Frequent Subgraph Mining of Personalized Signaling Pathway Networks Groups Patients with Frequently Dysregulated Disease Pathways and Predicts Prognosis

Arda Durmaz*

Systems Biology and Bioinformatics Graduate Program,
Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA
Email: arda.durmaz@case.edu

Tim A. D. Henderson*

Department of Electrical Engineering and Computer Science,
Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA
Email: tadh@case.edu

Douglas Brubaker

Department of Biological Engineering,
Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139
Email: dkb50@mit.edu

Gurkan Bebek†

Center for Proteomics and Bioinformatics, Department of Nutrition,
Department of Electrical Engineering and Computer Science,
Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA
Email: gurkan.bebek@case.edu

Motivation: Large scale genomics studies have generated comprehensive molecular characterization of numerous cancer types. Subtypes for many tumor types have been established; however, these classifications are based on molecular characteristics of a small gene sets with limited power to detect dysregulation at the patient level. We hypothesize that frequent graph mining of pathways to gather pathways functionally relevant to tumors can characterize tumor types and provide opportunities for personalized therapies.

Results: In this study we present an integrative omics approach to group patients based on their altered pathway characteristics and show prognostic differences within breast cancer ($p < 9.57E-10$) and glioblastoma multiforme ($p < 0.05$) patients. We were able validate this approach in secondary RNA-Seq datasets with $p < 0.05$ and $p < 0.01$ respectively. We also performed pathway enrichment analysis to further investigate the biological relevance of dysregulated pathways. We compared our approach with network-based classifier algorithms and showed that our unsupervised approach generates more robust and biologically relevant clustering whereas previous approaches failed to report specific functions for similar patient groups or classify patients into prognostic groups.

Conclusions: These results could serve as a means to improve prognosis for future cancer patients, and to provide opportunities for improved treatment options and personalized interventions. The proposed novel graph mining approach is able to integrate PPI networks with gene expression in a biologically sound approach and cluster patients in to clinically distinct groups. We have utilized breast cancer and glioblastoma multiforme datasets from microarray and RNA-Seq platforms and identified disease mechanisms differentiating samples.

Supplementary information: Supplementary methods, figures and tables are available at *arXiv.org*.

*Co-first Author
†Corresponding Author

1. Introduction

Personalized medicine aims to tailor treatment options for patients based on the makeup of their diseases. In the case of cancer, the genetic makeup of tumors is characterized to identify unique tendencies and exploit vulnerabilities of these tumors. However, identifying genomic alterations and molecular signatures that better describe or classify cancer to accomplish this goal has been challenging. Furthermore complex disease phenotypes, such as cancer, cannot be fully explained by individual genes and mutations. Recent studies have explored various approaches to uncover the molecular network signatures of cancers including multivariate linear regression[1] or factor graphs[2] to combine information flow based approaches with copy numbers and DNA methylation data. These techniques identified patient loci with high risk of disease along with genes that are dysregulated for various cancers.[3,4] Gene expression profiles and (in some cases) DNA methylation or metabolomics data have also been used to identify subtypes of the disease.[3–7] However prognostic classification of tumors still requires attention and it is an important step toward identifying most effective approaches in precision medicine.

Glioblastoma multiforme (GBM) is the most common form of malignant brain tumor in adults. GBM is characterized by a median survival of one year and an overall poor prognosis.[8] There have been numerous attempts to classify GBM by differential gene expression to identify clinically and prognostically relevant subtypes.[9,10] Previously methylation status of the *MGMT* promoter is suggested to be associated with tumor response of gliomas to alkylating agents and later associated with increased survival.[11,12] More recently The Cancer Genome Atlas (TCGA) project also provided supporting findings of the methylation status of the *MGMT* promoter as a prognostic marker through analysis of high dimensional data for 206 GBM tumors.[13] Further work utilizing the TCGA data classified GBM by aberrations and gene expression of *EGFR, NF1, and PDGFRA/IDH1* into four subtypes, Classical, Mesenchymal, Neural, and Proneural.[14] These classifications implied strong relationships between subtypes and neural lineages as well as response to aggressive therapy. Though these studies introduced GBM classification, there remained a need to classify dysregulations in tumors more specifically by survivability. While earlier approaches have focused on identifying gene sets,[10,15–18] these had little impact on finding dysregulated pathway segments. For instance, using nearest shrunken centroid classification method,[18,19] or clustering algorithms,[14] gene sets that stratify samples were identified, yet functionally these were not strongly related. Hence, they present little potential for improved treatment opportunities for patients.

Breast Invasive Carcinoma (BRCA) is the most diagnosed cancer among woman consisting of multiple sub classes with distinct clinical outcomes. Previously, 5 subtypes were identified using expression profiles of and later applied to develop predictors by manually selected genes.[6,20,21] Consecutive studies identified differing number of subtypes similar to inital identification. For instance using expression profiles Sotiriou *et al.* identified 6 subtypes further separating luminal-like and basal-like groups.[22,23] Furthermore a comprehensive study integrating multiple omics data to identify unified classification of the breast cancer samples provided strong evidence for 4 subtypes; *Basal, Her-2 enriched, Luminal-A, Luminal-B*.[4] However studies incorporating network or pathway information either used manual selection of pathways or produced limited results. For instance Gatza *et al.* identified 17 subgroups

using pathway based classification with mixed intrinsic subtype signatures.[24]

We describe an integrative omics approach based on frequent subgraph mining (FSM) that brings Protein-Protein Interaction (PPI) networks and gene expression data together to infer molecular networks that are dysregulated in patient samples. We tested our approach using gene expression data for both glioblastoma and breast cancer datasets collected with microarray and next generation sequencing (NGS) approaches. The networks inferred from FSM not only stratify patients into clinically-relevant subtypes, but also provides significant prognostic differences. Our results suggest that a network-based stratification of patients is more informative than using gene-level or feature-based data integration. Identifying personalized dysregulated signaling networks will offer effective means to diagnose and treat patients.

2. Methods

The proposed method uses a novel approach to integrate mRNA expression profiles and PPI networks to identify personalized dysregulated signaling pathways. We hypothesize that dysregulated sub-pathways observed in cancer can discriminate between tumors types which lead to different patient outcomes. We utilized publicly available datasets to develop and validate a method to detect altered molecular signatures in canonical pathways. Our classifications better distinguish patient prognosis in biologically relevant terms than previous studies.[14,25,26]

Our approach is to construct personalized networks of PPIs for cancerous tumors based on mRNA expression data. Section 2.1 details the construction of these networks called *dysregulated signaling pathways*. A network is constructed for each of the patients in each of the datasets used in Section 3. Personalized networks are mined using a new algorithm called QSPLOR (queue explorer) to identify a subset of frequently occurring subgraphs with 4 to 8 proteins as detailed in Sections 2.2 and 2.3. Finally, Non-Negative Matrix Factorization is used to cluster the patients via the frequently occurring subgraphs (Section 2.4 and 2.5).

In Section 3 the clusters are shown to separate patients into short-term and long-term survival groups. The methodology presented has the potential to stratify patients based on their molecular signatures, improve delivery of therapies and assist clinicians and researchers alike to better assess patient prognosis.

2.1. *Dysregulated Signaling Pathways*

Dysregulated Signaling Pathways are labeled graphs (Section 2.2) where vertices represent proteins and edges represent dysregulated activation/inhibition interactions. They are constructed from mRNA expression data (Section 3) and known PPI data.[27,28]

Dysregulation is computed by constructing a matrix $\mathbf{P}$, where $\mathbf{P}_{i,a}$ is the standard score of expression level of gene a for patient i. Then an *interaction matrix* $\mathbf{S}$ constructed from $\mathbf{P}$ in Equation 1. In Equation 1 (ab) represents two genes a and b such that the protein encoded by a interacts with the protein encoded by b. The variable i represents a particular patient.

$$\mathbf{S}_{(ab),i} = \sqrt{\mathbf{P}_{i,a}^2 + \mathbf{P}_{i,b}^2} \tag{1}$$

To determine if the relationship between two genes a and b is dysregulated for patient i the *z-score* for each interaction is computed. In Equation 2, $\mu(\mathbf{S}_{(ab),.})$ and $\sigma(\mathbf{S}_{(ab),.})$ respectively

refer to the mean and standard deviation of the dysregulation scores for genes a and b.

$$Z(\mathbf{S})_{(ab),i} = \frac{\mathbf{S}_{(ab),i} - \mu(\mathbf{S}_{(ab),\cdot})}{\sigma(\mathbf{S}_{(ab),\cdot})} \tag{2}$$

If $Z(\mathbf{S})_{(ab),i} > c$ then an edge $a \to b$ is included in the graph for patient i indicating a and b are dysregulated. In Section 3 the constant c, the z-score threshold, was set to 2 to mine for dysregulation.

2.2. *Frequent Subgraph Mining*

Frequent Subgraph Mining (FSM) is a data mining technique which looks for repeated subgraphs in a graph database. As in Inokuchi *et al.*,[29] the database $\mathcal{D}$ is a set of transactions where each "transaction" is the dysregulated signaling pathways for a patient. FSM detects signaling sub-pathways which are dysregulated in multiple patients.

A dysregulated signaling pathway is a directed labeled graph G consisting of a set of vertices V, a set of edges $E = V \times V$, a set of labels L, and a labeling function which maps vertices (or edges) to labels $l : V|E \to L$. A graph $H = (V_H, E_H, L, l)$ is a subgraph of $G = (V_G, E_G, L, l)$ if $V_H \subseteq V_G$ and $E_H \subseteq E_G$.

A graph H is a subgraph of G ($H \sqsubseteq G$) if there is an injective mapping $m : V_H \to V_G$ s.t.

(1) All vertices in H map vertices in G with the same label: $\forall\ v \in V_H\ [l(v) = l(m(v))]$
(2) All edges match: $\forall\ (u, v) \in E_H\ [(m(u), m(v)) \in E_G]$
(3) All edge labels match: $\forall\ (u, v) \in E_H\ [l(u, v) = l(m(u), m(v))]$

Such a mapping m is known as an *embedding*. The problem of determining if a graph H is a subgraph of G is called the *subgraph isomorphism problem* and is NP-Complete.[30] The *frequency* of a subgraph H is the number of graphs (transactions) in $\mathcal{D}$ which H *embeds* into.

The subgraph relationship $\cdot \sqsubseteq \cdot$ induces a *partial order* on the subgraphs of the graphs in $\mathcal{D}$. That partial order is referred to as the *subgraph lattice*. If the subgraphs in the lattice are all *connected* it is known as the *connected subgraph lattice*. The connected subgraph lattice of $\mathcal{D}$ can be viewed as a graph $\mathcal{L}_\mathcal{D} = (V_\mathcal{L}, E_\mathcal{L})$. The vertices $V_\mathcal{L}$ are all of the connected subgraphs of G. If u and v are both vertices of $\mathcal{L}_\mathcal{D}$ then there is an edge between u and v if and only if $u \sqsubseteq v$ and v and be constructed from u by adding one edge and at most one vertex. The k *frequent connected subgraph lattice* k-$\mathcal{L}_\mathcal{D}$ contains only those subgraphs of graphs in $\mathcal{D}$ which are present in at least k graphs in the graph database $\mathcal{D}$. The leaf nodes of the k-$\mathcal{L}_\mathcal{D}$ are the *maximal frequent subgraphs*.

The objective of frequent subgraph mining is to discover the vertices of k-$\mathcal{L}_\mathcal{D}$. If a subgraph does have at least k transactions it is embedded in, it is known as a *frequent subgraph*. Since finding a frequent subgraph requires repeated subgraph isomorphism queries the problem complexity of FSM is exponential. The number of steps in frequent subgraph mining is bounded from above by $\mathcal{O}(2^g g^h)$ where g is the size of the graph and h is the size of the largest frequent subgraph. The term 2^g is an upper bound on the number of subgraphs of g. Tighter bounds can be obtained if one has more specific knowledge of the graph. The term g^h is an upper bound on number of steps to check if a graph of size h is a subgraph of g.

We present QSPLOR, a new algorithm to find a subset of frequent subgraphs in Section 2.3. It is used to find frequently dysregulated signaling sub-pathways. QSPLOR uses a fixed

```
1  # param start: frequent single vertex subgraphs
2  # param score: a function to score queue items
3  # param max_size: the max size of the queue
4  # param min_sup: int, amount of support
5  # returns: a generator of frequent subgraphs
6  def qsplor(start, score, min_sup):
7      while not start.empty():
8          queue = [ start.pop() ]
9          while not queue.empty()
10             lattice_node = take(queue, score)
11             kids = lattice_node.extend(min_sup)
12             for ext in kids: add(queue, score, ext, max_size)
13             yield subgraph
14 def add(queue, score, item, max_size):
15     queue.append(item)
16     while len(queue) >= max_size:
17         i = argmin(score(idx, queue) for idx in sample(10, len(queue)))
18         queue.drop(i)
19 def take(queue, score):
20     i = argmax(score(idx, queue) for idx in sample(10, len(queue)))
21     return queue.take(i)
```

Fig. 1. QSPLOR: a new algorithm for mining a subset of frequent subgraphs.

amount of memory and a user defined scoring heuristic to guide the search. The algorithm only reports the maximal frequent subgraphs found for compactness. We report only a subset, and not all of frequently dysregulated signaling pathways because (i) it is much faster to report only some of the frequent subgraphs and (ii) using a greater number of frequent subgraphs does not necessarily lead to a more discriminating clustering of samples in our analysis.

There have been a variety of FSM algorithms developed over the last two decades and there are several recent surveys available.[31,32] In recent years interest in collecting representative subsets of frequent subgraphs has emerged.[33,34] Both studies employ random walks on the frequent connected subgraph lattice to collect a sample of the frequent subgraphs. Finally, Leap Search[35] was proposed to find interesting patterns as defined by an objective function.

2.3. *QSPLOR: Mining a Subset of Frequent Subgraphs*

Figure 1 shows pseudo code for QSPLOR a new algorithm to mine a subset of frequent subgraphs. It proceeds as a graph traversal of k-$\mathcal{L}_{\mathcal{D}}$ (the k frequent connected subgraph lattice of the graph database). It begins the traversal at each lattice node representing a frequent subgraph containing only one vertex. At each outer step it initializes a queue with one of the starting lattice nodes. Then in each inner step it removes an item of the queue. The `take` function removes one item from a uniform sample of the queue such that a user supplied scoring function is maximized.

On line 11, the lattice node is extended. This involves finding all possible one edge extensions to the subgraph represented by the lattice node. The ones that are frequent are returned by the extend method. After the extensions are found they are added to the queue with the `add` method. If the queue is at the maximal size after the addition, one item from the queue is dropped. The dropped item is from a uniform sample of the queue and minimizes the user supplied score function. After all extensions have been processed the subgraph is output.

The key to our algorithm is the user supplied scoring function which guides the traversal. The simplest scoring function simply returns a uniform random number. This will cause the traversal to be unguided. Complex scoring functions can prioritize certain labels or structures.

The best general scoring functions are those that prioritize *queue diversity* such that traversal is encouraged to explore as much of the lattice as possible. We use a distance function which captures both structural and labeling differences between graphs as the scoring function for this paper. See the supplementary methods for more details on QSPLOR.

2.4. *Non-Negative Matrix Factorization*

Clustering via Non-Negative Matrix Factorization (NMF) is used to partition patients into subgroups. Section 3 shows that the partitions are prognostically discriminative between the patient subgroups. NMF method was first proposed by Lee and Seung[36] with the aim of decomposing images into explanatory basis vectors. NMF has also been used on gene expression data.[37] For a description of our usage of NMF see the supplementary methods.

2.5. *Clustering Metrics*

Use of NMF requires careful evaluation of the results. Since NMF is based on random initialization of the initial stratification we have applied consensus clustering approach. Using R package NMF[38] we have applied method *'nsNMF'* and random seed with 150 runs. To identify best clustering rank k cophenetic correlation coefficient, silhouette values, residual metrics are evaluated. Cophenetic correlation coefficient is first suggested by Brunet *et al.*[37] to quantify the stability of the clusters. It is calculated as the correlation between sample distances obtained from consensus matrix and the cophenetic distances obtained from hierarchical clustering of the consensus matrix. Brunet *et al.* suggested to choose the ranks where cophenetic correlation coefficient starts to decrease. Silhouette is another method for quantifying cluster stability.[39] The values range between -1 and 1. Intuitively the average silhouette value represents how similar each sample is to the cluster the sample belongs to and how distant from neighbor clusters. Clustering with silhouette values > 0.7 are considered strong as patterns. Residual is the error of the NMF method. Since the method produces an approximation of the original matrix, the residuals represent how close the factorization is to the original data. Note that the residuals decrease naturally as the rank of factorization increases since more variables are added to represent the original matrix.

2.6. *Data Sources*

PPI networks were downloaded from Reactome(v56). Reactome is an expert curated publicly available repository which stores multiple types of relations including reactions, indirect and direct complexes.[27,28] Gene expression data was obtained from previously published studies and TCGA using UCSC Cancer Browser.[40] Clinical data is obtained from both TCGA and corresponding publications (See Figure 2).

3. Results

3.1. *Breast Cancer (Microarray)*

Curtis *et al.*[41] used genomic variations to identify novel subgroups in breast cancer and validated on a sample of 995 patients. Using the same discovery dataset we were able to identify 5 groups with significant differences in survival. QSPLOR mined 145 sub-pathways, with 4-8 proteins each, dysregulated in at least 25 patients.

Fig. 2. Summary of Data including sample and network numbers, median days and interquartile range, sample count of alive and dead event status. In this study both microarray (MA) and RNA-Seq data for breast cancer (BRCA) (MA: [41] and RNA-Seq:[4]) and late stage brain tumors (GBM) (MA:[14] and RNA-Seq:[42]) was utilized.

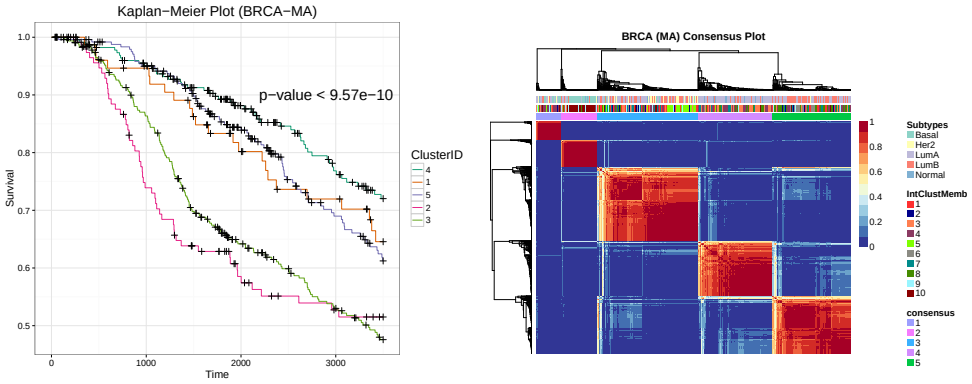
DataSet	Patients	Sub-Pathways	Median Days	Alive/Dead
BRCA MA	995	145	1449	645/350
BRCA RNA-Seq	200	200	1230	685/106
GBM MA	197	553	375	22/175
GBM RNA-Seq	163	548	335	50/113

Consensus clustering and utilization of clustering metrics identified 5 patient groups. The clustering results are similar to clustering of patient samples reported in Curtis *et al.*[41] Identified clusters 1 and 2 matched with clusters 10 and 5 respectively in Curtis *et al.* study as shown in Figure 3b. Furthermore given clusters also match with Basal and Luminal B intrinsic subtypes with further stratification. Compared to previously established subtypes based on the PAM50 classifier, identified clusters are significantly separated in terms of survival(Figure 3a). Enrichment analysis for Reactome pathways in short survivor group revealed pathways that are functionally relevant or predictor of poor survival, i.e. Nonsense-Mediated Decay (NMD),[43] SRP Dependent cotranslational protein targeting to membrane,[44] Selenocysteine synthesis,[45] Signaling by WNT.[46] In contrast, long survivor group was enriched in Neuronal System,[1,45] GABA receptor activation,[47] Signaling by GPCR[48] (See Supplementary Tables S1-S5).

3.2. *Breast Cancer (RNA-Seq)*

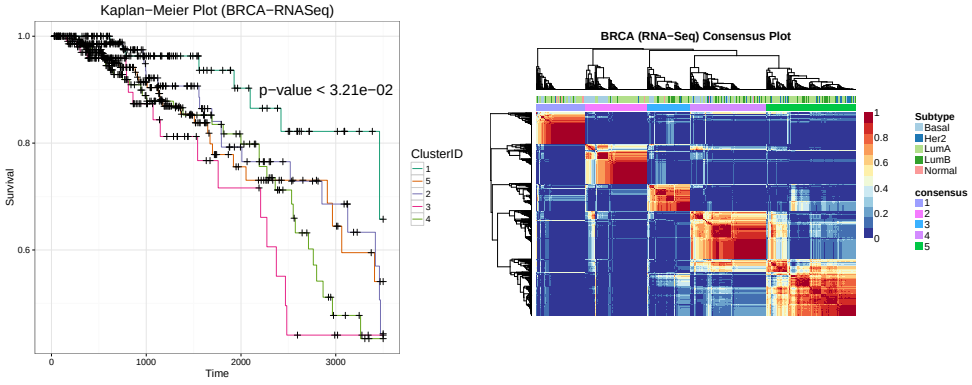
To test the proposed method on breast cancer with data from a different platform, we obtained 791 RNA-Seq samples from TCGA with matching clinical data. QSPLOR identified 200 dysregulated subgraphs. Note that the dataset was not filtered based on prior treatment or patient characteristics hence a heterogeneous dataset was utilized in contrast with breast cancer microarray dataset above. The clustering identified 8 clusters based on cophenetic correlation coefficient and silhouette values. However 8 clusters did not result in significant survival differences hence we have utilized 5 clusters to test whether informative groups were obtained with significant survival differences ($p < 0.05$) (Figure 4a). Reactome pathway enrichment for short survivor group resulted in processes related to cellular division; Mitotic Prometaphase, Separation of Sister Chromatids, Activation of ATR in response to replication stress. Furthermore APC/C-mediated degradation of cell cycle proteins and mitotic proteins pathways were significantly dysregulated. Long survivor group was enriched in immune system related processes; MHC class II antigen presentation, TCR signaling, Cytokine signaling.

We have applied the subgraphs found in microarray dataset to RNA-Seq dataset to check cross-platform application of the proposed method. We were able to identify 5 clusters with significant survival differences. The identified clusters 3 and 4 matched previously identified Basal and Her2 subtypes respectively with further stratification (Figure S16). Pathway enrichment for short and long survivor groups resulted in Keratin metabolims, Signaling by Rho GTPases, Signaling by WNT, Gastrin-CREB signaling pathway via PKC and MAPK, Axon guidance for short survivor group and Signaling by GPCR, EGFR, VEGF, FGFR4, Interleukin-2 signaling for long survivor group (See Supplementary Tables S11-S15).



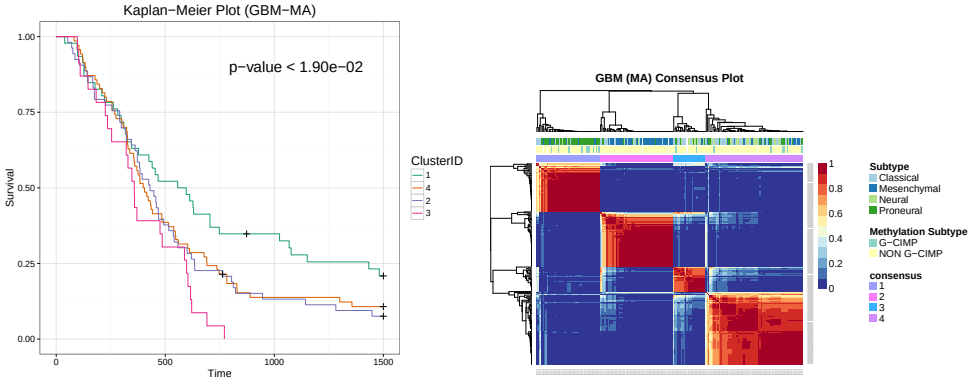
(a) (b)

Fig. 3. Results for breast cancer data analysis used in Curtis *et al.*.[41] **(a)** The Kaplan-Meier plot for 5 groups are shown (Log-rank test $p-value < 9.57E-10$).The x-axis represents days of survival. **(b)** Consensus clustering obtained using NMF is shown. Top bars show novel subtypes clusters, intrinsic subtypes and classification. IntClustMemb shows clusters identified in the Curtis *et al.* study



(a) (b)

Fig. 4. **(a)** Kaplan-Meier and consensus clustering results for breast cancer data obtained from TCGA (Log-rank test $p - value < 3.21E - 02$). Survival is represented as days. **(b)** Top bar in figure shows intrinsic subtypes previously defined, lower bar shows our novel pathway based groups.



(a) (b)

Fig. 5. (a) Survival and consensus clustering results for glioblastoma multiforme microarray data used in.[14] Survival is represented as days and there is a significant difference (Log-rank test $p - value < 1.9E - 02$). (b) Top bar in consensus clustering shows previous classification of GBM patients.

3.3. *Glioblastoma Multiforme (Microarray)*

Using 11861 genes from GBM microarray dataset[14] our method revealed 4 clusters with statistically significant stratification in survival curves ($p-value < 0.05$). The long survivor group 1 consists mostly of proneural subtypes, which also supports the biological implication of our method. A new stratification is visible in Figure 5b for the short survivor group 3.

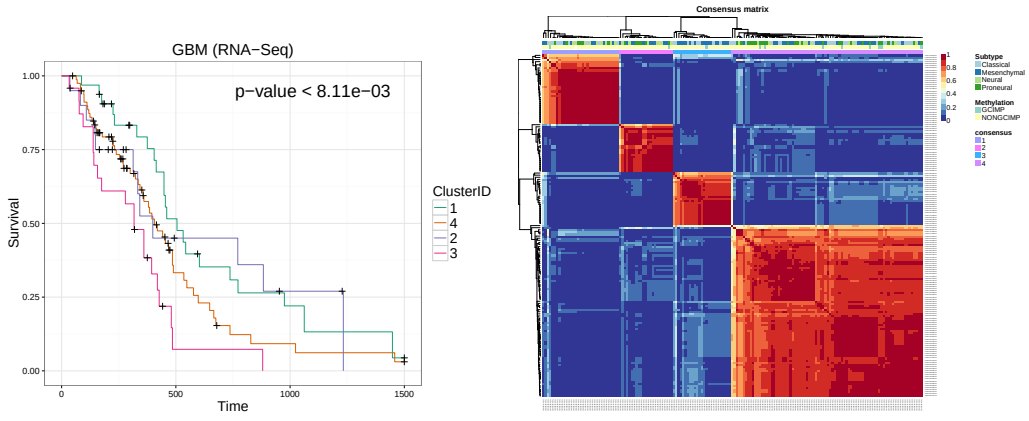
To identify biological implications, we conducted over-representation analysis for Reactome pathways. The long survivor group revealed pathways related to extracellular matrix organization and immune system; axon guidance, collagen degradation, TNFSF mediated activation cascade. The short survivor group was enriched in cell cycle related pathways including: replication, strand elongation and repair. Group 2 shows enrichment for trafficking of GPCR signaling, the Glutamate neurotransmitter release cycle, signaling by Wnt, Gastrin-CREB signaling pathway via PKC and MAPK. Group 4 shows enrichment for respiratory electron transport chain, mitochondrial translation and translation related processes. Overall, the analysis suggests new targets to study for GBM therapy (See Supplementary Tables S16-S19).

3.4. *Glioblastoma Multiforme (RNA-Seq)*

Using GBM data from TCGA[42] which included 15739 genes, our method revealed 4 groups with significant survival *(p-value <0.01)* stratification clustered based on 548 identified subgraphs. As in the microarray data analysis, mesenchymal groups were mostly clustered together in group 3 including the classical subtype. Group 4 is comprised of multiple subtypes suggesting a new classification scheme (Figure 6b). Pathway enrichment results may reveal new biomarkers. Short survivor group 3 was enriched in processes related to cell division; Mitotic prometaphase, Separation of Sister Chromatids, G2/M Transition, DNA Replication. In contrast, long survivor group 1 based on 1 year survival is enriched in Assembly of the primary cilium, Cytokine Signaling in Immune System, Gastrin-CREB Signaling pathway via PKC and MAPK, VEGFA-BEGFR2 Pathway and RET signaling. Interestingly Assembly of the primary cilium is found to be associated with GBM tumors[49,50] (See Supplementary Tables S20-S23).

4. Validation

We compared our method against 2 recently published work integrating PPI and pathway information; *Pathifier* and *NCIS*. (Details of the methods are given in supplementary document) Pathifier identified 6 groups with significant differences in survival (Figure S14a). The number of samples in each group does not suggest biologically relevant clustering (n = 6, and the larger clusters are not significant in terms of survival). The separation distances between groups are not robust with cophenetic correlation coefficient 0.61(Figure S14b). NCIS[25] identified 4 previously established subtypes in the GBM microarray dataset in conjunction with a curated PPI network. The network was constructed by the authors from Reactome, NCI-Nature Curated PID, and KEGG. It consists of 11,648 genes, 211,794 interactions matching 7,183 genes in the GBM dataset. The identified subtypes are similar to established subtypes and have significant differences in survival. However, it is not clear how the patients are clus-



(a) (b)

Fig. 6. (a) Kaplan-Meier and (b) consensus clustering results for glioblastoma multiforme samples obtained from TCGA. The RNA-Seq data set showed significant survival difference (Log-rank $p - value < 8.11E - 03$)

tered since previously identified subtypes do not provide overall significant survival difference (Figure S4). Using the data from NCIS study we have identified 5 clusters (based on the clustering metrics) which show separation of survival curves (Figure S15a). We were able to cluster previously proposed mesenchymal and proneural subtypes with further stratification of mesenchymal group (Figure S15b). Based on the survival analysis, proneural clustered groups show the longest survival curves in agreement with previous findings. These results suggest that the proposed method performed better than the NCIS and Pathifier algorithms in terms of significance of survival stratification and relevance of the identified genes and pathways which can be used as precursor targets for future therapeutic studies.

5. Discussion

The proposed method aims to integrate PPI data with gene expression data using a novel approach. In this study we were able to identify networks that play predictive role in clinical outcome and also networks that crosstalk between the established pathways. A crucial development for improving current prognostic methodologies. The presented method is also more general as it does not require apriori identification of important genes.

Several studies have investigated molecular correlation of prognosis and clinical subclasses in GBM. Earlier studies have identified tumor grade as one of the strong predictors of disease outcome,[51] such as *TP53* mutation and *EGFR* amplifications were claimed to stratify patients into subgroups,[52,53] while a later study contests the validity of this classification.[54] Further studies have identified various gene sets that would separate the patient samples by their molecular characterization,[10,15–18] and some have reported prognostic value of these gene sets. However, most of these have identified different sets of genes, a consensus on the functional delivery has not been reached. These proposed subtype classification methods also identified different sets of patient subtypes, classifications greatly rely on selected patient groups and sample size.

Overall the results suggest possible targets and pathways for cancer progression, mecha-

nisms and survival. Additionally enrichment using long and short survivor groups from RNA-Seq data resulted in similar gene targets. Note that results are 'reversed' for RNA-Seq dataset compared to microarray analyzed samples, however since the stratification is based on dysregulation, the method includes both overexpression or underexpression. Hence genes are categorized as possible markers rather than specific targets for long or short survival.

Our validation of the results we presented here, which reproduced similar survival curves over independent studies, presents great potential for prognostic value for this method. Moreover, finding significant mechanisms that can describe the underlying effects of survival and treatment responses can be easily done within these parameters and provide candidate pathways for therapeutic intervention. While follow up studies are needed to further asses the prognostic value, and possible effect of treatments, analysis that we have conducted provide an initial look of the biological mechanisms underlying in these patient groups with different survival which are also supported by various studies.

Gathering multiple omics datasets to better characterize individuals and associating these with extensive phenotype information has been the hallmark achievement of recent years.[3,4,14,41,42] These datasets have paved the road to improved personalized medicine, promising better disease characterization and diagnosis, identification of patient-specific treatment options and improved monitoring of patients in need. While personalized medicine offers great benefit to individuals, the computational approaches to integrate these multiple omic datasets and statistical methods to leverage the underlying disease and patient traits is still under development. This study tackled this problem of integration network data with transcriptomics data to identify classification scheme for both breast and late stage brain tumors (GBM). Our method can be used to group patients in an unsupervised manner, and have prognostic value. The significant separation of patient samples will allow further studies and utility, since these classifications are based on functionally related frequently altered pathway segments. In the future, we plan to investigate the utility of this method for other cancer types, integrating additional genomic features and investigate its value in improving treatment options.

Acknowledgments

Thank you Leigh Henderson for thoughtful discussions and reading drafts of this paper. This research was partially supported by a Grant from NIH/NCRR CTSA KL2TR000440 to GB.

References

1. Q. Li *et al.*, *Cell* **152**, 633 (Jan 2013).
2. C. J. Vaske *et al.*, *Bioinformatics* **26**, i237 (2010).
3. TCGA, *Nature* **474**, 609 (2011).
4. TCGA, *Nature* **490**, 61 (2012).
5. K. Holm *et al.*, *Breast Cancer Res* **12**, p. R36 (2010).
6. T. Sørlie *et al.*, *PNAS* **98**, 10869 (2001).
7. S. Tardito *et al.*, *Nat Cell Biol* **17**, 1556 (Dec 2015).
8. H. Ohgaki and P. Kleihues, *Acta neuropathologica* **109**, 93 (2005).
9. Y. Liang *et al.*, *PNAS* **102**, 5814 (2005).
10. C. L. Nutt *et al.*, *Cancer research* **63**, 1602 (2003).
11. M. Esteller *et al.*, *New England Journal of Medicine* **343**, 1350 (2000).

12. M. E. Hegi *et al.*, *New England Journal of Medicine* **352**, 997 (2005).
13. TCGA, *Nature* **455**, 1061 (Oct 2008).
14. R. G. Verhaak *et al.*, *Cancer cell* **17**, 98 (2010).
15. H. Colman *et al.*, *Neuro-oncology* , p. nop007 (2009).
16. W. A. Freije *et al.*, *Cancer research* **64**, 6503 (2004).
17. J. M. Nigro *et al.*, *Cancer research* **65**, 1678 (2005).
18. H. S. Phillips *et al.*, *Cancer cell* **9**, 157 (2006).
19. R. Tibshirani *et al.*, *PNAS* **99**, 6567 (2002).
20. C. M. Perou *et al.*, *Nature* **406**, 747 (2000).
21. J. S. Parker *et al.*, *J Clin Oncol* **27**, 1160 (Mar 2009).
22. C. Sotiriou *et al.*, *PNAS* **100**, 10393 (2003).
23. C. Fan *et al.*, *New England Journal of Medicine* **355**, 560 (2006).
24. M. L. Gatza *et al.*, *PNAS* **107**, 6994 (2010).
25. Y. Liu *et al.*, *BMC bioinformatics* **15**, p. 1 (2014).
26. Y. Drier, M. Sheffer and E. Domany, *PNAS* **110**, 6388 (2013).
27. D. Croft *et al.*, *Nucleic acids research* **42**, D472 (2014).
28. M. Milacic *et al.*, *Cancers* **4**, 1180 (2012).
29. A. Inokuchi *et al.*, An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, in *Principles of Data Mining and Knowledge Discovery*, jul 2000 pp. 13–23.
30. S. A. Cook, The complexity of theorem-proving procedures, in *ACM Symposium on Theory of Computing*, (ACM Press, New York, New York, USA, 1971).
31. C. Jiang, F. Coenen and M. Zito, *The Knowledge Engineering Review* **28**, 75 (mar 2013).
32. H. Cheng, X. Yan and J. Han, Mining Graph Patterns, in *Frequent Pattern Mining*, (Springer International Publishing, 2014) pp. 307–338.
33. V. Chaoji, M. Al Hasan, S. Salem, J. Besson and M. J. Zaki, *Stat. Anal. Data Min.* **1**, 67 (2008).
34. M. Al Hasan and M. J. Zaki, Output Space Sampling for Graph Patterns, in *Proceedings of VLDB*, (VLDB Endowment, aug 2009).
35. X. Yan, H. Cheng, J. Han and P. S. Yu, Mining Significant Graph Patterns by Leap Search, in *Proceedings of ACM SIGMOD ICMD*, 2008.
36. D. D. Lee and H. S. Seung, *Nature* **401**, 788 (1999).
37. J.-P. Brunet, P. Tamayo, T. R. Golub and J. P. Mesirov, *PNAS* **101**, 4164 (2004).
38. R. Gaujoux and C. Seoighe, *BMC bioinformatics* **11**, p. 1 (2010).
39. P. J. Rousseeuw, *Journal of computational and applied mathematics* **20**, 53 (1987).
40. J. Z. Sanborn *et al.*, *Nucleic acids research* , p. gkq1113 (2010).
41. C. Curtis *et al.*, *Nature* **486**, 346 (Jun 2012).
42. C. W. Brennan *et al.*, *Cell* **155**, 462 (Oct 2013).
43. L. B. Gardner, *Mol Cancer Res* **8**, 295 (Mar 2010).
44. J. Simões, F. M. Amado, R. Vitorino and L. A. Helguero, *Oncoscience* **2**, 487 (2015).
45. R. L. Schmidt and M. Simonović, *Croat Med J* **53**, 535 (Dec 2012).
46. G.-B. Jang *et al.*, *Sci Rep* **5**, p. 12465 (2015).
47. S. Z. Young and A. Bordey, *Physiology (Bethesda)* **24**, 171 (Jun 2009).
48. A. Singh, J. J. Nunes and B. Ateeq, *Eur J Pharmacol* **763**, 178 (Sep 2015).
49. J. J. Moser, M. J. Fritzler and J. B. Rattner, *BMC cancer* **9**, p. 448 (2009).
50. J. J. Moser, M. J. Fritzler and J. B. Rattner, *BMC clinical pathology* **14**, p. 1 (2014).
51. M. D. Prados and V. Levin, Biology and treatment of malignant glioma., in *Semin Oncol*, 2000.
52. A. von Deimling, D. N. Louis and O. D. Wiestler, *Glia* **15**, 328 (1995).
53. K. Watanabe *et al.*, *Brain pathology* **6**, 217 (1996).
54. Y. Okada *et al.*, *Cancer research* **63**, 413 (2003).