Pacific Symposium on Biocomputing 26:261-272 (2021)

Frequent Subgraph Mining of Functional Interaction Patterns Across Multiple Cancers

Arda Durmaz [1,5], Tim A. D. Henderson [2], and Gurkan Bebek [1–4,*]

[1] Systems Biology and Bioinformatics Graduate Program,
[2] Computer and Data Sciences Department,
[3] Center for Proteomics and Bioinformatics,
[4] Nutrition Department,
Case Western Reserve University, 10900 Euclid Ave., Cleveland OH 44106, USA
[5] The Department of Translational Hematology and Oncology Research, Taussig Cancer Institute,
Cleveland Clinic, 9500 Euclid Ave., Cleveland, OH 44195, USA
[*] Correspondence should be addressed to gurkan.bebek@case.edu.
{axd497,tadh,gurkan.bebek}@case.edu

Abstract. Molecular mechanisms characterizing cancer development and progression are complex and process through thousands of interacting elements in the cell. Understanding the underlying structure of interactions requires the integration of cellular networks with extensive combinations of dysregulation patterns. Recent pan-cancer studies focused on identifying common dysregulation patterns in a confined set of pathways or targeting a manually curated set of genes. However, the complex nature of the disease presents a challenge for finding pathways that would constitute a basis for tumor progression and requires evaluation of subnetworks with functional interactions. Uncovering these relationships is critical for translational medicine and the identification of future therapeutics. We present a frequent subgraph mining algorithm to find functional dysregulation patterns across the cancer spectrum. We mined frequent subgraphs coupled with biased random walks utilizing genomic alterations, gene expression profiles, and protein-protein interaction networks. In this unsupervised approach, we have recovered expert-curated pathways previously reported for explaining the underlying biology of cancer progression in multiple cancer types. Furthermore, we have clustered the genes identified in the frequent subgraphs into highly connected networks using a greedy approach and evaluated biological significance through pathway enrichment analysis. Gene clusters further elaborated on the inherent heterogeneity of cancer samples by both suggesting specific mechanisms for cancer type and common dysregulation patterns across different cancer types. Survival analysis of sample level clusters also revealed significant differences among cancer types ($p < 0.001$). These results could extend the current understanding of disease etiology by identifying biologically relevant interactions.

Supplementary Information: Supplementary methods, figures, tables and code are available at https://github.com/bebeklab/FSM_Pancancer.

Keywords: Frequent Subgraph Mining, Pan-Cancer, Transcriptomics; Proteomics

1 Introduction

Cancer is an inherently complex and heterogeneous disease. New technologies provided a comprehensive list of genomic and epigenetic aberrations for tumor growth and proliferation (1–4). This knowledge base could provide a more comprehensive view of how signaling events alter homeostasis within cells, between cells, or the microenvironment. The multiple omics measurements collected could be integrated to identify mechanisms or specific functions relevant to cancer (5) where shared

genomic features across cancers have been identified (1, 6, 7), some of which were through integrative methods to analyze multiple -omics datasets (8–11), While these gene-centric approaches report valuable insights, the biology behind their prognostics or stratification might be more complicated, leading to poor treatment options or reproducibility. For example, gliomas with mutated *IDH1* and *IDH2* have improved prognosis compared to gliomas with wild-type *IDH* (12). As a result, mutant-selective *IDH1* inhibitors were developed, but this drug strategy could make tumor progression worse (13–16). Other arguments are made over the validity of geneset-based biomarkers (17–19). Random genesets were shown to stratify patients into subgroups, contradicting the use of these geneset based methods (20, 21). Pathway-based approaches, on the other hand, could uncover functionally relevant mechanisms of oncogenic alterations to improve treatment options (4).

The availability of pan-cancer data allowed the simultaneous analysis of multiple cancer types. However, the multifaceted view of cancer hinders these efforts to uncover comprehensive maps of cancer for each cancer type. Sanchez-Vega et al. (4) were able to map 57% of tumors to at least one expert-curated signaling pathway targetable by currently available drugs. The ten expert-curated pathways in this study are a great resource but do not cover the alterations across all cancers. Leiserson et al. (22) focused on gene-level perturbations to find subnetworks common across cancer types but the identified subnetworks are not restricted to cover the same set of samples, which can mask subpopulations of samples with different genes mutated in the given subnetwork. An unsupervised approach that mines networks for a dynamic group of patients could bring a more comprehensive map and would provide improved insight into our understanding of tumor growth and treatment opportunities.

One of the commonly used methods in graph data mining is frequent subgraph mining (FSM). FSM provides a means to extract frequently occurring patterns in a graph database. For instance in the setting for protein-protein interactions (PPIs), one can define a graph for each cancer patient based on expressed proteins and mine for commonly occurring interactions across patients (23). FSM has been widely used in a variety of applications, including the identification of common metabolic pathways and clusters (24–26). Multiple algorithms have been developed to overcome challenges inherently present in subgraph mining regarding both memory and subgraph isomorphism issues (27–30). The general approach for mining frequently occurring patterns in a graph database is to grow candidate patterns either in depth-first search or breadth-first search manner and check whether the required support is achievable. One drawback of using FSM-based methods is the computational requirement since the subgraph isomorphism problem is NP-Complete (31).

One other methodology for utilizing global network topology is the random walks with restarts (RWR) on finite graphs. RWR algorithm is the simulation of a random walker jumping from node to node in the interaction network with the given parameters similar to the PageRank algorithm (32). A modified version of this approach is used to prioritize the local neighborhood by allowing the random walk to restart from specified seed nodes. This approach has been widely used for candidate gene prediction or disease-disease similarity measurements (33–35).

In this paper, we describe an integrative -omics approach to pan-cancer analysis using FSM coupled with biased random walks utilizing genomic alterations, gene expression, and PPI networks. We use FSM to identify frequently occurring interaction patterns to provide a better understanding of functional alterations across multiple cancer types while accounting for the complex interaction topology of cancer. Our goal is to integrate PPI networks with somatic alterations and gene expression profiles to infer molecular networks representing dysregulation in cancers. More specifically, we extract subnetworks that are frequent in the population and in close proximity to the mutated

genes. In our analysis, we investigate TCGA samples for 32 cancer types. We present patient clusters across all cancer types as well as patient classifications of individual cancers based on these networks. We identify mechanisms that are shared across tumor types and unique to individual cancers.

2 Methods

2.1 Pan-cancer Dataset and Omic Databases

We have downloaded TCGA single nucleotide variation (SNV) data from UCSC Xena (36). Additionally, we have filtered out samples with mutations of more than 800 to reduce the possible effects of hypermutators. PPI network was downloaded from StringDB version 10.5 and filtered to include edges with confidence scores > 0.4 with the remaining number of nodes being 17473 (37). Pathways were downloaded from the Reactome database. We excluded pathways with genes of less than 8 (38).

2.2 Biased Random Walks with Restarts

Biased random walks are applied to each sample separately by considering the mutated genes as seeds hence prioritizing local neighborhood of genomic alterations (See Supplementary Method Section S1.1). In this process, nodes with high degrees will intrinsically have increased probability values/traversed more often, to capture nodes with a statistically significant association with the seed set of nodes, we compared these results to a null distribution generated by applying the biased random walks to thousand randomly generated seed sets keeping the number of seeds equal to the original seed set. p-values for each node are obtained by comparing the steady-state probability vector to the null distribution per gene. Multiple hypothesis testing corrections are done using the Bonferroni method and genes with p-values < 0.1 are kept. Restart probability for the biased random walk is chosen as 0.6 to not restrict the networks towards the neighbors of seed sets. However, note that restart probability can be fine-tuned specifically to each network but 0.6 generally performs well across biological networks. Furthermore, since biased random walks can also identify spurious significant nodes solely due to the topology of the network, we have extracted connected components with the number of genes > 3.

2.3 Frequent Subgraph Mining

We developed an efficient method to sample for frequently occurring subgraphs across pan-cancer samples (Algorithm 1). The goal of frequent subgraph mining is to discover all subnetworks of graphs in the database which recur at least k times (39, 40). The database is a collection of undirected gene networks assembled as described in Section 2.1. The parameter k in Algorithm 1 is called the *minimum support*. A subgraph is considered "frequent" (and *supported*) if it recurs at least k times.

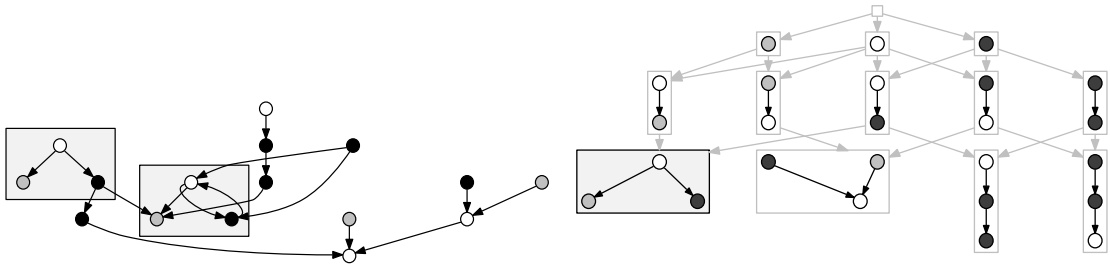
In this analysis pipeline, we have applied biased random walks over the PPI network for each sample separately using the somatic alterations as seed sets. Following the RWR, FSM can be applied with two approaches; Mining a single graph database generated by merging the RWR results over all the samples or mining graph databases generated separately for each sample. A Sample-specific network can be generated by filtering the combined network to include nodes found significant for the current sample. Simply this approach will result in subnetworks with the specified

```
Param: Mutation Matrix V of size g × p        // Each column is a set of mutations from one sample
Param: Interaction Network W of size g × g    // Interaction matrix stored as an adjacency matrix
Param: Minimum Support k                      // minimum network size to be discovered
Result: Set of k-frequent subgraphs S
// biased random walk profiles
for i ← 1 to p do
    P ← RWR(V[, i], W)                // perform RWR per sample using each sample's mutation set.
    CC ← ConnComp(P)                  // Find the connected component
    RES ← Append(CC)                  // Save this network for this patient
end
// sample-specific graph databases
for i ← 1 to p do
    PD ← []
    PD[i] ← RES[i]                    // For all RWR networks identified with mutations of the patient
    for j ← 1 to p do
        for e ∈ RES[j] do
            if e ∈ RES[i] then
                PD[j] ← Append(e)     // Merge to create a sample-specific network database
            end
        end
    end
    D ← Append(PD)
end
// frequent subgraph mining of sample-specific graph databases
for i ← 1 to p do
    PD ← D[i]                         // for each patient's network
    for h ∈ GetCand(PD)               // collect frequent subgraphs
    do
        if GetSupp(h) ≥ k             // if support for subgraphs is larger than k
        then
            S ← Append(h)             // include this subnetwork in the results
        end
    end
end
return S
```

Algorithm 1: High-level algorithm for the proposed framework. The input matrices V and W have sizes $g \times p$ and $g \times g$ respectively. g is the number of genes and p is the number of samples in the pan-cancer data. k is the minimum number of samples a frequent subnetwork recurs. The algorithm returns the set of k-frequent subgraphs.

support that are also present in the current sample. FSM results for sample-specific networks are then merged and duplicate networks are filtered. We have chosen to run FSM in sample-specific approach since applying FSM over an all-sample database (a single graph database including all the edges from all the samples) will lead to bias in the identified subgraphs due to the subset of the samples having a high number of dysregulated patterns.

Applying the algorithm above to our problem naively is not practical. It involves solving several difficult sub-problems, including candidate subgraph generation and subgraph isomorphism. Furthermore, many frequent subgraphs would overlap with each other (41) returning exponentially large similar subgraphs (42). Our FSM approach resolves these problems in two ways. First, it uses a highly optimized method for candidate generation which prunes unsupported supergraphs (39). Second, instead of collecting all frequent subgraphs, a sample of graphs is collected using the



(a) *A colored (labeled) digraph.* ***(b)*** *A 2-frequent connected subgraph lattice.*

Fig. 1: *Figure (b) is a connected subgraph lattice of the graph in Figure (a) including only the subgraphs with 2 or more embeddings in Figure (a). The boxed nodes in the graph show the embeddings of the boxed subgraph in the lattice. In the figure, the colors (black, gray, and white) are standing in for labels on the vertices (Adapted from (39)).*

GRAPLE algorithm (42). GRAPLE models the set of frequent subgraphs as a *lattice* where the graphs in the lattice are connected by their subgraph and supergraph relationships (see Figure 1). Frequent subgraphs are sampled from the lattice by taking random walks on the lattice. For full details see (39, 42), a related approach can be found in (43).

We have extensively tested the FSM algorithm to validate our approach and also compared it to previous methods (43–45). We tested parameter k on various benchmark datasets (Supplementary Table S3) and validated k's effect on run time and subnetwork discovery. We ran these simulations in a non-heuristic mode to recover all subnetworks. We comprehensively compared our performance with GRAMI (Supplementary Table S4). GRAMI finds a slightly different number of patterns because it uses undirected graph search (otherwise GRAMI's run time suffers). Our tool outperformed GRAMI.

2.4 Integrating Gene Expression Measurements to FSM Framework

We integrated the somatic mutations with gene expressions using the same -omics dataset and interaction network from (46). The integration of gene expression is done in two steps. First, in the biased random walk step, the transition probabilities are assigned based on the euclidean norm of z-scores of interacting genes. This scheme prioritized genes with high dysregulation compared to the population in addition to seed sets. Furthermore, for functional relevance, we have applied dimension reduction followed by clustering with PAM and pathway enrichment (PAM-Clusters, Figure S10) (47). To apply the dimension reduction, each identified subgraph is assigned an average dysregulation score (matrix of frequent subgraphs vs samples) (23).

2.5 Functional Analysis

To associate biological mechanisms with frequent subgraphs, we utilize clustering, non-linear dimension reduction, pathway enrichment, and survival analysis. Since FSM is done in a sample-specific manner, identified subgraphs contain redundant interactions (repeated interactions across multiple subnetworks). We apply greedy clustering to remove these redundant interactions by grouping highly connected nodes (48). In this process, we find high modularity partitions of our networks.

For survival analysis, we utilized unsupervised clustering using the frequent subgraphs as features. Frequent subgraphs mined using gene expression integration are assigned dysregulation scores

using the average euclidean norms of standardized gene expressions for a single fsg $\sqrt{\sum_i^n g_i^2}$ and samples are clustered using PAM on dimension reduced space (47, 49). For FSGs identified using only SNVs, we assigned the frequency of matching genes in the FSG and the sample as a score and employed PAM. However note that for clustering samples with matching gene frequencies as scores, we did not use non-linear dimension reduction.

3 Results

3.1 Pan-cancer Subgraphs

FSM has identified 43k unique subgraphs with sizes between 6-60 edges across the 90% of the pan-cancer dataset with support 20 corresponding to 0.3% of samples (Figure 2). Identified subgraphs covered more than 40% of the genes in the protein-protein interaction networks.

(a) Frequent subgraph with highest frequency *(b) Frequent subgraph with largest size*

Fig. 2: *Sample frequent subgraphs mined from the pan-cancer dataset. Each edge in the given subgraph is represented in at least 20 common set of samples.*

3.2 Pathway Enrichment

To elaborate on the functional relevance of the identified subgraphs, we have clustered the merged subgraph network using a greedy approach (FSG-Clusters) (48). More specifically, frequent subgraphs are merged into a single network, and clustering is applied. This method can be seen as filtering the initial protein-protein interaction network to include edges that show frequent interaction patterns. However, note that this scheme is not similar to simply filtering the edges that have minimum support level but in subgraph space. A total of 106 clusters was identified (See Suppl. tables). To filter clusters without functional relevance, we have removed clusters with node size smaller than 10 and larger than 400. Pathway enrichment analysis using the Reactome Pathways has identified a total of 620 significant subnetworks using a p-value threshold of 0.01, including previously identified mechanisms: PI3K Cascade, Cytokine Signaling, DNA Repair, Signaling by NOTCH (Figure 3).

3.3 Disease Enrichment

To evaluate the representation of cancer types in identified clusters we have done enrichment analysis for each patient as well (Figure S3). Multiple clusters showed few over-representation in terms

Fig. 3: *Top three enrichment results of identified clusters sorted by root pathways (FSG-Clusters).*

of predefined disease types. These clusters also showed few or no pathway enrichment which might suggest small subnetworks stratifying patients in combination with broad dysregulation patterns.

Samples with lower-grade glioma (LGG) are represented across different clusters similar to breast cancer samples. However, increased representation for LGG samples in clusters 85 and 63 is visible. Cluster 85 is mostly associated with CSF2RA-B metabolism, which are cytokines related to macrophage, granulocyte differentiation, and production. An earlier study showed how intercellular microglia polarization signaling through CSF2 (GM-CSF) and IFNG are the molecules that drive microglia towards the M1 phenotype (50). Cluster 63, on the other hand, is related mostly to NOTCH signaling, p75NTR degradation through NRIF interactions (Figure S11). In contrast, breast cancer patients show increased representation in clusters 23, 24, 35, and 44. Given clusters correspond to lipid metabolism (known risk factor for developing cancer (51)), membrane trafficking, cytoskeletal related processes, *SEMA3A*, *SEMA4D* signaling, which might related to increased Metastasis in breast cancer (52). Patients with skin disorders are mainly represented in clusters 47 and 102. Pathway enrichment for the clusters identifies degradation of the extracellular matrix, O-linked glycosylation, and collagen biosynthesis. On the other hand, uveal melanoma patients are enriched for cluster 89, which shows dysregulation in GPCR signaling, the main biological processes impacted by the recurrent mutations in uveal melonama (53). Thyroid cancer patients show the most specific enrichment for cluster 80, showing functional relevance in the regulation of RAS by GAPs, and MAPK pathways, key signaling pathways in both initiation and progression of medullary thyroid carcinoma (54). Prostate cancer patients are mainly enriched for clusters 35, 44, and 23, showing enrichment for Rho GTPase activation of PAK, cleavage of cell adhesion proteins through apoptosis, *SEMA3A*, and *SEMA4D* related signaling. Head and neck cancer patients also show dysregulation across a large number of clusters discovered but show the highest enrichment for cluster 106, similar to breast cancer and LGG patients.

3.4 Comparison of Pan-cancer FSM Networks

Oncogenic Signaling Pathways of Pan-cancer To further elaborate on the utility of the proposed method we have compared the genes in identified frequent subgraphs to previously established expert-curated pathways (4). We have recovered 65% of genes covering 90% of pathways reported in the curated list including *EGFR, TP53, PIK3CA, PTEN* matching various mechanisms. To further compare against previously curated pathways, we have utilized cancer hallmark genesets (55, 56). Frequent subgraphs cover 100% of the hallmark gene sets with at least 1 overlapping gene. Interestingly FSG clusters cover multiple pathways and pathways are covered by multiple FSG clusters as well both for oncogenic signaling pathways and hallmark gene sets. This further elaborates on the complexity of cancer and the interaction topology (Figure S1). We identified additional genes, novel to the curated pathway database as well suggesting the importance of system-level identification of functional mechanisms and the complexity of cancer progression (See Suppl. Figure S2).

HotNet2 Pan-cancer Subnetworks We also compared our method to HotNet2 (22), which aims to find subnetworks significantly enriched in given alterations across the pan-cancer dataset. However, the main difference is that HotNet2 focuses on gene-level perturbations and looks for subnetworks covering a wide range of samples in the dataset. More specifically in the subnetworks identified by HotNet2, different subsets of samples can show alterations in different nodes of the subnetwork. In contrast, our methodology aims to identify subnetworks for all samples meaning that in the identified subnetworks a common set of samples show dysregulation for all the nodes in the subnetwork. When we compare to HotNet2 subnetworks, we observe that clusters 63, 70, 80, and 90 correspond to 5 subnetworks out of 15 relating to $BRAF$, RAS, $PIK3CA$ subnetwork, $KDM6A$, $MLL2$, $MLL3$ subnetwork, SWI/SNF complex, $BAP1$ complex and cell adhesion networks respectively using overrepresentation analysis (See Suppl. Tables.). However, a comparison of FSGs prior to clustering results in 12 subnetworks to be significantly enriched. This suggests that different groups of patients show dysregulation in separate parts of a larger network that are combined into a single cluster based on intermediary interactions.

3.5 Functional Classification of Pancancer Samples

We calculated dysregulation scores for each subnetwork to stratify the cancer samples. We set the support level (k) to 8 for this purpose to increase the number of samples identified during the FSM run since with larger support of 20, many of the samples drop out. As expected, the number of unique frequent subgraphs increased dramatically to 135k, increasing the noise inherent in the frequent subgraph space (14k unique genes). However, dimension reduction shows clear separation of cancer types (See Figures 4, 5 and Suppl. Figures S4, S5 and S6).

While some cancers are spread across multiple clusters (e.g. BRCA), some cancers were separated based on tissue, which reflects implicit biological processed and their alterations (e.g. Uveal melanoma, brain tumors, LIHC, PCPG, THCA etc.) (See also Suppl. Figure S12). Most importantly, survival differences (Figure 4) clearly exist across cancers and cancer subtypes. LGG is split into clusters 11 and 14, where 14 represents GBM-like LGG samples with significant survival differences (57). BRCA clusters 3, 4, 5, 6, 8, and 26 show significant survival differences in these groupings, which reflect previous findings (17, 58).

Significant features between clusters are obtained by comparison of subnetwork dysregulation scores using 1 vs all approach with p-value threshold after Bonferroni correction set as 0.01. Pathway enrichment is done on genes in the significant subnetworks. Pathway enrichment also shows

functionally relevant mechanisms. For instance, clusters 9, 21, and 22 representing Stomach Adenocarcinoma, Rectum Adenocarcinoma, and Colon Adenocarcinoma are significantly enriched for genes related to O-linked glycosylation **(Fig.S6)**. However separately from Rectum and Colon Adenocarcinomas, Stomach Adenocarcinoma is highly enriched for *Defective CSF2RA/CSF2RB causes pulmonary surfactant metabolism dysfunction* pathway, which has been previously associated with Stomach Adenocarcinomas. Interestingly, there is a clear separation of functional mechanisms between clusters: 1, 5, 6, 8, 9, 10, 20, 21, 22, 26 (Group 1), and the rest. More specifically, the second group of cancers is all associated with mechanisms related to signaling events such as RAF/MAP kinase cascades, FGFR signaling, and PI3K Cascade, but the first group is not. These results provide a strong validation for the FSM approach presented.

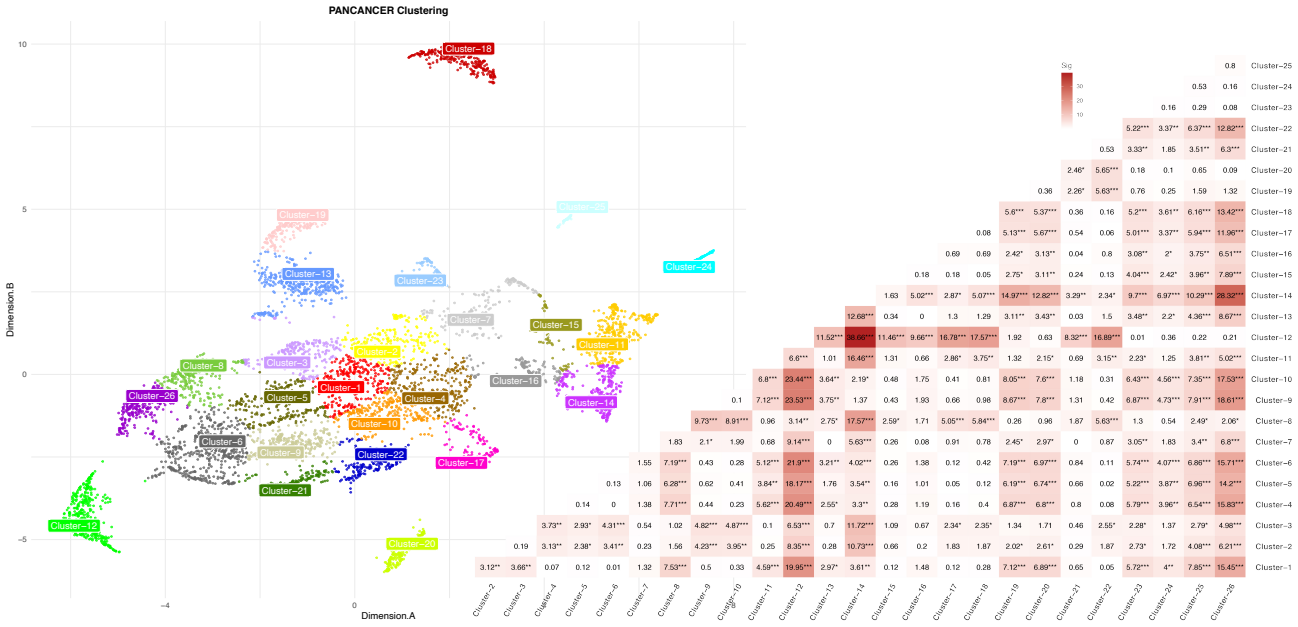


Fig. 4: Left: *UMAP dimensionality reduction on scored frequent subgraph matrix. Samples clusters are labeled and colored based on labels.* ***Right:*** *Pairwise survival differences using log-rank test are shown for FSM patient clusters shown on the left.*

3.6 Analysis of Single Cancers using Pan-cancer Frequent Subgraphs

We have shown further utility of FSGs mined using the pan-cancer dataset to stratify patients into subtypes. We have applied FSG level clustering using PAM and identified significant survival differences **(Fig. S7)**. The significant results were seen for two separate cancers, Lower Grade Gliomas (LGG) and Uterine Cancer suggest that subnetworks mined using the pan-cancer dataset is able to capture subtype-specific functional networks. This further shows the comprehensive nature of our networks identified in this framework.

3.7 Single Cancer Analysis with the FSM Framework

While individual cancer analysis using the pan-cancer FSGs are possible (as shown above Section 3.6), the FSM framework we present can be applied to a single cancer type as well. For this

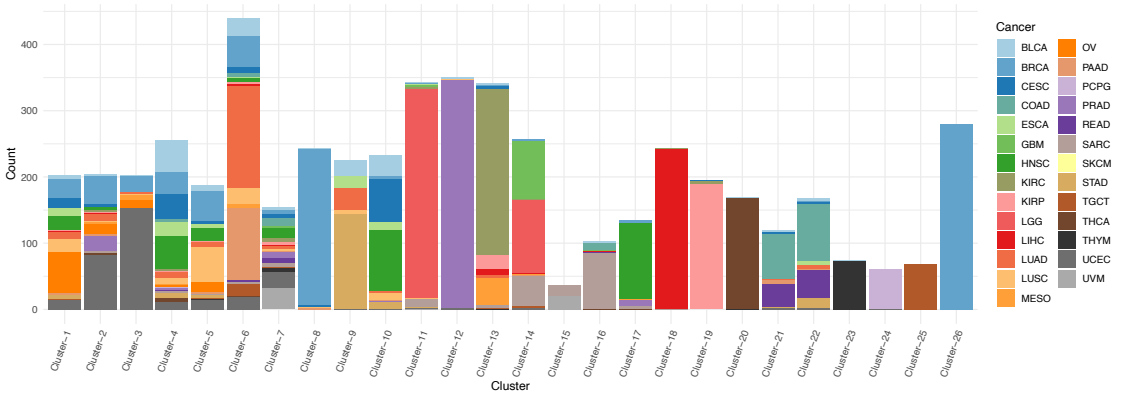


Fig. 5: *Disease profile for each UMAP cluster is shown. The number of patients for each cancer type is stacked on each bar. TCGA disease codes are listed in Supplementary Table S1.*

purpose we analyzed glioblastoma multiforme (GBM) samples only. In a recent study, we used a more simplified FSM framework to cluster individual cancer types and successfully found subtypes for breast cancer and GBM (23). Using our new approach, we have identified 1.2k frequent subgraphs with a total of 5 clusters representing the frequent subgraph network and covering 60% of the GBM samples. The spectrum of the pathway dysregulation in the clusters corresponded to Cytokine Signaling, TRAF6 mediated IRF7 activation, PI3K/AKT signaling, and PIP3 signaling. Interestingly, cluster-2 covered a large fraction of dysregulated pathways and, cluster-4 was enriched specifically for the AKT related pathways. However clusters 3 and 5 showed no pathway enrichment which requires further analysis.

3.8 Survival Differences of Patients Represented in PAM-Clusters

We have investigated the patient groups that correspond to each cluster identified using gene expression and SNP datasets. Pairwise comparison of survival curves show high significance between clusters **(Fig.5)**. For example, cluster-8, which is represented mostly by BRCA patients, shows a significant difference when compared against clusters 1, 4, 5, 6 that are composed of mixed disease types of OV, UCEC, HNSC, LUSC, LUAD, BRCA. Furthermore, the difference between clusters 8 and 26 for BRCA patients only might represent subtype differences as well. Similarly clusters 11 and 14 represent 2 distinct LGG patient clusters with significant survival differences. Interestingly however BRCA, LUAD, LUSC, HNSC, UCEC, and OV cancer types are heterogeneously divided into different clusters suggesting common molecular mechanisms driving the diseases and requires further investigation.

4 Discussion

We have applied frequent subgraph mining coupled with random walk with restarts to the pan-cancer dataset. The application of the FSM with patient-level constraints allowed us to extract interaction patterns functionally relevant to cancer progression. Identified patterns might prove useful for novel targeting strategies especially patient-specific targets due to increased sensitivity in regulatory pattern identification.

The approach proposed in the context of mining functionally important subgraphs is more efficient compared to our initial methodology published (23) both in terms of runtime and coverage.

Biased random walks significantly decrease the search space by reducing the number of edges per patient and applying the FSM separately for each patient as given above ensures that each sample is represented. Furthermore, the use of biased random walks allowed us to increase the sensitivity of our approach by considering the mutational signatures as a network. More specifically, each graph database is obtained based on the mutated genes but frequent subgraphs do not necessarily contain mutated genes but are associated with mutated genes. Additionally, as given above the proposed approach is more comprehensive in comparison with other methods available since gene-level enrichment-based methods or prior knowledge do not take into account the complex interaction patterns relevant to cancer progression.

In comparison to previous methods and established biomarkers, the proposed method underlines the complex interaction patterns present in defining different cancer groups. For example, SEMA3A has been previously associated with breast cancer metastases through the promotion of osteoblast differentiation in MCF-7 cell lines (59). Colony-stimulating factor has also been associated with glioma progression previously and identification of CSF2RA is an important observation (60). p75 neurotrophin receptor also is a crucial regulator of glioma progression leading to cytoskeletal modifications (61). Analysis of GBM patients only increased the sensitivity of frequent subgraphs. PI3K/AKT is responsible for drug resistance for malignant glioma patients, suggesting a critical biomarker in targeted therapies (62).

Furthermore, we have shown that the proposed approach is able to elucidate increased functional relevance by strictly enforcing frequency requirements hence decreasing false positives in contrast with previously established methods that either focus on gene-level approaches or do not consider the underlying topology of the patient data.

Finally, our approach is able to stratify patients of individual cancers based on pancancer frequent subgraphs. In this unsupervised approach, we were able to find significant survival differences in patient groups of LGG and Uterine Cancer. This further validates our approach and shows utility for future cancer studies.

Bibliography

[1] Ciriello, G. et al. *Nature genetics*, 45(10):1127, 2013.
[2] Lawrence, M.S. et al. *Nature*, 499(7457):214, 2013.
[3] Kandoth, C. et al. *Nature*, 502(7471):333, 2013.
[4] Sanchez-Vega, F. et al. *Cell*, 173(2):321–337, 2018.
[5] Werner, H.M.J., Mills, G.B. and Ram, P.T. *Nat Rev Clin Oncol*, 11(3):167–176, 2014.
[6] Hoadley, K.A. et al. *Cell*, 158(4):929–944, 2014.
[7] Hoadley, K.A. et al. *Cell*, 173(2):291–304, 2018.
[8] Bailey, M.H. et al. *Cell*, 173(2):371–385, 2018.
[9] Tamborero, D. et al. *Scientific reports*, 3:2650, 2013.
[10] Hofree, M. et al. *Nature methods*, 10(11):1108, 2013.
[11] Shen, R., Olshen, A.B. and Ladanyi, M. *Bioinformatics*, 25(22):2906–2912, 2009.
[12] Cohen, A.L., Holmen, S.L. and Colman, H. *Curr Neurol Neurosci Rep*, 13(5):345, May 2013.
[13] Dolgin, E. *Cancer Discov*, 9(8):992, Aug 2019.
[14] Sulkowski, P.L. et al. *Sci Transl Med*, 9(375), 02 2017.
[15] Johannessen, T.C.A. et al. *Mol Cancer Res*, 14(10):976–983, 10 2016.
[16] Tateishi, K. et al. *Cancer Cell*, 28(6):773–784, Dec 2015.
[17] van 't Veer, L.J. et al. *Nature*, 415(6871):530–6, Jan 2002.

[18] Paik, S. et al. *N Engl J Med*, 351(27):2817–26, Dec 2004.
[19] Parker, J.S. et al. *J Clin Oncol*, 27(8):1160–7, Mar 2009.
[20] Venet, D. et al. *PLoS Comput Biol*, 7(10):e1002240, 2011.
[21] Dhawan, A. et al. *bioRxiv*, page 203729, 2017.
[22] Leiserson, M.D. et al. *Nature genetics*, 47(2):106, 2015.
[23] Durmaz, A. et al. In *PSB 2017*, pages 402–413. World Scientific, 2017.
[24] Koyutürk, M., Grama, A. and Szpankowski, W. *Bioinformatics*, 20(suppl_1):i200–i207, 2004.
[25] Huan, J. et al. In *CSB*, pages 227–238. World Scientific, 2006.
[26] Zhang, X. and Wang, W. In *null*, page 32. IEEE, 2007.
[27] Kuramochi, M. and Karypis, G. In *Data Mining, 2001. ICDM 2001*, pages 313–320, 2001.
[28] Yan, X. and Han, J. In *Data Mining, 2002. ICDM 2003.*, pages 721–724. IEEE, 2002.
[29] Nijssen, S. and Kok, J.N. In *Proceedings of the tenth ACM SIGKDD*, pages 647–652, 2004.
[30] Ranu, S. and Singh, A.K. In *Data Engineering, 2009. ICDE'09.*, pages 844–855. IEEE, 2009.
[31] Garey, M.R. *Computers and intractability*, 1979.
[32] Can, T., Çamoglu, O. and Singh, A.K. In *Proceedings of the 5th international workshop on Bioinformatics*, pages 61–68. ACM, 2005.
[33] Köhler, S. et al. *AJHG*, 82(4):949–958, 2008.
[34] Erten, S. et al. *BioData mining*, 4(1):19, 2011.
[35] Guo, H. et al. *Scientific reports*, 5:10857, 2015.
[36] Goldman, M. et al. *bioRxiv*, page 326470, 2018.
[37] Szklarczyk, D. et al. *NAR*, 43(D1):D447–D452, 2014.
[38] Fabregat, A. et al. *NAR*, 46(D1):D649–D655, 2017.
[39] Henderson, T.A.D. *Frequent subgraph analysis and its software engineering applications*. Doctoral dissertation, Case Western Reserve University, 2017.
[40] Cheng, H. et al. In *Frequent Pattern Mining*, pages 307–338. Springer Publ., 2014.
[41] Yan, X. and Han, J. In *Proceedings of the Ninth ACM SIGKDD*, pages 286–295, 2003.
[42] Henderson, T.A.D. and Podgurski, A. In *International Workshop on Software Analytics*. ACM, 2016.
[43] Chaoji, V. et al. *Stat. Anal. Data Min.*, 1(2):67–84, jun 2008.
[44] Saha, T.K. and Hasan, M.A. In *2014 IEEE BigData*, pages 72–79, Oct 2014.
[45] Al Hasan, M. and Zaki, M. In *SIAM 2009*, volume 2, pages 646–657, 12 2009.
[46] Li, T. et al. *Nature methods*, 14(1):61, 2017.
[47] McInnes, L., Healy, J. and Melville, J. *arXiv preprint arXiv:1802.03426*, 2018.
[48] Blondel, V.D. et al. *J. Stat. Mech. Theory Exp.*, 2008(10):P10008, 2008.
[49] Rdusseeun, L. and Kaufman, P.J. 1987.
[50] Li, W. et al. *CNS Neurol Disord Drug Targets*, 12(6):750–62, Sep 2013.
[51] Long, J. et al. *Am J Cancer Res*, 8(5):778–791, 2018.
[52] Yang, Y.H. et al. *PLoS One*, 11(2):e0150151, 2016.
[53] Vivet-Noguer, R. et al. *Cancers (Basel)*, 11(7), Jul 2019.
[54] Cote, G.J., Grubbs, E.G. and Hofmann, M.C. *Recent Results Cancer Res*, 204:1–39, 2015.
[55] Subramanian, A. et al. *PNAS*, 102(43):15545–15550, 2005.
[56] Liberzon, A. et al. *Cell systems*, 1(6):417–425, 2015.
[57] Chen, R. et al. *Neurotherapeutics*, 14(2):284–297, 04 2017.
[58] Shimoni, Y. *PLoS Comput Biol*, 14(2):e1006026, 02 2018.
[59] Shen, W.W. et al. *International journal of clinical and experimental pathology*, 8(2):1584, 2015.
[60] Mueller, M.M., Herold-Mende, C.C. et al. *Am. J. Pathol.*, 155(5):1557–1567, 1999.
[61] Johnston, A.L. et al. *PLoS biology*, 5(8):e212, 2007.
[62] Stupp, R. et al. *NEJM*, 352(10):987–996, 2005.
[63] Elseidy, M. et al. *Proc. VLDB Endow.*, 7(7):517–528, March 2014.