Driver Subnetwork Identification by Community Detection and Network Dismantling Methods

Fatma Ulkem Bas, Sertan Ali Dikli, Nurcan Tunçbağ

During the tumor evolution, cancer driver mutations accumulate on some specific genes (oncogenes, tumor suppressors) which promote cell proliferation. Driver mutations generally show themselves as a deleterious mutation in the tumor-suppressor genes, such as p53. Similarly, proto-oncogenes can be mutated and turn into oncogenes, which, as the name implies, means the cancer-inducing genes being expressed higher than their trending levels. These biomarkers altogether are called cancer driver genes in the extent of this study.

These specific genes and the driver network that they affect need to be understood and studied to help cancer research. Network models provide an insight into the mathematical projection of biological processes. Hypothetically, this might lead to an unbiased prediction of the affected pathways, response to cancer drugs, cellular state upon different mutations throughout the lifetime of the cell, gene expression.

Cancer is a family of diseases that can lead to a typical phenotype that is uncontrollable cell proliferation, which shows a considerable variation among tissues. The network structures and the driver genes may differ across tissues. Therefore, tissue-specific modeling is crucial in cancer.

This study uses network analysis and perturbations to predict driver genes and their associated subnetworks. For this purpose, we consider topological aspects such as node removal costs, centrality measures, or their localization tendencies in the network (e.g., the most populated region of the network by the drivers). Here, we constructed a pipeline to reveal driver subnetworks. Starting on a given tissue-specific protein-protein interaction (TSPPI) network, we trim the network to an optimal point without losing the driver genes. We performed a blind method in which the driver genes are not given initially to the algorithm until the testing period. We aimed to prevent a bias towards the most studied driver genes. We may summarize the pipeline as a method that uses a graph-based approach to predict genes that drive tumor progression.

In this study, we performed first the Leiden community detection algorithm, followed by Generalized Network Dismantling with Reinsertion, and finally Personalized PageRank (PPR). We obtained the TSPPI networks for different tissues such as breast, colon, kidney, lung, ovary, and skin from the TissueNet V2 database and the driver gene information related to the corresponding tissues from the Cancer Genome Interpreter (CGI) Biomarkers database for testing.

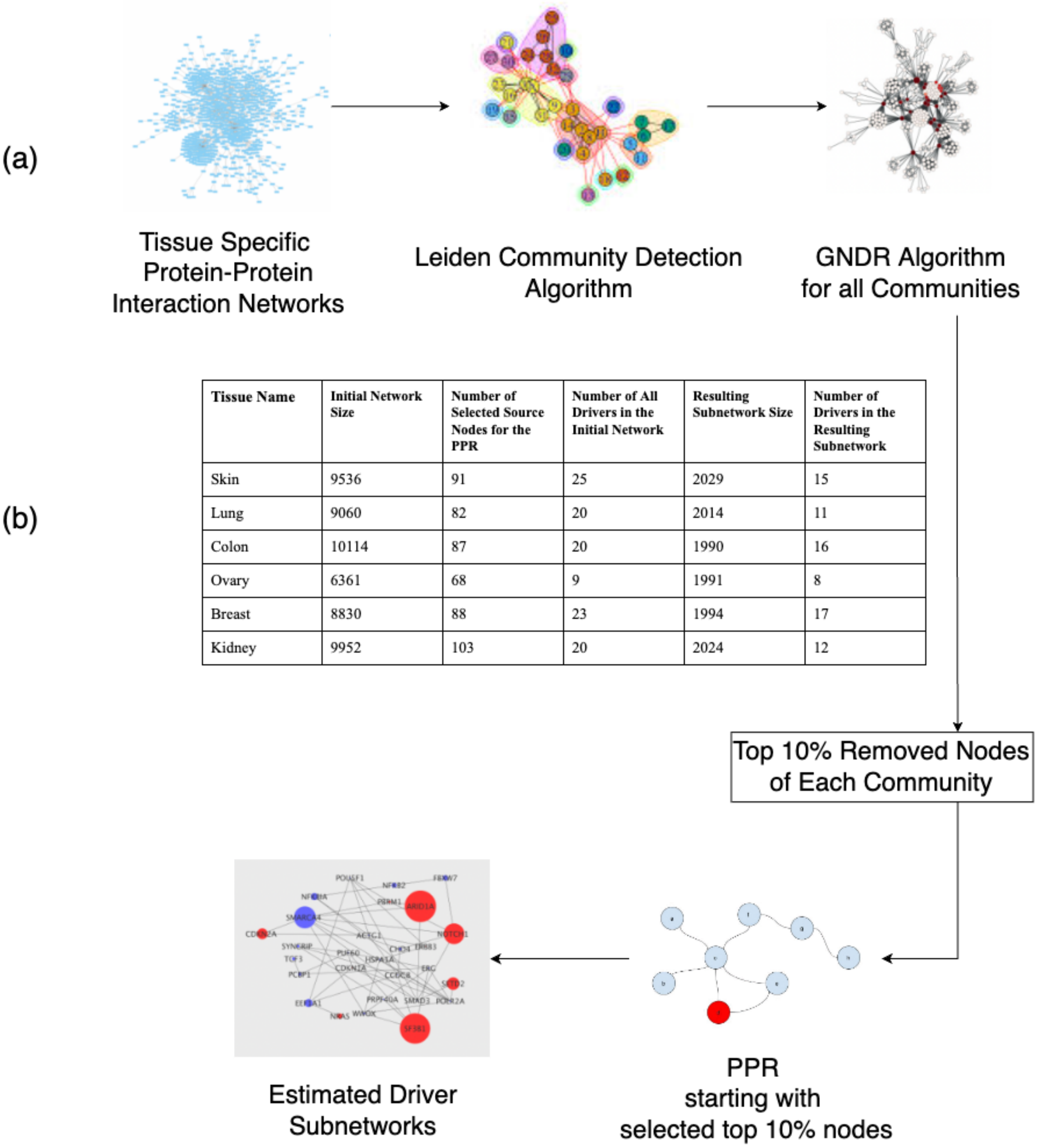
Leiden algorithm is a community detection method for the networks depending on the topological connections within each communities' members; in other words, by using this method, we created subgroups of nodes that reveals the modules The tissue-specific communities were pre-processed to continue with the Generalized Network Dismantling with Reinsertion (GNDR). In summary, these two steps provide a node elimination strategy depending on the topological costs of removed nodes and return a list of removed nodes from the reference interactome resulting in the size reduction of the largest connected component.

We selected the top 10% of removed node lists of each community from each TSPPI as the source nodes for the Personalized PageRank (PPR) algorithm. Then, we conducted the PPR algorithm on the original TSPPI network. Finally, we removed around 70% of the nodes losing less than 20% of the driver genes. Throughout the processes, the driver gene information was hidden from the pipeline and only used for the testing period.

Utilizing the algorithms mentioned above with optimized parameters, we were able to reduce the initial size to 25% of each initial network, and the known driver genes were searched among these sets of genes and confirmed that more than half of the driver genes were still present in the remaining network as shown in Table 1. This algorithm found 74, 80 and 89 percent of the known driver genes for ovary, colon and breast cancer tissues, while reducing the network sizes to 31, 19.7 and 22.5 percent respectively. However; for the kidney, skin and lung tissues the driver counts were 60, 60 and 55 while the corresponding network sizes are 20, 21 and 22. In this view point, we can say the algorithm has been more succesful for the former 3 tissues, while the latter 3 tissues has a similar but less success. More detailed numbers can be found in the Figue 1.b as a table.

We conducted pathway enrichment analysis and the results show that our subgraph show an enrichment in breast cancer tissue specific pathways such as Estrogen-dependent gene expression, and breast cancer pathways with the adjusted p-values of $4.3x10^{-16}$ and $1.07x10^{-7}$. As well as many other cancer related pathways such as Mitotic G1 phase and G1/S transition with an adjusted p-value of $1.38x10^{-18}$.

To conclude, these procedures can be used to reduce the computational burden of the following studies and provide promising candidates for the unknown driver genes.


(a)
Tissue Specific Protein-Protein Interaction Networks
Leiden Community Detection Algorithm
GNDR Algorithm for all Communities
(b)
Top 10% Removed Nodes of Each Community
POU5F1
FBXW7
NFKB2
NFKBIA
PBRM1
ARID1A
SMARCA4
CDKN2A
ACTG1
NOTCH1
SYNCRIP
CHD4
ERBB3
PUF60
TCF3
HSPA1A
CDKN1A
ERG
PCBP1
CCDC8
SETD2
EEF1A1
PRPF40A
SMAD3
POLR2A
NRAS
WWOX
SF3B1
Estimated Driver Subnetworks
PPR starting with selected top 10% nodes

Tissue Name	Initial Network Size	Number of Selected Source Nodes for the PPR	Number of All Drivers in the Initial Network	Resulting Subnetwork Size	Number of Drivers in the Resulting Subnetwork
Skin	9536	91	25	2029	15
Lung	9060	82	20	2014	11
Colon	10114	87	20	1990	16
Ovary	6361	68	9	1991	8
Breast	8830	88	23	1994	17
Kidney	9952	103	20	2024	12

Figure 1. (a) Overview of the pipeline. **(b)** Tissue specific protein-protein interaction network sizes and driver gene counts obtained from TissueNet v2 and CGI Biomarkers Database, respectively and estimated driver subnetwork sizes resulting from the overall procedure.