degree of mutual exclusivity between gene pairs. Heat diffusion-based methods employing vertex weights to assign and diffuse gene coverages have been proposed in the context of cancer module identification in various previous work (Vandin *et al.*, 2011a; Leiserson *et al.*, 2014; Bersanelli *et al.*, 2016; Yang *et al.*, 2017a; Reyna *et al.*, 2018). To our knowledge, this is the first method to employ edge-weighted diffusion to incorporate degrees of coverage and mutual exclusion via edge weights. We furthermore provide a novel heuristic based on split-and-extend to crop large modules to sizes determined by the structure of the network, while maintaining high coverage and mutual exclusivity. We compare the output module sets of MEXCOWalk with those of three alternative knowledge-based methods and show that it provides better results in terms of recovering known cancer genes, enrichment for mutations in specific cancer types, and the accuracy in classifying normal and tumor instances.

2 Methods

In the following subsections we provide the problem definition and a description of our MEXCOWalk algorithm.

2.1 Problem Definition

We provide a novel combinatorial optimization problem definition to detect driver modules in cancer. Such a definition is not only important for algorithmic purposes but also to serve as a measure of performance for alternative methods suggested for the problem.

Let S_i denote the set of samples for which gene g_i is mutated. Let $G = (V, E)$ represent the PPI network where each vertex $u_i \in V$ denotes a gene g_i whose expression gives rise to the corresponding protein in the network and each undirected edge $(u_i, u_j) \in E$ denotes the interaction among the proteins corresponding to genes g_i, g_j. Henceforth we assume that g_i denotes both the gene and the corresponding vertex in G.

Let $M \subseteq V$ be a set of genes denoting a *module*. We define the mutual exclusion of M as,

$$MEX(M) = \frac{|\bigcup_{\forall g_i \in M} S_i|}{\sum_{\forall g_i \in M} |S_i|}$$

and the coverage of M as,

$$CO(M) = \frac{|\bigcup_{\forall g_i \in M} S_i|}{|\bigcup_{\forall g_i \in V} S_i|}.$$

Let $P = \{M_1, M_2, \ldots M_r\}$ be a set of modules. Let $RS(M_q)$ denote the relative size of a module M_q with respect to the total size, that is $RS(M_q) = \frac{|M_q|}{|\cup_{\forall M_t \in P} M_t|}$. We define the mutual exclusion score and the coverage score of a set of modules, so that each module M_q contributes its share proportional to its relative size $RS(M_q)$ for the former, whereas for the latter the contribution of M_q is proportional to the normalized value of $1 - RS(M_q)$. Intuitively, a large module with high mutual exclusion score should be rewarded, since as the size of the module increases the chances of achieving better mutual exclusion decrease. Analogously, a small module with high coverage score should be rewarded. Thus we define the mutual exclusion score of P as,

$$MS(P) = \sum_{\forall M_q \in P} RS(M_q) \times MEX(M_q).$$

The coverage score of P is defined as

$$CS(P) = \sum_{\forall M_q \in P} \frac{1 - RS(M_q)}{\sum_{\forall M_t \in P} 1 - RS(M_t)} \times CO(M_q),$$

if $|P| > 1$ and $CS(P) = CO(M_1)$, if $|P| = 1$.

For a graph H and a set M_q of genes, let $H(M_q)$ denote the subgraph of H induced by the vertices corresponding to genes in M_q.

Cancer driver module identification problem: Given as input a PPI network G, S_i for each gene g_i, integers $total_genes$, and min_module_size, find a disjoint set of modules P that maximizes the *driver module set score* defined as,

$$DMSS(P) = MS(P) \times CS(P) \tag{1}$$

and that satisfies the following:

1. For each $M_q \in P$, $G(M_q)$ is connected.
2. $|\bigcup_{\forall M_q \in P} M_q| = total_genes$.
3. $min_{\forall M_q \in P} |M_q| = min_module_size$.

Theorem 1. *Cancer driver module identification problem is NP-hard.*

Algorithm 1 *MEXCOWalk*

Input: PPI network $G = (V, E)$, S_i for each gene g_i, integers $total_genes, min_module_size$ and threshold θ with $0 < \theta \leq 1$.
Output: Set of driver modules P.

```
//1. Weight Assignment with MEX and CO
```
construct G_w by assigning a weight to each $g_i \in V, e \in E$
```
//2. Edge-Weighted Random Walk
```
construct G_d by applying weighted-random walk on G_w
```
//3. Constructing Set of Driver Modules
//Initial Candidate Modules
```
repeat
  $P = SCC(G_d)$
  remove $M_q \in P$ with $|M_q| < min_module_size$
  remove min-weight edge from G_d
until $|\bigcup_{\forall M_q \in P} M_q| == total_genes$
```
//Split-and-extend on Large Modules
```
$split_size = \max_{\forall G_d(M_q)} outdeg(G_d(M_q))$
for each $M_q \in P$ with $|M_q| > split_size$ **do**
  remove M_q from P and let $L = \{G_d(M_q)\}$
```
  //Split
```
  while L not empty **do**
    remove G_c from L and let v' be max outdegree vertex in G_c
    remove $IN(v')$ from G_c and insert it into $leaf_q$ or $seed_q$
    for each $M_j \in SCC(G_c)$ **do**
      insert M_j into one of L, $leaf_q$, or $seed_q$
    end for
  end while
```
  //Extend
```
  for each M_i in $leaf_q$ **do**
    merge M_i with appropriate $M_j \in seed_q$
  end for
  insert modules in $seed_q$ into output set of modules P
end for

Proof. The transformation is from *Set Packing* which is shown to be NP-complete (Garey and Johnson, 1979). In the Set Packing problem, given a collection C of finite sets and a positive integer $K \leq |C|$, the problem is to find out whether C contains at least K mutually disjoint sets. The problem is NP-hard even when the size of each set is at most 3, which can easily be extended to the setting where the size of each set is exactly 3. Given an input to the Set Packing problem within this setting in the form of K and C such that for each $S \in C$, $|S| = 3$, we generate G as a complete graph on $|C|$ vertices, corresponding to the set of genes, such that each finite set in C corresponds to a set of samples S_i for which gene g_i is mutated. We set both $total_genes$ and min_module_size to K. The answer to the Set Packing problem is Yes, if and only if the maximized score of the cancer module identification problem is exactly $\frac{3 \times K}{|\bigcup_{\forall g_i \in V} S_i|}$.

2.2 MEXCOWalk Algorithm

Since the problem is computationally intractable we provide a polynomial-time heuristic approach based on vertex and edge-weighted random walks on special graphs that incorporate mutual exclusion and coverage information as vertex and edge weights in the H.Sapiens PPI network. The pseudocode of our method is provided in Algorithm 1. There are three main steps of the algorithm, each of which is described in detail in the following subsections.

2.2.1 Weight Assignment with MEX and CO

Given a PPI network $G = (V, E)$, we first construct a weighted graph G_w that contains properly defined weights for vertices and edges. For each $g_i \in V$ we assign a weight, $w(g_i) = CO(\{g_i\})$, thus the weight corresponds to the mutation frequency of a gene. It represents the heat to be diffused from that vertex during the random walk procedure.

The weight of an edge incident on a g_i should on the other hand reflect the ratio of heat transferred to g_i's neighbors at each step of the random-walk. We formulate it so as to mimic the optimization goal defined in the problem definition. One option could be to define the weight solely in terms of the gene pair g_i, g_j. However such a simple weighting scheme may not be sufficient in practice, since the co-occurrence of a pair in a module increases the chances of the genes in their neighborhoods to coexist in the same module as well. This is especially important for the contribution of mutual exclusion in the edge-weight, as pairwise mutual exclusion values are almost always close to 1. In order to reflect these observations we consider an edge-weighting scheme where contribution of mutual exclusion is computed within the vertex neighborhoods. More specifically, let $N_e(g_i)$ denote the *extended neighborhood* of g_i, that is $N_e(g_i) = \bigcup_{\forall (g_i, g_j) \in E} g_j \cup \{g_i\}$. The contribution of mutual exclusion to the edge weight, $MEX_n(g_i, g_j)$ is the average of $MEX(N_e(g_i))$ and $MEX(N_e(g_j))$. Thus the weight of an edge (g_i, g_j) is defined as,

$$w(g_i, g_j) = MEX_n(g_i, g_j) \times CO(\{g_i\}) \times CO(\{g_j\}).$$

The contribution of coverage is computed as a product so as to reduce the chances of a single gene with large coverage dominating the weights of incident edges. Furthermore, it allows the algorithm to favor more balanced coverages among equal-sized modules; coverage of 100 patients with a module containing a pair of genes, one covering 99 and the other only 1, is less preferable than a module with a pair where each gene covers 50 patients. Finally we note that, to further strengthen the impact of mutual exclusion on edge-weights, we introduce a threshold θ, so that for pairs with MEX_n score less than θ, edge weights are assigned to 0.

2.2.2 Edge-Weighted Random Walk

Once G_w is constructed after vertex and edge weight assignments, we apply an insulated heat diffusion process on G_w that can also be described as a random walk with restart on the graph. The random walk starts from a gene g_s. At each time step, with probability $1 - \beta$, the random surfer follows one of the edges incident on the current node with probability proportional to the edge weights. Otherwise, with probability β, the walker restarts the walk from g_s. Here β is called the restart probability. The transition matrix that corresponds to this process is defined as follows:

$$T_{ij} = \begin{cases} \frac{w(g_i, g_j)}{\sum_k w(g_k, g_j)}, & \text{if } (g_i, g_j) \in E \\ 0, & \text{otherwise} \end{cases}$$

T_{ij} can be interpreted as the probability that a simple random walk will transition from g_j to g_i. The random walk process can then be described as a network propagation process by the equation, $F_{t+1} = (1 - \beta)TF_t + \beta F_0$, where F_t is the distribution of walkers after t steps and F_0 is the diagonal matrix with initial heat values, that is $F_0[i, i] = CO(g_i)$. One strategy to compute the final distribution of the walk is to run the propagation function iteratively for increasing t values until F_{t+1} converges (Hofree *et al.*, 2013). Another strategy, which we chose to employ in our implementation, is to solve this system numerically using the equation, $F = \beta(I - (1 - \beta)T)^{-1}F_0$ (Leiserson *et al.*, 2014). The edge-weighted directed graph G_d is constructed by creating directed edge $[g_i, g_j]$ with weight $F[i, j]$, for every pair $i \neq j$.

We note that the idea of random walks with restart has been employed in the context of cancer module identification in previous work (Vandin

et al., 2011a; Leiserson *et al.*, 2014; Bersanelli *et al.*, 2016; Yang *et al.*, 2017a; Reyna *et al.*, 2018). However as the concept of edge weights is absent, the transition probabilities in those studies are only based on the degrees of the vertices. In our case, the transition probabilities reflect the edge weights which in turn model the contribution of a pair of genes to the maximization score, when placed in the same module. Similar to the previous methods employing heat diffusion we assign $\beta = 0.4$.

2.2.3 Constructing Set of Driver Modules

We have two main steps. We employ strongly connected components (SCC) as a primitive in both of the steps, similar to Hotnet2 and Hierarchical Hotnet. We first create an initial set of candidate modules. For this, we iteratively remove the smallest weight edge from G_d, add the strongly connected components (SCC) of G_d into initial module set P, and remove all modules of size less than min_module_size from P, until the total number of genes in P decreases to $total_genes$.

Next we process large candidate modules via a *split-and-extend* procedure, where large modules to be thus processed are determined by a network property of the initial modules. We define the $split_size$ to be the outdegree of any vertex in any of the subgraphs induced by the modules. Any initial candidate module M_q of size greater than the $split_size$ goes through the split-and-extend procedure.

The idea is to first extract *seed* modules that satisfy certain size and connectivity criteria, and extend them with small *leaf* modules. Given a directed graph G_c, let $IN(v')$ denote the *isolated neighborhood* of v' in G_c, that is $w \in IN(v')$, if and only if $w \in N_e(v')$ and for any directed edge $[w, x]$ or $[x, w]$, $x \in N_e(v')$. The split phase of a module M_q consists of removing $IN(v')$ from $G_d(M_q)$, where v' is the vertex with largest degree in $G_d(M_q)$. Assuming its size is not less than min_module_size, $IN(v')$ is a seed module to be extended in the next phase, otherwise it is a leaf module that is to be attached to an appropriate seed module. The remainder of $G_d(M_q)$ goes through a SCC partitioning. Any resulting component of size larger than the $split_size$ goes through the same split process, any component of size less than min_module_size becomes a leaf module, and any other component in between these two sizes becomes a seed module. In the extend phase, each leaf module is merged with the seed module with which it has maximum number of connections in $G_d(M_q)$.

We note that the graph G_d and the original PPI network G may differ, even when we consider the undirected version of G_d for comparison. Since the decisions made by the module construction step described in this subsection are based on the connections in G_d, it should be noted that the extracted modules may not exactly satisfy Constraint 1, that is the connectivity constraint of our main problem. However this should be considered a benefit rather than a drawback. It is well-known that PPI networks are incomplete and contain many false positives and negatives. Network prediction/smoothing methods based solely on network topology have been suggested previously. Many such methods are unsurprisingly based on random walks and have been shown to perform quite well on PPI networks including those of the H.Sapiens (Lei and Ruan, 2013; Alkan and Erten, 2017). Thus the directed graph G_d, especially after dilutions via minimum weight edge removals, not only embeds mutual exclusion and coverage information as relevant to cancer module identification, but also is indirectly a smoothed version of the original PPI network G.

3 Evaluation Results

We implemented the MEXCOWalk algorithm in Python. The source code, useful scripts for evaluations, and all the input data are freely available as part of the supplementary material. We compare MEXCOWalk results against those of three other existing knowledge-based cancer driver module identification methods: Hotnet2 (Leiserson *et al.*, 2014), MEMCover (Kim *et al.*, 2015), and Hierarchical Hotnet (Reyna *et al.*, 2018). The first two benchmark algorithms are chosen as representatives of their respective subcategories; Hotnet2 is one of the most popular benchmark methods based on optimizing coverage via a heat-diffusion heuristic and MEMCover is a popular algorithm among those optimizing mutual exclusion as well as coverage via a greedy seed-and-growth heuristic. Hierarchical Hotnet is chosen as a third benchmark method, as it is one of the most recent cancer driver module identification methods.

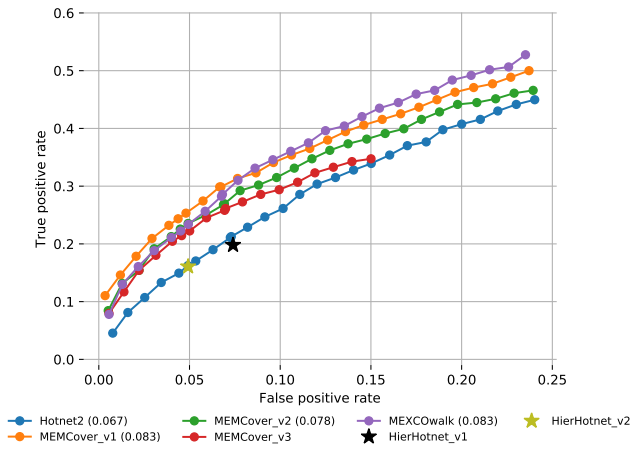

Fig. 1. The fraction of recovered CGC genes for each total_gene value is shown with a ROC plot. AUROC values are written in parentheses for each applicable method.

3.1 Input Data

All four methods, including MEXCOWalk, assume same type of input data in the form of mutation data of available samples and a H.Sapiens PPI network. We employ somatic aberration data from TCGA, preprocessed and provided by (Leiserson *et al.*, 2014). The preprocessing step includes the removal of hypermutated samples and genes with low expression in all tumor types. After the filtering, the dataset contains somatic aberrations for 11,565 genes in 3,110 samples. The mutation frequency of a gene g_i is calculated as the number of samples with at least one single nucleotide variation (SNV) or copy number alteration (CNA) in g_i divided by the number of all samples. As for the H.Sapiens PPI network, we used the HINT+HI2012 network (Leiserson *et al.*, 2014). This is a combination of two interactome databases: HI-2012 prepublication data in human HI2 Interactome database2 (HI2012) (Yu *et al.*, 2011) and high quality interactomes database (HINT) (Das and Yu, 2012). We execute each of the four algorithms on the largest connected component of this combined network that consists of 40,704 interactions among 9,858 proteins.

3.2 Parameter Settings

Regarding MEXCOWalk, we have settings for three parameters: the mutual exclusion threshold θ, the $total_genes$, and the min_module_size. We examine the distribution of MEX_n scores of all edges to determine a meaningful range of values for the mutual exclusion score threshold θ and conclude that only settings of $\theta \geq 0.5$ are in that range; details can be found in the Supplementary Document. In the main document, we present results for $\theta = 0.7$, as this setting provides the best results in terms of our main optimization goal defined in Equation 1. The results with other threshold values of $0.5, 0.6, 0.8, 0.9$ are available in the Supplementary Document. Regarding the $total_genes$ parameter, within the scope of the evaluations discussed in the main document,

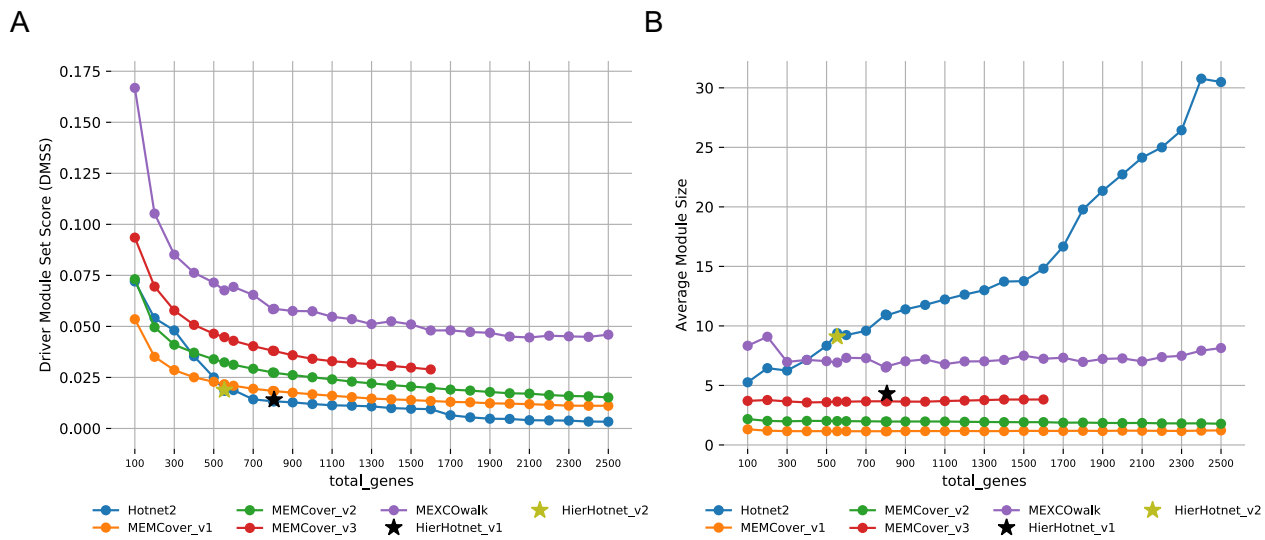

Fig. 2. A) DMSS evaluations of output modules of MEXCOwalk, MEMCover, Hotnet2, and Hierarchical Hotnet for increasing values of $total_genes$. B) Average modules sizes in the outputs of MEXCOwalk, MEMCover, Hotnet2, and Hierarchical Hotnet for increasing values of $total_genes$.

we view it as the main independent variable; we obtain the results of each evaluation under the settings $total_genes = 100, 200, \ldots, 2500$. Finally, we set min_module_size to 3 for the results discussed in the main document, as this constitutes a nontrivial mimimum module size compatible with the problem definition. Further results applying to the settings of $min_module_size = 6, 9, 12$ are presented in the Supplementary Document.

For Hotnet2, we obtain results for varying values of $total_genes = 100, 200, \ldots, 2500$, with the default value of $min_module_size = 3$. We present results of Hierarchical Hotnet with the default setting of its *clustering parameter* δ, which outputs modules of size greater than one, with a total of 806 genes. Since some of these modules may contain modules with two genes, we generate a filtered version as well, where all such modules are removed, resulting in modules with a total of 554 genes. In what follows, we refer to the former version as *HierHotnet_v1* and the latter version as *HierHotnet_v2*.

For MEMCover, mutual exclusion scores are obtained from type-restricted permutation test with all pan-cancer samples, that is the TR_test. Because confidence scores are not available for HINT + HI2012 network, we set the confidence score of all edges to 1 when calculating the edge weights for the MEMCover model. We set the coverage parameter k to its default value of 15. MEMCover introduces a parameter, $f(\theta)$, that is used to control the trade-off between the output number of modules and the average weights within each module. It indirectly controls the module sizes; the smaller $f(\theta)$, the larger the modules output by MEMCover in general. We consider three settings for the MEMCover algorithm, referred to as *MEMCover_v1*, *MEMCover_v2*, and *MEMCover_v3*, respectively. For the first one, we assign $f(\theta) = 0.584$, which is achieved by setting θ parameter (not to be confused with the θ we employ in MEXCOWalk) to 40%, as recommended in the original paper. For the second one, we assign $f(\theta) = 0.03$, which is the setting that minimizes the percentage of size one and size two modules. Finally, the last one corresponds to the setting where $f(\theta) = 0.03$ and all modules of size < 3 are removed. To obtain results with varying $total_genes$ from 100 to 2500 we consider the modules formed by the first $total_genes$ many genes output by each version, since the order MEMCover outputs the modules reflects the algorithm's quality preferences. Values of $total_genes$ larger than 1600 are not available for *MEMCover_v3* as it outputs 1684 genes in total.

3.3 Static Evaluations

Most of the existing driver module identification methods employ *static evaluations*, where the union of the genes in all the modules are compared against a reference set of cancer genes. Such evaluations measure the capability of an algorithm in dissecting cancer-related genes throughout the modules it provides, without regard for the specific modularity of the output. COSMIC Cancer Gene Census (CGC) database (Forbes *et al.*, 2017) is one such reference containing 616 genes with mutations that have been causally implicated in cancer. For consistency with previous work, our first evaluation compares the algorithms based on their ability to recover these known cancer genes. Figure 1 plots the Receiver Operating Characteristic (ROC) curves of the set of genes in the union of modules each algorithm provides with respect to the CGC genes. *MEMCover_v1* and our model has the same Area Under the ROC (AUROC) value of 0.083. *MEMCover_v2* ranks the third. Although the areas are not comparable, *MEMCover_v3* outputs provide worse TP rates than those of *MEMCover_v2*. Finally, Hotnet2 and Hierarchical Hotnet recover fewer known cancer genes than the rest of the methods.

3.4 Modular Evaluations

Most cancer driver module identification studies evaluate their output modules via analysis methods including functional enrichment, survival analysis, or literature verification. However in general these evaluations are not systematic due to lack of evaluation metrics that quantify these types of analysis and as such are not amenable for comparisons among alternatives. Without such metrics it is not possible to provide a fair judgment of different methods as far as the quality of the modules is of concern, and not just the quality of the static gene set. We provide three modularity-based metrics and evaluate the output module sets of alternative methods based on these metrics.

3.4.1 Driver Module Set Score

Our first evaluation metric is the main optimization goal of the cancer driver module identification problem, that is the driver module set scores ($DMSS$) defined in Equation 1. Fig. 2-A shows that MEXCoWalk discovers modules that have better $DMSS$ values than the module sets of all the other methods. The difference is much more dramatic for

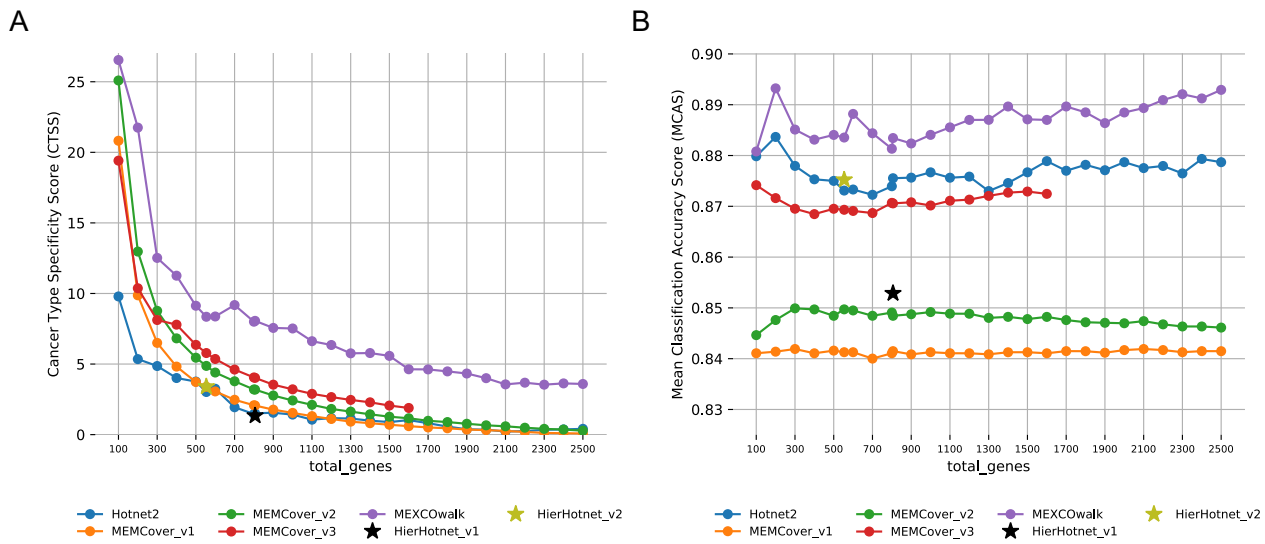

Fig. 3. A) CTSS evaluations of output modules of MEXCOwalk, MEMCover, Hotnet2, and Hierarchical Hotnet for increasing values of $total_genes$. B) MCAS evaluations of output modules of MEXCOwalk, MEMCover, Hotnet2, and Hierarchical Hotnet for increasing values of $total_genes$.

smaller $total_genes$ values such as 100 and 200. Those of Hierarchical Hotnet and Hotnet2 are among the worst, especially for settings of $total_genes > 500$. *MEMCover_v1* performs worse than the two other MEMCover versions, as it provides many size 1 and size 2 modules. This finding demonstrates another merit of the $DMSS$ definition; if there are many small modules, assuming the mutual exclusion does not decrease substantially by enlarging the modules, then our optimization score function prefers outputs with larger modules. Consider for instance, the following special case where we have 10 genes under consideration, each covering x out of a total of y samples. The output consisting of a set of modules each containing one gene has a $DMSS$ of x/y. On the other hand, assuming a MEX score of m for every pair of genes, the output with any pair of genes per module has a $DMSS$ of $2m^2x/y$. This implies that the latter is a more preferable module set than the former, as long as $m > \sqrt{1/2}$. It corresponds to the case where upto almost 58% of samples covered by a gene to be in the intersection of samples covered by another gene.

Analyzing the outputs of the methods with respect to the average module sizes in Fig. 2-B, we can observe that all MEMCover versions, especially *v1* and *v2*, provide module sets with very small average module sizes; for almost all values of $total_genes$, MEMCover_v1 provides an average size of almost 1.2 genes per module, MEMCover_v2 provides an average of almost 1.9 genes per module, and MEMCover_v3 provides an average of almost 3.7 per module. MEXCOWalk provides module sets with average sizes ranging between 6.5 and 9, whereas Hotnet2 module sizes significantly increase proportional to $total_genes$. The range of average module size values for Hotnet2 is approximately between 5 and 30.

3.4.2 Cancer Type Specificity Score

Our second modularity-based evaluation metric is defined with respect to cancer type specificity. We test an output module set in terms of enrichment for mutations in a specific cancer type using Fisher's exact test. For a module M, let S_M denote the set of patients where at least one of the genes in M is mutated. For a cancer type t, let ${S_M}^t$ denote the subset of patients in S_M diagnosed with cancer type t. Assuming n_t denotes the number of patients of cancer type t in the whole dataset, we calculate the Fisher's exact test with the following entries in the contingency table in row-major order:$|{S_M}^t|$, $n_t - |{S_M}^t|$, $\sum_{t' \neq t} |{S_M}^{t'}|$,$\sum_{t' \neq t} n_{t'} - |{S_M}^{t'}|$. We use the Bonferroni method for multiple testing correction.

Let $P = \{M_1, M_2, \ldots M_r\}$ be a set of modules. For each module $M_q \in P$, the described process results in a p-value for every cancer type t, denoted with ${p_q}^t$. We define the *cancer type specificity score* of P as the average $-\log$ of best p-value per module. More formally,

$$CTSS(P) = \frac{\sum_{M_q \in P} -\log(\min_{\forall t}({p_q}^t))}{r} \quad (2)$$

Fig. 3-A shows the CTSS scores of the module sets provided by the methods under consideration. Compared to the other methods, MEXCOWalk provides a larger CTSS value for every setting of $total_genes$, indicating that the output modules are strongly enriched for particular cancer types. We also observe that module sets of MEMCover versions perform better than those of Hotnet2 and Hierarchical Hotnet.

3.4.3 Mean Classification Accuracy Score

We examine the predictive value of an output set of modules in classifying tumor and normal samples of TCGA pan-cancer data consisting of 12 cancer types. For this, we employ k-nearest-neighbor classifier using Euclidean distance with $k = 1$, where the features are the expression values of the set of genes in a module. To evaluate the predictive performance of a module M_q, we use 10-fold stratified cross-validation accuracy, denoting it with $Acc(M_q, c)$ for a fold c. We can then define the Mean Classification Accuracy Score of a set of modules P as,

$$MCAS(P) = \frac{\sum_{M_q \in P} \sum_{c=1}^{10} Acc(M_q, c)}{10 \times r} \quad (3)$$

The plots of the MCAS scores of the module sets of all four methods for varying $total_genes$ are provided in t Fig. 3-B. MEXCOwalk consistently achieves the top accuracy for all settings of $total_genes$, implying that MEXCOwalk modules can more accurately perform tumor/normal classification than the other methods. Interestingly, Hierarchical Hotnet performs worse than Hotnet2. Among MEMCover models, MEMCover_v3 shows a better performance than MEMCover_v1 and MEMCover_v2, in contrast to their relative performances in recovering known cancer genes.

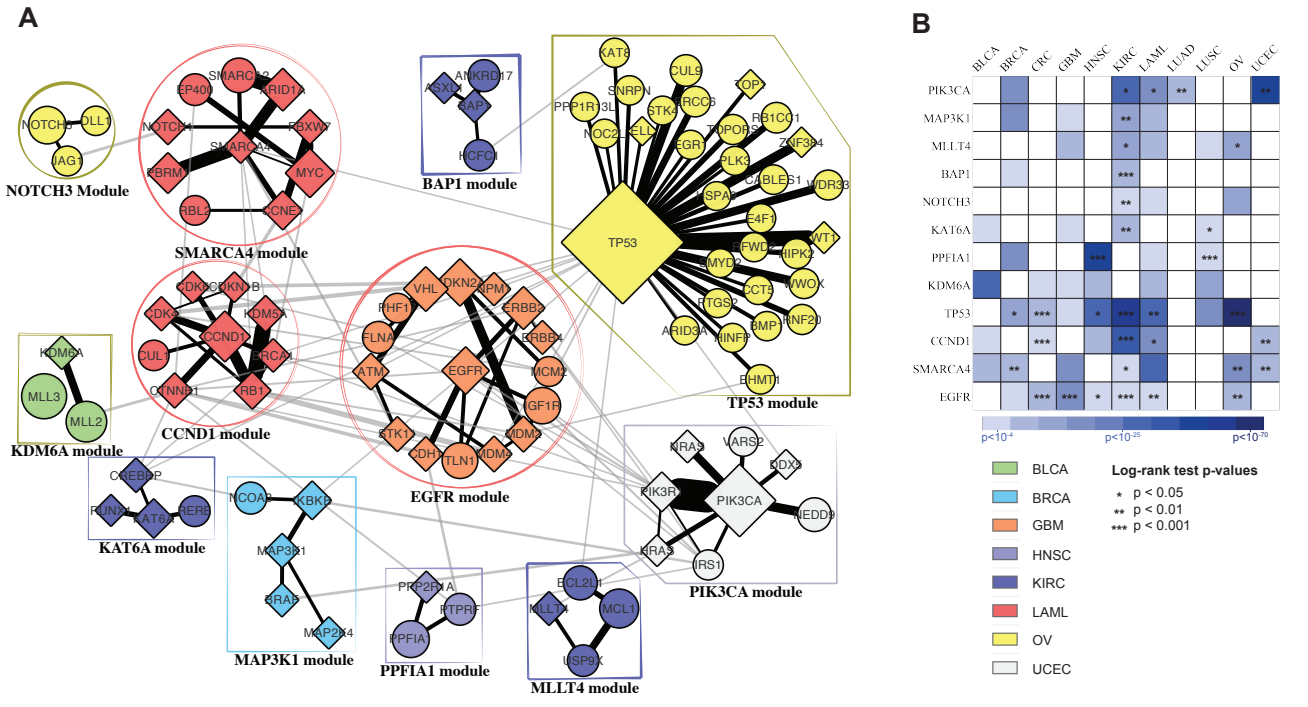

Fig. 4. A) MEXCOWalk output modules when $total_genes = 100$. Diamond shaped nodes correspond to CGC genes. Sizes of the nodes are proportional with mutation frequencies of corresponding genes. Edges within the module are colored black, whereas the edges between the modules are colored. Edge weights are reflected in the thicknesses of the line segments. Color of a module denotes the cancer type with the strongest enrichment for mutations in genes of that module. The legend for the color codes are shown on the right. Each module is named with the largest degree gene in the module. B) Results of cancer type specificity and survival analyses. Rows correspond to modules and columns correspond to cancer types. Colors of the matrix entries indicate the significance of enrichment for cancer types in terms of Fisher's exact test p-values. Stars indicate the significance of log-rank test p-values in survival analyses.

3.5 Analysis of MEXCOWalk Modules

Fig. 4-A shows the 12 modules that MEXCOWalk identifies when $total_genes$ is set to 100. The sizes of the modules range between 3 and 31, and their coverage values range between 5% to 50%. Note that the edges correspond to the PPI network edges, whereas the weight of an edge is the smaller of the weights of the corresponding directed edges from G_d as computed through edge-weighted random walk and thus represents the degree of mutual exclusivity and coverage assigned by MEXCOWalk.

Many of these modules are part of well known cancer-related pathways such as those centered at EGFR, TP53, PIK3CA, and CCND1. Analyzing the interactions between the modules, EGFR module can be seen as an important hub module between many important modules such as the TP53 module, CCND1 module, and the PIK3CA module; without the EGFR module these three modules would almost be isolated in the induced subgraph. The EGFR module contains several known cancer genes many of which are related to cell cycle control: VHL, CDKN2A, NPM1, ERBB2, ERBB4, MDM2, MDM4, STK11, CDH1, ATM. Seven cancer types are enriched for mutations in this module with GBM being the most significant enrichment; Fisher's exact test p-value is $= 4.9e - 20$. Indeed, EGFR gene is mutated in more than half of all GBM patients and anti-EGFR agents are already used for GBM treatment (Taylor *et al.*, 2012). However, resistance to these agents is a major problem suggesting that treatment strategies might benefit from targeting multiple genes in this module. This module also contains TLN1, which is not one of the known cancer genes listed in CGC. However it is mutated in 104 patients across 10 cancer types and it has previously been associated with tumorigenecity and chemosensitivity (Singel *et al.*, 2013; Fang *et al.*, 2016). We investigate whether the genes in this module are predictive of patient survival profiles by calculating a risk score for each patient as in Beer *et al.* (2002) and Shrestha *et al.* (2017). When we divide the GBM patients into two as training and test sets, the low-risk and high-risk thresholds that we identify from the training set are successful in stratifying the patients into low-risk and high-risk groups in the test set; the log-rank test p-value$= 0.0004$ (Fig. 4-B).

Our TP53 module includes 30 interactors out of 213 available in the HINT+HI2012 PPI network. TP53 shares the highest edge weight with WT1, which is a transcription factor that has roles in cellular development and cell survival. Another gene which has a large edge weight is CUL9. It is altered in 48 patients out of a total of 3110 samples, which would possibly make it easy to miss through single-gene tests.

The PIK3CA module identifies several genes in the PI3K pathway whose deregulation is critical in cancer development and progression (Karakas *et al.*, 2006). The module provides a chance to observe the importance of incorporating mutual exclusion in MEXCOWalk. Among all the interactions presented in the induced subgraph of 100 genes in all 12 modules, the one between PIK3CA and PIK3R has the largest weight. These genes are mutated in 602 and 155 patients respectively, although the overlap between the two patient sets is only 18 indicating the high mutual exclusivity between the pair of genes.

The CCND1 module is yet another fairly known cancer driver module (Kim and Diehl, 2010; Malumbres and Barbacid, 2009). Other than EGFR, it is the module that contains the most reference genes; all 9 genes in the module, except CUL1, are listed in the reference CGC database. It is shown that the mutations, amplification, and expression changes of these genes, which alter cell cycle progression, are frequently observed in a variety of tumors and contribute to tumorigenesis (Malumbres and Barbacid, 2009; Kim and Diehl, 2010). Indeed, we find significant

association of this module with patients' survival outcome in CRC, KIRC, LAML and UCEC types (Fig. 4-B).

One of the less known modules, the KDM6A module illustrates the advantage of using random walk to compensate for the incomplete edges in the PPI network data. MLL3 gene does not have interactions with MLL2 and KDM6A in the original PPI network. Nevertheless, the heat diffussion process of MEXCOWalk diffuses a large enough heat due mainly to the network topology, the coverage, and the mutual exclusion values of the pairs. As a result the three genes are connected with large edge weights in G_d, demonstrating the quality of MEXCOWalk modules even with missing interaction data, as verified by the the predicted interaction between MLL3 and KDM6A (Lee *et al.*, 2007).

4 Conclusion

In this study, we introduce a novel method, MEXCOWalk, that incorporates network connectivity, mutual exclusion, and coverage information to identify cancer driver modules. The optimization function employed by MEXCOWalk combines the mutual exclusion and coverage scores of modules after normalizing with suitable functions of module size. MEXCOWalk employs a vertex-weighted, edge-weighted random walk strategy where the edge weights reflect a novel combination of mutual exclusion and coverage. MEXCOwalk is able to output a set of modules with a predefined size, that is $total_genes$. This flexibility avoids *ad hoc* selection of an edge weight threshold and when applied to the other existing methods, it enables a robust comparison across different number of output genes. Another main contribution is to be able to split large modules in a systematic way, which becomes critical for large $total_genes$ values. Finally, apart from methodological contributions, the proposed modularity-based metrics fill an important gap in the literature.

Our results indicate that MEXCOWalk outperforms several state-of-the-art computational methods on TCGA pan-cancer data in terms of recovering known cancer genes, providing modules that are capable of classifying normal and tumor samples, and that are enriched for mutations in specific cancer types. Additionally, the risk scores determined with output modules can stratify patients into low-risk and high-risk groups in multiple cancer types. We also show that MEXCOwalk is robust against different settings of its parameters. In summary, MEXCOwalk identifies a number of known cancer modules as well as several putative ones. Further work on these modules may provide new insights into cancer biology.

In this article, MEXCOwalk is applied to TCGA pan-cancer data and HINT + HI2012 network. In the future, we plan to apply it to individual cancer types and network data from different PPI databases. Also, additional types of genetic and epigenetic aberrations can be included as they become available. Finally, adaptations of MEXCOwalk to include network density-related scores in edge weights in addition to mutual exclusion and coverage also constitute planned extensions of this work.

Acknowledgements

We thank Aissa Houdjedj for his help with the preparation of the figures.

Author contributions

Authors are listed in alphabetical order. C.E and H.K conceived the study and designed the overall methodology, and the strategy of analysis. E.H. developed the first implementation of MEXCoWalk. R.A. finalized the MEXCoWalk algorithm implementation and performed the experiments. I.B. performed the classification and survival analysis. All authors contributed to the preparation of the manuscript.

Funding

This work has been supported by The Scientific and Technological Research Council of Turkey [117E879 to H.K. and C.E.]

References

Alkan, F. and Erten, C. (2017). Rednemo: Topology-based ppi network reconstruction via repeated diffusion with neighborhood modifications. *Bioinformatics (Oxford, England)*, **33**(4), 537–544.

Babur, Ö., Gönen, M., Aksoy, B. A., Schultz, N., Ciriello, G., Sander, C., and Demir, E. (2015). Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biology*, **16**(1), 45.

Beer, D., Kardia, S., Huang, C., Giordano, T., Levin, A., Misek, D., Lin, L., G., C., Gharib, T., and Thomas, D. e. a. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, **8(8)**, 816–824.

Bersanelli, M., Mosca, E., Remondini, D., Castellani, G., and Milanesi, L. (2016). Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. *Scientific Reports*, **6**(1), 34841.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*, **22**(2), 398–406.

Dao, P., Kim, Y.-A., Wojtowicz, D., Madan, S., Sharan, R., and Przytycka, T. M. (2017). Bewith: A between-within method to discover relationships between cancer modules via integrated analysis of mutual exclusivity, co-occurrence and functional interactions. *PLoS Computational Biology*, **13**(10).

Das, J. and Yu, H. (2012). Hint: High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology*, **6**, 92.

Deng, Y., Luo, S., Deng, C., Luo, T., Yin, W., Zhang, H., Zhang, Y., Zhang, X., Lan, Y., Ping, Y., Xiao, Y., and Li, X. (2017). Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. *Briefings in Bioinformatics*. Exported from https://app.dimensions.ai on 2019/01/21.

Dimitrakopoulos, C. M. and Beerenwinkel, N. (2017). Computational approaches for the identification of cancer genes and pathways. Published online 11 November 2016.

Dopazo, J. and Erten, C. (2017). Graph-theoretical comparison of normal and tumor networks in identifying brca genes. *BMC Systems Biology*, **11**(1), 110.

Erten, S., Bebek, G., and Koyuturk, M. (2011). Vavien: An algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. *J Comput Biol*, **18**, 1561–74.

Fang, K., Dai, W., Ren, Y., Xu, Y., Zhang, S., and Qian, Y. (2016). Both talin-1 and talin-2 correlate with malignancy potential of the human hepatocellular carcinoma mhcc-97 l cell. *BMC Cancer*, **16(45)**, 2076–2079.

Forbes, S., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C., Ward, S., Dawson, E., and Ponting, L. e. a. (2017). Cosmic: somatic cancer genetics at high-resolution. *Nucleic Acids Res*, **45**, D777–D783.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA.

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature Methods*, **10(11)**, 1108–1115.

Karakas, B., Bachman, K., and Park, B. (2006). Mutation of the pik3ca oncogene in human cancers. *British Journal of Cancer*, **94**, 455–459.

Kim, J. K. and Diehl, J. A. (2010). Nuclear cyclin d1: An oncogenic driver in human cancer. *J Cell Physiol*, **220(2)**, 292–296.

Kim, Y.-A., Cho, D.-Y., and Przytycka, T. M. (2015). MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics*, **31**(12), i284–i292.

Lawrence, M., Stojanov, P., Polak, P., V Kryukov, G., Cibulskis, K., Sivachenko, A., L Carter, S., Stewart, C., H Mermel, C., Roberts, S., Kiezun, A., S Hammerman, P., McKenna, A., Drier, Y., Zou, L., H Ramos, A., J Pugh, T., Stransky, N., Helman, E., and Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**.

Lee, M., Villa, R., Trojer, P., Norman, J., Yan, K., Reinberg, D., Di Croce, L., and Shiekhattar, R. (2007). Demethylation of h3k27 regulates polycomb recruitment and h2a ubiquitination. *Science*, **318(5849)**, 447–50.

Lei, C. and Ruan, J. (2013). A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity. *Bioinformatics*, **29**(3), 355–364.

Leiserson, M. D. M., Blokh, D., Sharan, R., and Raphael, B. J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLOS Comput Biol*, **9**(5), 1–15.

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., Lawrence, M. S., Gonzalez-Perez, A., Tamborero, D., Cheng, Y., Ryslik, G. A., Lopez-Bigas, N., Getz,

G., Ding, L., and Raphael, B. J. (2014). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, **47**(2), 106–114.

Liu, B., Wu, C., Shen, X., and Pan, W. (2017). A novel and efficient algorithm for de novo discovery of mutated driver pathways in cancer. *Ann. Appl. Stat.*, **11**(3), 1481–1512.

Malumbres, M. and Barbacid, M. (2009). Cell cycle, CDKs and cancer: a changing paradigm. *Nat Rev Cancer*, **9**, 153–166.

Miller, C. A., Settle, S. H., Sulman, E. P., Aldape, K. D., and Milosavljevic, A. (2011). Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Medical Genomics*, **4**(1), 34.

Reyna, M., Leiserson, M., and Raphael, B. (2018). Hierarchical HotNet: identifying hierarchies of altered subnetworks. . *Bioinformatics*, **34**(17), i972–980.

Shrestha, R., Hodzic, E., Sauwewald, T., Dao, P., Wang, K., Yeung, J., Anderson, S., Vandin, F., Haffari, G., Collins, C., and Sahinalp, C. (2017). Hit'ndrive: patient-specific multidriver gene prioritization for precision oncology. *Genome Res*, **27(9)**, 1573–1588.

Singel, S., Cornelius, C., Batten, K., Fasciani, G., Wright, W., Lum, L., and Shay, J. (2013). A targeted rnai screen of the breast cancer genome identifies kif14 and tln1 as genes that modulate docetaxel chemosensitivity in triple-negative breast cancer. *Clin Cancer Res*, **19(8)**, 2061–2070.

Taylor, T. E., Furnari, F., and Cavenee, W. (2012). Targeting egfr for treatment of glioblastoma: Molecular basis to overcome resistance. *Curr Cancer Drug Targets*, **12(3)**, 97–209.

Vandin, F., Upfal, E., and Raphael, B. (2011a). Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol*, **18**, 507–22.

Vandin, F., Upfal, E., and Raphael, B. J. (2011b). *De Novo* discovery of mutated driver pathways in cancer. In *Research in Computational Molecular Biology - 15th Annual International Conference, RECOMB 2011, Vancouver, BC, Canada, March 28-31, 2011. Proceedings*, pages 499–500.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLOS Comput Biol*, **6**(1), 1–9.

Weinstein, J., Collisson, E., Mills, G., Shaw, K., Ozenberger, B., Ellrott, K., Sander, C., and et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, **45**(10), 1113–1120.

Yang, C., Ge, S.-G., and Zheng, C.-H. (2017a). ndmaSNF: cancer subtype discovery based on integrative framework assisted by network diffusion model. *Oncotarget*, **8(51)**, 89021–89032.

Yang, H., Wei, Q., Zhong, X., Yang, H., and Li, B. (2017b). Cancer driver gene discovery through an integrative genomics approach in a non-parametric bayesian framework. *Bioinformatics*, **33**(4), 483–490.

Yeang, C.-H., McCormick, F., and Levine, A. (2008). Combinatorial patterns of somatic gene mutations in cancer. *The FASEB Journal*, **22**(8), 2605–2622.

Yu, H., Tardivo, L., Tam, S., Weiner, E., and Gebreab, F. e. a. (2011). Next-generation sequencing to generate interactome datasets. *Nature Methods*, **8**, 478–480.

Zhang, J. and Zhang, S. (2018). The discovery of mutated driver pathways in cancer: Models and algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **15**(3), 988–998.