A PATHWAY GRAPH KERNEL BASED MULTI-OMICS APPROACH FOR PATIENT CLUSTERING

by
YASİN İLKAĞAN TEPELİ

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Master of Science

Sabancı University
August 2020

A PATHWAY GRAPH KERNEL BASED MULTI-OMICS APPROACH FOR PATIENT CLUSTERING

Approved by:

Asst. Prof. Dr. Öznur Taştan .....
(Thesis Supervisor)

Assoc. Prof. Dr. Esra Erdem .....

Asst. Prof. Dr. Yavuz Oktay .....

Date of Approval: 18/08/2020

ABSTRACT

A PATHWAY GRAPH KERNEL BASED MULTI-OMICS APPROACH FOR PATIENT CLUSTERING

YASIN İLKAĞAN TEPELI

COMPUTER SCIENCE & ENGINEERING
MSc. THESIS
AUG 2020

Thesis Supervisor: Asst. Prof. Öznur Taştan Okan

Keywords: Cancer, Multi-view clustering, Kernel methods, Graph kernels, Pathways, Multi-omics data,

Accurate classification of patients into molecular subgroups is critical for the development of effective therapeutics and for deciphering the underlining mechanisms for these subgroups. The availability of multi-omics data catalogs for large cohorts of cancer patients provides multiple views into the molecular biology of the tumors and the alterations that take place in patient genes such as mutations and differential expression patterns. At the same time, the molecular interaction networks provide the biological context for these alterations.

We develop PAMOGK (Pathway based Multi Omic Graph Kernel clustering framework) that integrates multi-omics patient data with existing biological knowledge on pathways. We use a novel graph kernel that evaluates patient similarities based on a single molecular alteration type in the context of a pathway. To corroborate multiple views of patients that are evaluated by hundreds of pathways and molecular alteration combinations, we use a multi-view kernel clustering approach.

Applying PAMOGK to kidney renal clear cell carcinoma (KIRC) patients results in four clusters with significantly different survival times (p-value $= 1.24e$-11). When we compare PAMOGK to eight other state-of-the-art multi-omics clustering methods, PAMOGK consistently outperforms these in terms of its ability to partition KIRC patients into groups with different survival distributions. The

discovered patient subgroups also differ with respect to other clinical parameters such as tumor stage and grade, and primary tumor and metastasis tumor spreads. The pathways identified as important are highly relevant to KIRC. We also extend our analysis to eight other cancer types with available mutation, protein and gene expression data. PAMOGK framework is available in github.com/tastanlab/pamogk

ÖZET

HASTA KÜMELEMESİ İÇİN YOLAK ÇİZGE ÇEKİRDEĞİ BAZLI BİR ÇOKLU-OMİK YAKLAŞIMI

YASIN İLKAĞAN TEPELI

BİLGİSAYAR BİLİMİ & MÜHENDİSLİĞİ
YÜKSEK LİSANS TEZİ
AĞUSTOS 2020

Tez Danışmanı: Dr. Öznur Taştan

Anahtar Kelimeler: Kanser, Çoklu-bakış kümeleme, Çekirdek metodları, Çizge çekirdeği, Yolaklar, çoklu-omik verisi

Hastaların moleküler altgruplara doğru sınıflandırılması, etkili tedavilerin geliştirilmesi ve bu alt gruplarda kansere neyin yol açtığını çözmek için önemlidir. Kanser hastalarının büyük kohortları için çoklu omik veri katologlarının erişilebilir olması, somatik mutasyon ya da farklı ifadelenme gibi hasta genlerinde gerçekleşen değişimleri kataloglayarak tümörlerin moleküler biyolojisine çoklu bakış sağlar. Aynı zamanda, moleküler etkileşim ağları da, bu değişimler için biyolojik bağlam sağlar.

Çoklu omik hasta verilerini yolaklardaki mevcut biyolojik bilgi ile birleştiren PAMOGK'u (Yolak tabanlı Çoklu-Omik Çizge Çekirdeği kümelemesi) geliştiriyoruz. Bir yolak bağlamında tek bir moleküler değişim tipine göre hasta benzerliklerini değerlendiren yeni bir çizge çekirdeği geliştiriyoruz. Yüzlerce yol ve moleküler değişiklik kombinasyonları ile değerlendirilen hastaların çekirdek olarak sunulmuş çoklu görüşlerini birleştirmek için çok görüntülü çekirdek kümeleme yöntemini kullanıyoruz.

Berrak hücreli böbrek kanseri (KIRC) hastalarına PAMOGK uygulanması, sağkalım süreleri önemli ölçüde farklı olan dört küme ile sonuçlanır (p-değeri $= 1.24e$-11). PAMOGK'u diğer sekiz en gelişmiş çoklu-omik kümeleme yöntemiyle karşılaştırdığımızda, PAMOGK, KIRC hastalarını farklı sağkalım dağılımları olan gruplara ayırabilme açısından sürekli olarak daha iyi performans gösterir. Bulunan

hasta alt grupları ayrıca tümör evresi ve derecesi ve primer tümör ve metastaz tümör yayılımları gibi diğer klinik parametrelere göre de farklılık gösterir. Önemli olarak tanımlanan yolaklar KIRC ile son derece ilgilidir. Analizimizi mutasyon, protein ve gen ifadesi verilerine sahip sekiz farklı kanser tipi ile genişletiyoruz. PAMOGK'a, github.com/tastanlab/pamogk adresinden ulaşılabilir.

ACKNOWLEDGEMENTS

Foremost, I would like to express my gratitude to my supervisor Asst. Prof. Dr. Oznur Taştan Okan for her support, understanding, patience and immence knowledge. I could not have imagined a better mentor for my studies. Besides my advisor, I would like to thank to my master thesis committee Esra Erdem and Yavuz Oktay for their critical evaluation and feedback.

I also thank co-authors of the paper, whose results are included in these thesis: Ali Burak Unal, Furkan Akdemir for their contribution, help and guidance during this work.

I would also like to thank to other TastanLab members and former members; Duygu Ay, Afshan Nabi, Gulden Olgun, Hamed Mohammadi, Berke Dilekoglu, Halil İbrahim Kuru, Halil Tuvan Gezer, Ege Alpay.

As they were with me during my graduate studies, I would like to thank to Ece, Polen, Yunus Emre, Bahar, Duygu, Hadi, Sahand, Simge, Elif, and Ali. My gratitude also goes to Arda, Uğur, İpek, Ömer, and İrem who always stand behind me with their moral support.

I owe special thanks to my mother Sevilay, my father Yusuf, my brother İlbey, my sister-in-law Özge, and my nephew Göktuğ for their endless support, love, and encouragement. Without them, none of these would be possible.

Last but not least, I'm grateful to Pınar for her understanding, encouragement, continuing support, and patience during my graduate studies. Her constant love, and support kept me miles away from stress and help me to finalize this thesis successfully.

I also thank TUBITAK for the project grant #117E140 and BIDEB for 2210-A scholarship program and Sabanci University for the tuition waiver.

Dedicated
to all children who have not been given an opportunity to receive a proper education...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

α:	Trade-off parameter of SmSPK
λ:	Trade-off parameter of MKKM-MR
$\mathbf{A}_g$:	Adjacency matrix of graph g
C_i:	Patient cluster i
D:	Number of types of molecular alterations
E_i:	Edge list of pathway i
H:	Loss over clustering assignments
I:	Identity matrix
$G_i^{(j)}$:	Undirected vertex labeled graph from pathway i and patient j
k:	Number of patient subgroups
$\mathcal{K}$:	Graph kernel function
$\mathbf{K}$:	Kernel matrix
$\mathbf{K}_\gamma$:	Best kernel matrix of multi-view kernel clustering
$\ell_i^{(j)}$:	Label set of graph from pathway i and patient j
M:	Number of pathways
N:	Number of patients
P:	Number of all pairs of shortest paths
$\mathcal{S}$:	Set of cancer patients
$\mathbf{S}_g^{(t)}$:	Vertex label matrix of g at time t
T:	Transpose
Tr():	Trace of a square matrix
V_i:	Vertex list of pathway i

LIST OF ABBREVIATIONS

AKKM:	Average kernel k-means
BLCA:	Bladder urothelial carcinoma
BRCA:	Breast invasive carcinoma
COAD:	Colon adenocarcinoma
HNSC:	Head and neck squamous cell carcinoma
GBM:	Glioblastoma multiforme
KIRC:	Kidney Renal Clear Cell Carcinoma
LAML:	Acute myeloid leukemia
LUAD:	Lung adenocarcinoma
LUSC:	Lung squamous cell carcinoma
OV:	Ovarian serous cystadenocarcinoma
READ:	Rectum adenocarcinoma
UCEC:	Uterine corpus endometrial carcinoma
LMMKM:	Localized multiple kernel k-means
MKKM-MR:	Multiple kernel k-means with matrix induced regularization
PAMOGK:	Pathway based Multi Omic Graph Kernel clustering
RBF:	Radial Basis Function
SmSPK:	Smoothed shortest path kernel
SNF:	Similarity Network Fusion
TCGA:	The Cancer Genome Atlas

Chapter 1

INTRODUCTION

Cancers are classified based on the tissue of origin. However, cancer is a genomically heterogeneous disease; within the same cancer type, patients bear different molecular alterations and these differences lead to which differences in the progression of the disease and response to therapies (Curtis et al., 2012; Weinstein et al., 2013; Müller et al., 2016). Discovering coherent subgroups of patients with similar molecular profiles is essential to developing better diagnostic tools and subtype specific treatment strategies. The cancer subtypes based on molecular alterations have the potential to guide the clinical decisions for improved therapies (Prasad et al., 2016). Knowledge of molecular subtypes is also key to gain insight into different mechanisms that yield these different subtypes to cancer. The problem of stratifying patients based on their molecular profiles is also relevant for complex diseases other than cancer.

Large scale characterization of patients with omics technologies opens up opportunities to better characterize each cancer (Verhaak et al., 2010; Toss and Cristofanilli, 2015; Curtis et al., 2012). Most of the early approaches rely on single omic data type such as gene expression; rather than multiple data types made available by these technologies. Each omic data presents a *view* into the tumour; combining these different views made available by multiple data types help reach a more detailed and holistic view of the cancer. A *view* typically refers to a feature space and each view stores unique information. A view can also be represented as knowledge graph or kernel matrix instead of feature space. When integrating these data, one would like to capture both the concordant and complementary information across different data sets. Therefore, simply combining data and input to a clustering algorithm does not suffice. To unify different views of the omic data types, several multi-omics clustering methods have been proposed (reviewed in Rappoport and Shamir, 2018a) to integrate the multi-dimensional data collected on patients. We also take

a multi-view clustering approach.

Although corroborating multi-omics data is important to construct a better view of patient similarities, it might not be sufficient to boost the signal as often since only a small fraction of molecular alterations is common among the patients. Analyzing molecular data in the context of molecular networks is a widely used approach to overcome this heterogeneity and sparsity problem (reviewed in Cowen et al., 2017). Therefore, in addition to integrating the multi-omics data, integration with the available prior knowledge is critical. To this end, in this thesis, we present PAMOGK (Pathway based Multi Omic Graph Kernel clustering), a multi-view kernel based clustering approach which integrates multi-omics patient data with pathways using graph kernels.

PAMOGK represents each patient as a set of vertex labeled undirected graphs, where each graph represents the gene interactions in a biological pathway, each vertex represents a gene, each edge represents the interaction between two genes and each vertex label is either discrete or continuous value that represents the patient specific molecular alterations. To quantify patient similarity over a pathway and to attain an omic view, we use a novel graph kernel, the smoothed shortest path graph kernel (SmSPK), whose first version was developed in Unal, 2019. While existing graph kernels are designed to capture the topological similarities of the graphs, SmSPK captures the similarities of the vertex label within the graph context. This allows us to capture patients' similarities that stem from the dysregulation of similar processes in the pathways. By utilizing multi-view kernel clustering approaches, PAMOGK stratifies patients into subgroups. PAMOGK also offers additional insights by showing how informative each pathway and the data type is to the clustering process based on the assigned kernel weights.

We apply our methodology to kidney renal cell carcinoma(KIRC) data made available through the Cancer Genome Atlas Project (TCGA) (Creighton et al., 2013). We integrate the patient somatic mutations, gene expression levels, and protein expression dataset. Compared to the state-of-the-art multi-omics clustering methods, PAMOGK consistently outperforms in terms of its ability to partition into groups with different prognosis. Extracting of the relative importance of pathways in the clustering process show that the discovered pathways that are relevant to KIRC. We also extract patient clusters by applying PAMOGK to other cancer types and evaluate the results. PAMOGK is available at https://github.com/tastanlab/pamogk.

The general framework of multi-view kernel clustering has been presented before in the thesis Unal (2019). This thesis is a follow up on that earlier work with the following specific contributions:

- The representation of the expression data type on the graphs have been updated with continuous labeling.
- Due to problems associated with the earlier pathway dataset used in Unal (2019), the pathway data resource is updated from KEGG (Kyoto Encyclopedia of Genes and Genomes[1]) to (NCI-PID) at NDEXBio(Schaefer et al., 2008).[2]
- Each step of the framework has been evaluated thoroughly and the decisions made are justified by comparison to alternate strategies. In the previous work, the proposed graph kernel was only compared with the radial basis kernel (RBF). Here, we compare SmSPK thoroughly with the state-of-the-art graph kernels. The set of multi-view kernel clustering methods that are evaluated are also expanded.
- In this thesis, we evaluate the framework with different state-of-the-art multi-omics methods for cancer subtyping, which was not conducted in Unal, 2019.
- We apply the framework to eight other cancer types.
- The entire framework is reimplemented in Python with improvements in several steps. [3]
- The updated framework performs better in terms of stratifying patients into well-separated clusters.

The thesis is organized as follows:

- In Chapter 2, we review the traditional and more sophisticated clustering methods that are widely applied to cancer subtyping. In this chapter, we also provide background information on the graph kernel methods to which we compared SmSPK and the different multi-view kernel clustering alternatives that are experimented in the PAMOGK framework.
- Chapter 3 introduces our proposed framework PAMOGK that includes the proposed graph kernel SmSPK and utilizes a multi-view kernel clustering method. We also describe our pathway, patient molecular datasets, and clinical datasets besides the node label assignment techniques.
- We present the results of applying PAMOGK to kidney cancer patients in

[1] https://www.genome.jp/kegg/

[2] https://ndexbio.org/#/networkset/8a2d7ee9-1513-11e9-bb6a-0ac135e8bacf

[3] Mustafa Furkan Akdemir contributed to the implementation

Chapter 4. The various different methods used in alteration mapping, kernel computation and multi-view kernel clustering steps of the framework; the graph kernel, the effect of smoothing, choice of multi-view kernel clustering method are described, evaluated, and discussed. We also present results of comparing PAMOGK with other state-of-the-art multi-omic clustering methods. This chapter also investigates the most informative pathways. Lastly in this section, we also apply PAMOGK to other cancer types and evaluate results in terms of how well patients are clustered.

- We conclude our work and discuss future directions in Chapter 5.

Chapter 2

RELATED WORK

In this chapter, we review the related clustering methods used for grouping patients based on omics data. We review the multi-omics clustering methods. We also discuss related work that make use of pathways in clustering tasks relevant to patient subtyping. Finally, we provide background on graph kernels methods and multi-view kernel clustering methods that have been used in this thesis.

2.1 Traditional Clustering Methods Used in Cancer Subytping

In this section, we review the traditional clustering algorithms that are used with a single-omic. Although there are a high number of methods regarding single-omic, here we focused on the base algorithms such as K-means, hierarchical clustering, consensus clustering, and spectral clustering rather than modified algorithms. Also, note that all of these algorithms can be used as multi-omics clustering methods if datasets are concatenated carefully.

K-means, especially the version that works with kernels (Schölkopf et al., 1998), is one of the most widely used clustering algorithms across different fields and tasks in bioinformatics. It basically partitions samples into k number of clusters where k is a positive integer specified by user. At each iteration, K-means assigns samples into clusters so that the distance between a cluster center and the samples that belong to that cluster is minimized. The objective function is then as follows:

$$\min_{S_k \in S} \sum_{k=1}^{K} \sum_{x \in S_k} ||x - S_k||^2 \tag{2.1}$$

where S_k is the clustering assignments for cluster k, $S = \{S_1, \cdots, S_K\}$ is the set of clusters, K is the number of clusters, and x is a sample.

While modified versions of K-means (Nidheesh et al., 2017; Handhayani and Hiryanto, 2015; Kannan et al., 2016) are used for clustering in genomics, the algorithm is also used as a part of many cancer subtyping methods (Ren et al., 2015; Eason et al., 2018).

Hierarchical Clustering builds a hierarchy between possible clusters. There are two types of hierarchical clustering: agglomerative (Jain and Dubes, 1988)) and divisive (Kaufman and Rousseeuw, 1990). Agglomerative hierarchical approach, which is widely used in finding cancer subtypes, accepts each sample as a cluster at the beginning and combines the clusters that are similar at each step using a similarity metric between clusters and a dissimilarity metric between samples. On the other hand, in the divisive approach, it starts with one cluster which consists of all samples, and this cluster is partitioned at each step.

Hierarchical clustering has been utilized widely in clustering approaches in bioinformatics and cancer subtyping problems (Eisen et al., 1998; Eason et al., 2018; Lapointe et al., 2004; Sotiriou et al., 2003; Bertucci et al., 2005).

Consensus Clustering (Monti et al., 2003a) is an robust approach that combines multiple clustering results from clustering methods. It needs at least one clustering method to work with. The chosen clustering methods are applied to different bootstrapped groups of samples, or patients groups multiple times; and each result is combined at a consensus matrix that shows the frequency of being in the same cluster for each of the sample or patient pairs across different runs. Then the consensus matrix can be used as a similarity matrix, or converted into a dissimilarity matrix to be utilized in clustering.

Consensus clustering is often used in cancer subtyping with other methods such as K-means, hierarchical clustering or non-negative matrix factorization (NMF) (Gan et al., 2018; Eason et al., 2018; Ren et al., 2015)

Spectral Clustering (D. Zhou and Burges, 2007) is graph-based clustering method that make use of K-means. As a first step using methods like the k-nearest neighbor, a weighted similarity graph is constructed between samples. As a second step, a Laplacian matrix $L = D - W$ is constructed between pairs of samples where D is the diagonal degree matrix, and W is the edge weight of sample nodes in the similarity graph. Then the eigenvectors of each sample from L are used as sample features and used in the K-means algorithm to cluster the samples.

Spectral clustering or modified spectral clustering is used in many approaches that aim to cluster samples in single and multi-omics data (Shi and Xu, 2017; Jiang et al., 2019; John et al., 2019; Eason et al., 2018).

Non-negative matrix factorization (NMF) (Lee and Seung, 1999) is a method that assumes data can be represented in a lower dimension. Following this assumption, the input matrix X with dimension $n \times p$ is formed by multiplication of two non-negative $(n \times k)$ W and $(k \times p)$ H matrices:

$$X \approx WH. \tag{2.2}$$

The W and H matrices are found by minimizing the Frobenius Error: $||X - WH||_2^2$. Then the matrix W is used to cluster the samples with lower dimensional data. Later, by minimizing Frobenius error for each dataset separately and adding a new common constraint term to minimization, Jialu Liu et al. (2013) proposed sparse multi NMF for multi-omics data.

NMF is a widely used technique in genomics. The NMF itself or modified versions of NMFs are also utilized in clustering cancer patients (Frigyesi and Höglund, 2008; Ma et al., 2019; Brunet et al., 2004).

2.2 Multi-Omics Clustering Methods

In this section, we review the multi-omics clustering methods, which we compared with our method to cluster cancer patients. Note that we use the term omic here instead of the view, but most of these methods are known as multi-view clustering methods. These multi-omics clustering methods have been reviewed in (Rappoport and Shamir, 2018a) to integrate the multi-dimensional data collected on patients. These methods mainly include three approaches: Early integration, late integration, and intermediate integration.

2.2.1 Early Integration Methods

These types of integration methods apply their algorithms after concatenating the different data types as one data. However, this approach equally weights each dataset and suffers from the curse of dimensionality as the higher dimensional datasets can dominate others in the clustering. While some methods just do concatenation and clustering, several approaches try to overcome the problems of early

integration. The traditional single omic clustering methods can be utilized as a multi-omics clustering approach after data concatenation. There are also more sophisticated methods that are statistical models and assume latent lower-dimensional distribution of data such as **LRACluster and iCluster**.

In **LRACluster** (Wu et al., 2015), a probabilistic model is used to model different types of omics. The probability density function defines the distribution of data over parameters. For binary data, the Bernoulli distribution; for numeric values, Gaussian distribution; and for count data, Poisson distribution is used. After modeling data with distributions, the method finds a low-rank approximation of model parameters by minimizing the sum of minus likelihood functions of different omics. It also uses nuclear norm on the low-rank approximation matrix as regularization. Finally, at each step, the low-rank approximation is clustered using k-means. LRACluster is applied to 11 different cancer types using gene expression, somatic mutation, copy number variation, and DNA methylations as omics. They compared their method with iCluster+ (Mo, S. Wang, et al., 2013) and observed that their method performs better in terms of accuracy, silhouette width, and time.

iCluster(R. Shen et al., 2009) is an early integration method that assumes a latent lower-dimensional distribution of data. Although the main idea of the method is based on lower dimensional distribution, since it concatenates all multi-omics data before applying the method, it can be considered as early integration. Method jointly estimates $(k \times n)$ cluster indicator matrix $Z = (z_1, z_2. \cdots, z_k)'$ by using the model $X_i = W_i \times Z + \epsilon_i$ where X_i is the $p_i \times n$ dataset matrix, W_i is $(p_i \times k)$ coefficient matrix, ϵ_i is the normally distributed independent error matrix, n is the number of samples, k is the number of clusters, and p_i is the number of genes or proteins in omic i. The likelihood of the datasets is maximized using expectation maximization and regularization. K-means is applied to matrix Z at each step to get clustering assignments. The method is applied to both breast and lung cancer separately using gene expression and DNA copy number change.

Later on, **iCluster+** (Mo, S. Wang, et al., 2013) is proposed to deal with categorical and count data in addition to real-valued data. Additionally, a faster method with bayesian regularization, **iClusterBayes** (Mo, R. Shen, et al., 2017) is proposed which does not requires any parameter tuning unlike iCluster+.

2.2.2 Late Integration Methods

A second strategy is to deploy late integration approaches (Rappoport and Shamir, 2018a). In this case, the samples are clustered with each omic data type separately,

and the ensemble's cluster assignments are combined into a single clustering solution. The consensus clustering by Monti et al., 2003b is frequently used for cancer subtyping (Hayes et al., 2006; Verhaak et al., 2010). We review the other methods that also fall into this category. These approaches have the drawback that they do not capture the correlations between the different data types. This strategy leads to poor clustering when each view individually contains a weak signal.

PINS (Nguyen et al., 2017) is proposed as a late integration method, but unlike other late integration methods, it uses the original input when combining clustering results from different data types. It first does perturbation clustering for each omic separately. It then constructs a square connectivity matrix (samples $\times$ samples) for each omic where the value is 1 if patients are in the same clusters and 0, otherwise. Then, connectivity matrices averaged to get a resulting connectivity matrix, or the voting principle is applied to matrices with a threshold to find clusters. Furthermore, they looked whether they could divide the clusters into subclusters. Additionally, perturbation is applied to find the optimal number of clusters. This method is applied to different types of cancer datasets such as KIRC, GBM, LAML, LUSC, BRCA, and COAD from The Cancer Genome Atlas (TCGA) to find patient subgroups within these cancer types. mRNA expression, DNA methylation, and miRNA expression are used as multi-omics. The algorithm is compared against SNF(B. Wang et al., 2014), iCluster+(R. Shen et al., 2009), Consensus clustering(CC)(Monti et al., 2003a), and max silhouette (Rousseeuw, 1987) using Cox regression (Therneau and Grambsch, 2000) p-value as an evaluation metric and observed to be successful on different cancer types.

2.2.3 Intermediate Integration Methods

To overcome the problems of both early and late integration, several intermediate approaches have been proposed.

MCCA (Witten and Tibshirani, 2009) is a modified version of Correlation Canonical Analysis(CCA) (Hotelling, 1936) to make CCA available for multiple views or datasets. The original CCA method finds a linear combination of two omics, respectively X^1 and X^2. It tries to find two projection vectors a^1 with p dimension and a^2 with q dimension that maximizes the correlation between projected vectors of two omics such that:

$$\max_{a^1,a^2} corr(X^1a^1, X^2a^2) \tag{2.3}$$

Here, the projected vectors $U^1 = X^1a^1$ and $U^2 = X^1a^1$ are called canonical variates,

and each pair of canonical variate U_k^1 and U_k^2 are defined with projection vectors a_k^1 and a_k^2. Also, a new pair of canonical variate should be uncorrelated with the canonical variates that are found before. Later, these canonical variates are used for clustering. While there are different types of CCA methods that utilize Bayesian theorem, kernel, or deep learning; the one that supports multiple views is multiple canonical correlation analysis **MCCA** which is using the sum of pairwise correlations. This method is applied to the diffuse large B-cell lymphoma patients. They partitioned DNA copy number data into 24 and use these as multiple datasets.

SNF (Similarity Network Fusion by B. Wang et al., 2014) is one of the similarity-based methods. As a first step, for each omic, it creates a similarity matrix, then using this similarity matrix it creates a similarity network where nodes are the samples, and edge weights are the values in the similarity matrix. Then an iterative procedure based on message passing theory is applied to each similarity network. In this method, each network is updated with the information passed from other networks. When convergence happens, all networks look similar, which is also the resulting similarity network. By using this technique, while weak similarity connections will disappear, the strong and common ones will stay and will have a strong connection in the resulting graph. In the end, the resulting similarity network is converted into a similarity matrix, and spectral clustering is applied.

As an application, SNF is applied to cancer patients of glioblastoma, kidney, breast, lung, and lung cancer from TCGA using the DNA methylation, mRNA, and miRNA expression data. Results show that fusing these datasets with SNF increases performance compared to single-omic experiments.

rMKL-LPP (regularized Multiple Kernel Learning with Locality Preserving Projections by Speicher and Pfeifer, 2015) is a similarity based method that make use of dimension reduction. This algorithm applies dimension reduction to each dataset or omic by preserving the locality which is preserving the similarity between a sample and it's neighbors. The method performs the minimization below to find projection vectors α and kernel weights $\beta = \{beta_1, \cdots, beta_M\}$ to form a linear combination of kernels:

$$\begin{aligned} &\text{minimize}_{\alpha,\beta} \sum_{i,j=1}^{N} \left\| \alpha^T \mathcal{K}^i \beta - \alpha^T \mathcal{K}^j \beta \right\|^2 w_{ij} \\ &\text{subject to } \sum_{i,j=1}^{N} \left\| \alpha^T \mathcal{K}^i \beta \right\|^2 d_{ij} = \text{ const.} \\ &\|\beta\|_1 = 1 \\ &\beta_m \geq 0, \quad m = 1, 2, \ldots, M \end{aligned} \tag{2.4}$$

where $\mathcal{K}^i$ is a matrix that consists of similarity values of a sample i over kernels

and other samples, M is the number of kernels, $\|\beta\|$ stands for the regularization of kernel weights, $\mathbf{W}$ is the matrix consists of w_{ij} which is 1 if i and j are neighbor(k-nearest), 0 otherwise; $\mathbf{D}$ is the constraint matrix consists of d_{ij} which prevents trivial solutions. The terms w_{ij} and d_{ij} are used to preserve locality. The K-means is applied to the resulting representation, and the number of clusters is chosen by using silhouette width.

This method is applied to different cancer types from TCGA such as GBM, KRCCC, LSCC, COAD, and BIC, compared to SNF (B. Wang et al., 2014) and observed to be successful.

2.3 Use of Pathways in Related Prediction Tasks

As the molecular networks are widely used, and pathway graphs become well-defined, some methods utilize biological pathways for different purposes.

Although it is used in a classification task, a new method, Pathway-Induced Multiple Kernel Learning (**PIMKL** by Manica et al., 2019) which utilizes pathways and multiple kernels to classify cancer patients is proposed. The method first defines subnetworks from the biological interaction network in which they use pathway gene-sets. For each subnetwork defined by a pathway, they combine the topology information of genes with the molecular measurements of patients and form a similarity matrix. To do this, they map the molecular measurements from node label space to edge label space which measures the interaction between pathway nodes. Then, they compute the similarity matrix between patients using the normalized Laplacian matrix. This step is called as pathway induction. After defining multiple kernels, they combine these kernels to do classification using EasyMKL (Aiolli and Donini, 2015) which finds the linear combination of these kernels.

PARADIGM (Vaske et al., 2010) is a pathway-based probabilistic approach that uses factor graphs to cluster cancer patients. For each pathway, multi-omics data that belong to a patient are combined to infer integrated pathway activity score, which is the degree of alteration of that patient on the specific pathway. To infer the activity score, the method maximizes the likelihood of the pathway factor graph by utilizing the activation level of genes from the genomic data of the patient. Then these scores form a patient$\times$pathway matrix. As an application, they cluster glioblastoma and breast cancer patients using uncentered correlation hierarchical clustering with centroid linkage using copy number and gene expression data.

2.4 Graph Kernel Approaches

In this section, we review the different graph kernels method to extract patient×patient kernels from patient graphs. The graph kernels we review in this section compares different graphs and find similarities between these graphs. Since our graphs are attributed graphs with continuous values, we choose graph kernels that can deal with continuous attributes. Unless stated otherwise, for the implementations of graph kernels, we used GraKel(Siglidis et al., 2018) graph kernel library.

2.4.1 Shortest Path Graph Kernel

The shortest path graph kernel by Borgwardt et al. (Borgwardt and Kriegel, 2005) is similar to our method in terms of techniques to examine graph. At first, they convert an node-labeled input graph $G = (V, E)$ where each edge has weight as one into a shortest path graph $S = (V, E_s)$ where V is the set of vertices, E is the set of edges, E_s is the set of edges of the new graph and $E_s \subseteq E$. Then, they label the new edges with the length of the shortest path between the corresponding vertices. Afterwards, they define the shortest path kernel function between $S_i = (V_i, E_i)$ from G_i that belongs to patient i and $S_j = (V_j, E_j)$ from G_j that belongs to patient j as follows:

$$\mathbf{K}(S_i, S_j) = \sum_{e_i \in E_i} \sum_{e_j \in E_j} k^{(1)}_{walk}(e_i, e_j) \tag{2.5}$$

where $k^{(1)}_{walk}(e_i, e_j)$ is a positive semi-definite kernel on the edge walks of length 1. For labeled graphs, it is defined as:

$$\begin{aligned} k^{(1)}_{walk}(e_i, e_j) =& k_v(\ell(v_i), \ell(v_j)) k_e(\ell(e_i), \ell(e_j)) k_v(\ell(u_i), \ell(u_j)) + \\ & k_v(\ell(v_i), \ell(u_j)) k_e(\ell(e_i), \ell(e_j)) k_v(\ell(u_i), \ell(v_j)) \end{aligned} \tag{2.6}$$

where $e_i = \{v_i, u_i\}$, $e_j = \{v_j, u_j\}$, k_v is the kernel that compares labels of two vertex and k_e is the kernel that compares shortest path lengths and $\ell()$ stands for the label of a node or an edge. Both k_v and k_e is calculated with dirac kernel.

The most significant difference between is while Shortest path graph kernel compares the topology of the graph and only look for labels of end vertices of shortest paths, SmSPK uses the graph structures and shortest paths as a context. Within this context, it compares the labels of nodes on these paths separately. Originally this method is not time efficient, as it takes $O(n^4)$ time. Although for unlabeled and

labeled graphs, different speed-up techniques are implemented, for the continuous attributes, there is no speed-up technique. It is expected that this graph kernel will not be efficient in terms of time for dense and large graphs.

2.4.2 Propagation Graph Kernel

Propagation Kernel proposed by (Neumann et al., 2016) compares labels of all node pairs between two label-propagated graphs. At each iteration, first, labels of the graph are propagated using $P_{t+1} = TP_t$ where T is the transition matrix and P_t is n(number of node) $\times$ p(number of attributes) node attribute matrix in step t. Secondly, it calculates kernel value for two propagated graphs i and j using the following equation:

$$K(G, G') = \sum_{u \in P} \sum_{v \in P'} k(u, v). \tag{2.7}$$

In the equation, u and v are the nodes of graphs G_i, G_j at iteration t. $k(u, v)$ is calculated using a bin schema where bins are created for node information, and the number of elements in these bins is compared between two graphs. For an efficient calculation, the authors used the Locality Sensitive Hashing method for this kernel method. As a result, this method compares the number of nodes with similar information in a graph by propagating node information, unlike our method, which compares the label information of the same nodes in different contexts. Although patients bear alteration on very different genes, the propagation kernel might find all patients similar since non-altered genes will have label 0, and they generally constitute more than half of the nodes in the graph.

2.4.3 Graph Hopper Kernel

Graph Hopper kernel, like shortest path graph kernels and our graph kernel, compares the shortest paths in graphs. It can deal with labeled and attributed graphs. The basic notation is given as:

$$K(G, G') = \sum_{\pi \in P} \sum_{\pi' \in P'} k_p(\pi, \pi') \tag{2.8}$$

where P is all pairs of shortest paths in graph G and $k_p(\pi, \pi')$ is the kernel which compares the paths π and π'. $k_p(\pi, \pi')$ kernel can be computed as summation of

node kernels along these paths whose lengths are equal:

$$k_p(G, G') = \sum_{j=1}^{|\pi|} k_n(\pi(j), \pi'(j)) \tag{2.9}$$

The node kernel k_n can be delta kernel for labeled graphs and linear kernel or Gaussian kernel for continuous attributed graphs. As we combine 2.8 and 2.9 we can see that this kernel is the weighted sum of node kernels for each vertex pairs between two graphs where the weight is the number of times two vertices are in the same order in shortest paths belongs to two graphs. Like propagation kernel and unlike our graph kernel, this method finds similarity even if the similarly labeled nodes are far away since they are comparing all nodes from two graphs and takes into account how much they are common in terms of being on same length path. Since it does not compare specific nodes that belong to the same genes, it might find all patients as similar if there is at least 1 labeled gene.

2.4.4 Wasserstein Weisfeiler Lehman Graph Kernel

Togninalli et al., 2019b proposed a graph kernel that can deal with attributed graphs with continuous values by combining Wasserstein distance and Weisfeiler Lehman Graph Kernel schema. The method consists of 3 steps: Calculation node embeddings of graphs, calculating Wasserstein distance between graphs using calculated node embeddings, and converting distance to the kernel. In the first step, using Weisfeiler Lehman schema, they find label or attribute of node v_i, $x^h(v_i) =$ at the step h of the schema. While calculating $x_h(.)$, for the labeled graphs they used the same strategy that is used in original Weisfeiler Lehman Graph Kernel (Shervashidze et al., 2011), they also proposed new strategy that is suitable for continues attributed graphs (eq. 5 in (Togninalli et al., 2019b)). At each iteration of Weisfeiler Lehman, they find

$$X_G^h = [x_h(v_1), \cdots, x_h(v_{n_G})] \tag{2.10}$$

as WL features where G is the graph, v_i is the i^{th} vertex and n_G is the number of nodes in graph G. By concatenating these features in each step, they define node embedding matrix of graph G with dimension $(n_G \times m(H+1))$ where m is the number of attributes, and H is the number of iterations.

In the second step, using embeddings of nodes, a ground distance is calculated between graphs. While for the labeled graphs, normalized Hamming distance is used,

Euclidean distance is utilized for attributed graphs. Then, modified Wasserstein distance is calculated using minimization below:

$$\text{Wasserstein Distance}(X, X') = min_P < P, M > \tag{2.11}$$

where M is the ground distances between elements of X and X' that is found before, P is the joint probability between X and X', and $<,>$ is the Frobenius dot product.

Lastly, the kernel is calculated using an instance of a Laplacian kernel for a set of graphs:

$$K = e^{-\lambda D_w^{fWL}} \tag{2.12}$$

where D_w^{fWL} is the Wasserstein distance using Weisfeiler Lehman embedding schema. In terms of performance in the paper, the WWL performs better than both Graph Hopper kernel and traditional Weisfeiler Lehman graph kernel.

2.5 Multi-View Kernel Clustering Methods

Our framework includes multi-view kernel clustering methods in the last step. After forming different kernels that reflect the patient's similarities, from different data types or pathways, a multi-view kernel clustering method is needed to cluster patients using these kernels. Therefore, in this section, we have examined different types of multi-view kernel clustering methods. In these methods, each kernel matrix is considered as a view for samples(patients) to a cluster. These methods aim to combine these kernels and form a single resulting kernel matrix in an unsupervised manner. The most general and popular approach is to find the weight for each kernel(view) and take the weighted average to form the resulting kernel using these weights. These methods mostly formulate an optimization problem where they minimize a loss function to find these weights. For this linear combination strategy, we choose to analyze, Kernel K-Means, Multi-view Kernel K-Means with Matrix Induced Regularization (X. Liu, Dou, et al., 2016), and Localized Multiple Kernel K-Means (Gönen and Margolin, 2014).
On the other hand, the linear combination of kernels is not the only multi-view kernel combination technique. We also extract a kernel fusing part of the Similarity Network Fusion approach and use it as a multi-view kernel clustering approach.

2.5.1 Average Kernel K-Means Method

K-means method which works on feature space is base clustering algorithm for most of the clustering algorithms. This algorithm proposes each cluster with the center of the cluster and minimizes a cost function which includes the sample distances to the center of their clusters. It is possible to apply these method with kernels. Kernel k-means (KKM) (Schölkopf et al., 1998) is proposed to apply K-Means algorithm with kernels and that version solves the optimization problem below 2.13 to find clusters using one kernel:

$$\begin{aligned} \min_{\mathbf{H}\in\mathbb{R}^{n\times k}} \quad & \mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{H}\mathbf{H}^{\mathrm{T}})) \\ \text{s.t.} \quad & \mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I}_k \end{aligned} \tag{2.13}$$

Here $\mathbf{K}$ represents the kernel matrix, $\mathbf{H}$ is the sum of the square loss of over relaxed cluster assignment matrix and $\mathbf{I}_x$ is x-by-x identity matrix. Since it accepts a single kernel matrix, we input the average of the kernel matrices. We will refer to this method as average kernel k-means (AKKM). In other words, the weight of each m kernel matrices becomes $\frac{1}{m}$ when we compute the weighted combination of these kernel matrices. We chose AKKM as the base method to compare with other multi-view kernel clustering algorithms.

2.5.2 Multiple Kernel K-Means with Matrix-Induced Regularization

One of the best performing models which is also applicable to our framework was Multiple Kernel K-Means with Matrix-Induced Regularization (MKKM-MR). Its main purpose is to deal with Multiple Kernel K-Means(MKKM)'s lack of ability to detect relations between kernels. To reduce redundancy among kernel matrices and enhance the diversity of the selected kernel matrices, MKKM-MR (X. Liu, Dou, et al., 2016) uses the matrix-induced regularization and the objective is to minimize sum-of-squared loss over the cluster assignments. The algorithm solves the following optimization problem:

$$\begin{aligned} \min_{\mathbf{H}\in\mathbb{R}^{nxk},\gamma\in\mathbb{R}_+^m} \quad & \mathrm{Tr}(\mathbf{K}_\gamma(\mathbf{I}_n - \mathbf{H}\mathbf{H}^{\mathrm{T}})) + \frac{\lambda}{2}\gamma^{\mathrm{T}}\mathbf{M}\gamma \\ \text{s.t.} \quad & \mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I}_k \\ & \gamma^{\mathrm{T}}1_m = 1 \end{aligned} \tag{2.14}$$

Here, k is the number of clusters, n denotes the number of samples, m is the number of kernel matrices. $\mathbf{H}$ is the relaxed clustering assignment matrix, $\gamma = [\gamma_1, \gamma_2, \cdots, \gamma_m]$

are the weights of input kernel matrices. $\mathbf{K}_\gamma$ is the best kernel matrix, $\mathbf{M}$ is the matrix that measures the relation between kernel matrices. $\mathbf{I}_x$ is the x-by-x dimensional identity matrix, 1_m is m dimensional vector of ones. λ is the parameter that adjusts the trade-off between clustering cost and the regularization term.

In MKKM-MR, one crucial assumption, which can reduce the performance if it is not fulfilled, is that the method assumes the best kernel comes from the linear combination of all kernel matrices.

MKKM-MR is evaluated on well-known datasets like Oxford Flower[1][2], ProteinFold[3], UCI-Digital[4] and Caltech102[5]. It is compared to several algorithms, including average KKM, LMKKM, Multiple KKM, and evaluated in terms of accuracy and normalized mutual information. It outperforms many well-known strategies and methods.

2.5.3 Localized Multiple Kernel K-means

LMKMM is another powerful method that optimizes not only the weight of the kernel matrices but also the weight of the samples(Gönen and Margolin, 2014). We reimplemented LMKKM in Python, which is originally provided in Matlab and R. The objective function of LMKKM is as follows:

$$\begin{aligned} \min_{\mathbf{H}\in\mathbb{R}^{n\times k},\mathbf{\Theta}\in\mathbb{R}_+^{n\times p}} \quad & \mathrm{Tr}(\mathbf{H}^{\mathrm{T}}\mathbf{K}_\Theta\mathbf{H}-\mathbf{K}_\Theta) \\ \text{s.t.} \quad & \mathbf{H}^{\mathrm{T}}\mathbf{H}=\mathbf{I}_k \\ & \mathbf{\Theta}1_p=1_n \end{aligned} \tag{2.15}$$

, where $\mathbf{H}$ is relaxed cluster assignment matrix which includes arbitrary real numbers, $\mathbf{\Theta}=[\theta_1,\theta_2,\cdots,\theta_p]$ is the matrix of the set weights of samples in all kernel matrices in which θ_i is the weight vectors of samples in i^{th} kernel matrix, $\mathbf{K}_\Theta=\sum_{i=1}^{V}(\theta_i{\theta_\mathbf{i}}^T)\odot\mathbf{K}_i$ is the weighted sum of the kernel matrices where "$\odot$" represents the Hadamard product, $\mathbf{I}_k$ is k-by-k identity matrix, 1_x represents x dimensional vector of ones, p is the number of kernel matrices and n is the number of

[1] http://www.robots.ox.ac.uk/~vgg/data/flowers/17/

[2] http://www.robots.ox.ac.uk/~vgg/data/flowers/102/

[3] http://mkl.ucsd.edu/dataset/protein-fold-prediction

[4] http://ss.sysu.edu.cn/~py/

[5] http://mkl.ucsd.edu/dataset/ucsd-mit-caltech-101-mkl-dataset

samples in each kernel matrix. The authors utilized the method to cluster patients from colon and rectal cancer. When evaluated with normalized mutual information, purity, and the Rand index metrics, this method has a better performance compared to single-view kernel k-means or multiple kernel k-means.

Chapter 3

METHODOLOGY

Given a set $\mathcal{S}$ of N cancer patients, for which molecular profiles of the tumors are available, a set of D types of molecular alterations, for each alteration type, every patient's alterations, a set of M pathway graphs and a positive integer k, PAMOGK aims to stratify $\mathcal{S}$ into k subgroups through integrating pathways. Formally, we would like to find a partition $\mathcal{C} = /C_1, C_2, /cdots, C_k/$ of the set $\mathcal{S}$In this section, we detail the steps of PAMOGK and data processing used in our experiment. Let M be the number of pathways, D be the number of types of molecular alterations (mutations, altered expression, etc.) available for the patients, and N be the number of patients.

3.1 PAMOGK Overview

PAMOGK involves three main steps (Figure 3.1). In the first step, each pathway is represented with an undirected graph. Next, for a given molecular alteration type, i.e., somatic mutations, a patient's molecular alterations are mapped on the pathway. These alterations constitute the patient-specific node labels of the patient's graph. Thus, a "view" is constructed for each pathway-molecular alteration type pair. To assess a pair of patients' similarity under a view, in the second step, the novel graph kernel, SmSPK, is computed to quantify a patient pair's similarity over a pathway and a molecular alteration type. Each $N \times N$ kernel matrix constitute a *view* to the patient similarities. In the final step, to stratify cancer patients into meaningful subgroups, these multiple kernels are input to a multi-view kernel clustering algorithm. In the following sections, we elaborate on each step of PAMOGK with more technical details.

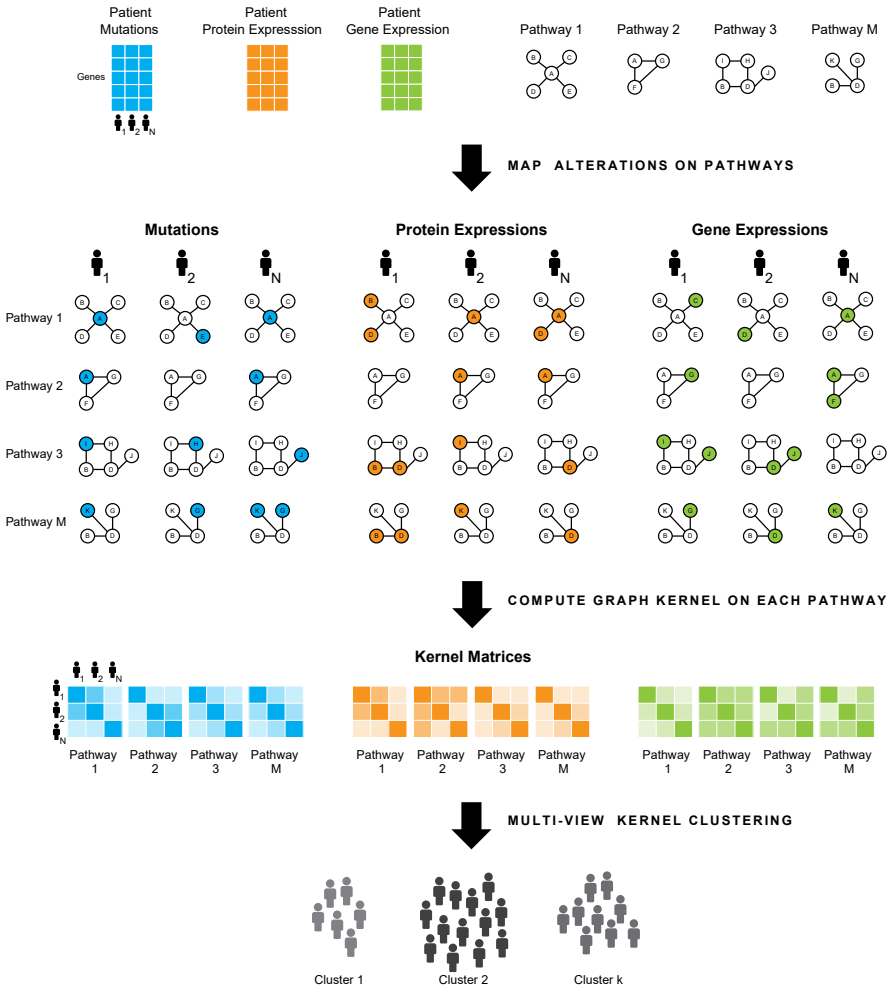


Figure 3.1 The PAMOGK framework (best viewed in color). PAMOGK takes different omic measurements (shown in different colors) and pathways as input. Note that pathway graphs are shown smaller than usual due to size constraints. Each pathway-omic pair constitute a view. In a view, each patient is represented with an undirected graph whose interactions are based on the pathway, and the node labels are molecular alterations of the genes for that patient. For each view, a patient-by-patient graph kernel matrix is computed to assess patient similarities under that pathway-alteration view. In the final step, these views are input to a multi-view kernel clustering method to obtain the patient clusters.

3.2 Step 1: Patient graph representation

We first convert each pathway to an undirected graph where nodes are genes, and an edge exists if there is an interaction between the two genes. For each pathway graph i and patient j, we define an undirected vertex-labeled graph $G_i^{(j)} = (V_i, E_i, \ell_i^{(j)})$. $V_i = \{v_1, v_2, \ldots, v_n\}$ is the set of n genes in the pathway i and $E_i \subset V_i \times V_i$ is a set of undirected edges between the genes in this pathway. The label set $\ell_i^{(j)} = \{l_1, l_2, \ldots, l_n\}$ is in the same order of V_i and represents the corresponding vertex's

label for patient j. For a specific pathway, the pathway graph structure is the same for all patients and is defined by the set of interactions in the pathway while the vertex labels are different and are based on each patient's molecular alterations. For a patient j, $\ell_i^{(j)}$ entries are assigned based on the patient's molecular alteration profile. For example, in the case of somatic mutations, if the corresponding gene k is mutated in patient j, the label of value 1 is assigned to this gene (node), and 0 otherwise. At the end of this step, we have $N \times M \times D$ labeled pathway graphs.

3.3 Step 2: Computing Multi-View Kernels with Graph Kernels

In this step, we would like to assess the similarities of the patients on a given pathway for a given molecular data type. For this, we resort to graph kernel functions. While typical kernels take vectors as input, a graph kernel takes two graphs as input and returns a real-valued number that quantifies the similarity of two input graphs: $\mathcal{K} : \mathcal{G} \times \mathcal{G} \mapsto \mathbb{R}$ (Vishwanathan et al., 2008). Powerful graph kernels are presented in earlier work (Feragen et al., 2013; Shervashidze et al., 2011; Borgwardt and Kriegel, 2005; Neumann et al., 2016; Togninalli et al., 2019b). However, these graph kernels are designed to compare graphs with different graph structures and to identify similarities and differences that arise from these different structures. In our case, though, we would like to compare graphs with identical topology but different node label distribution. The graphs' structures are identical because they are from the same pathway, and the label distributions are different because of the patient specific alterations. To assess the similarity of topologically identical graphs with different node label distribution, we devise a new graph kernel for our purposes.

Inspired from the shortest path graph kernel (Borgwardt and Kriegel, 2005), SmSPK makes use of all shortest paths of the graphs to characterize them. Both methods use the shortest paths of the graphs but in different ways with different end goals. The shortest path kernel (Borgwardt and Kriegel, 2005) compares the end vertices and lengths of the shortest paths in the graph to measure the similarity of the input graphs in terms of their topologies, whereas SmSPK compares the similarity of the node attributes on the shortest paths to capture node attribute similarities. In SmSPK we also smooth the node labels of a patient in the pathway so that if two patients have alterations in genes in close proximity, they contribute to the similarity even though the set of altered genes are not identical. To propagate node labels along the pathway, we use the random walk with restart, which is a common

strategy used in various tasks (reviewed in (Cowen et al., 2017)). For a single graph indexed by g, the label propagation is performed by employing the following formula for all patients:

$$\mathbf{S}_g^{(t+1)} = \alpha \mathbf{S}_g^{(\mathrm{t})} \mathbf{A}_g + (1-\alpha)\, \mathbf{S}_g^{(0)}, \tag{3.1}$$

where $\mathbf{S_g}^{(0)}$ is a patient-by-gene matrix which represents the labels of the vertices in the graph g at time $t = 0$ and each row (patient) is determined by $\ell_g^{(j)}$. $\mathbf{S_g}^{(t)}$ is the node label matrix at time t. $\mathbf{A}_g$ is the degree normalized adjacency matrix of the pathway graph g. $\alpha \in [0, 1]$ is the parameter that defines the degree of smoothing. We iterate over propagation until convergence is attained. We assign node attributes of the graph for each patient based on the final $\mathbf{S}$. Once we attain the label smoothed graphs of $G_g^{(i)}$ and $G_g^{(j)}$, we compute the similarities of these two graphs to each other as follows:

$$\mathcal{K}(G_g^{(i)}, G_g^{(j)}) = \sum_{p=1}^{P} \mathbf{s}_p^{(i)} . \mathbf{s}_p^{(j)} \tag{3.2}$$

Here, $\mathbf{s}_p^{(i)}$ is the vector that represents the labels of the vertices of the graph G_g on the shortest path p for patient i after smoothing, P is the number of all pairs of shortest paths on the graph. The above function is a valid kernel function, as the dot product is the linear kernel, and the kernel property is preserved under summation.

SmSPK is related to the propagation kernel, which also works on the core principle of propagating labels on the graph. The two kernels operate on the same principle of spreading information across the neighbors of a node but assess similarity in different ways. While SmSPK compares the labels of the same nodes on the same shortest paths, the propagation kernel compares node label probability distributions in the entire graph by comparing node label bins. This is not very beneficial in our case because having similar alterations in different parts of the graph contributes to the similarity of the graphs. Another key difference is that while SmSPK completes the smoothing step and then computes the similarity over the shortest paths, the propagation kernel computes similarity at every propagation step to capture the graph's structural differences.

For a given molecular alteration type and a pathway, we compute the SmSPK over all pairs of patients. The resulting matrix $\mathbf{K}$ is a symmetric $N \times N$ matrix, for which the i, j-th entry is the kernel function evaluated for patient i and patient j pair. By computing kernel matrices for each pathway and each molecular alteration type, we obtain $M \times D$ different kernel matrices. We normalize the kernel matrices by dividing the kernel matrix entry $\mathbf{K}(i, j)$ by $\sqrt{(\mathbf{K}(i, i) * \mathbf{K}(j, j)}$ so that all kernel entries are in the range 0 and 1.

3.4 Step 3: Multi-View Kernel Clustering

Each of the kernel matrices computed in the previous section represents a view of the patients' similarities. To integrate these views, we resort to existing multi-view kernel clustering approaches. The multi-view clustering approach allows to identify the clusters and and the weights associated with each of the views in an unsupervised manner. In the literature, there are many available multi-view kernel clustering methods (X. Liu, S. Zhou, et al., 2017). By considering the usability and efficiency of these algorithms, we choose four candidate algorithms to use in the multi-view kernel clustering step of PAMOGK. When we experimented with these four algorithms (see Section 4.3) and among them multiple kernel k-means with matrix-induced regularization (MKKM-MR) (X. Liu, Dou, et al., 2016) yielded the best clustering results. Thus the final model of PAMOGK uses MKKM-MR; yet, this step can be replaced by any multi-view clustering approach as long as the method accepts kernel matrices as input. In this section, for completeness, we provide a brief overview of the selected multi-view kernel clustering methods with which we experimented.

MKKM-MR minimizes the sum-of-squared loss over cluster assignments using matrix-induced regularization to reduce redundancy among kernel matrices and promotes the diversity of the selected kernel matrices.

Kernel k-means (KKM) (Schölkopf et al., 1998) is a simple but a strong baseline algorithm. It accepts a single kernel matrix, for this reason, we input the average of the kernel matrices available for the multiple views. We refer to this method as average kernel k-means (**AKKM**).

LMKMM (Gönen and Margolin, 2014) is another powerful method that optimizes not only the weights of the kernel matrices but also the weight of the samples. We reimplemented LMKKM in Python, which is originally provided in Matlab and R.

Additionally, **SNF** (B. Wang et al., 2014) is one of the multi-omics clustering methods that we review in the Related Work section. It calculates a similarity matrix of samples using an exponential kernel based on the view created by each data type separately and constructs a similarity network for each view. In these networks, samples are nodes and edge weights are the similarities. Through an iterative procedure based on the message passing algorithm, the networks are fused into a single network. In addition to comparing PAMOGK to the SNF algorithm in its original form, we also use the SNF as a possible multi-view clustering method to couple with SmSPK in the PAMOGK framework. Specifically, we compute the patient similar-

ities using SmSPK, fuse them with SNF fusion step, and cluster the samples with kernel k-means or spectral clustering and refer these two versions SNF-KKM and SNF-Spectral, respectively.

3.5 Dataset and Data Preprocessing

Table 3.1 Data sources and their download dates of datasets used in PAMOGK experiments.

Data	Source	Download Date
Somatic Mutations	https://www.synapse.org/#!Synapse:syn1701259	March 24, 2019
Gene Expression	https://www.synapse.org/#!Synapse:syn417925.5	April 24, 2019
Protein Expression	https://www.synapse.org/#!Synapse:syn416783.3	April 24, 2019
Clinical Data	https://www.synapse.org/#!Synapse:syn417024.7	April 24, 2019
Pathway Data	https://ndexbio.org/#/networkset/8a2d7ee9-1513-11e9-bb6a-0ac135e8bacf	April 24, 2019

3.5.1 Pathway data

As the pathway source, we use National Cancer Institute - Pathway Interaction Database (NCI-PID) at NDEXBio (Schaefer et al., 2008)[1]. NCI-PID is a curated database with focus on processes that are relevant to cancer research (download date: Apr 24, 2019). We filter out a pathway if it does not contain a gene that overlaps with the omic data gene list. This filtering results with 165 pathways. The pathway size descriptive statistics are provided in Table 3.2.

Table 3.2 Pathway Size Statistics of 165 pathways.

	Average	Median	Max. number of	Min. number of
Nodes	44.6	42	142	2
Edges	231.9	181	1277	1

3.5.2 Patient molecular and clinical data

The molecular and clinical data for KIRC are obtained from the TCGA PanCancer project (Weinstein et al., 2013). We retrieve the data directly from Synapse[2]. We

[1] https://ndexbio.org/#/networkset/8a2d7ee9-1513-11e9-bb6a-0ac135e8bacf

[2] https://www.synapse.org/#!Synapse:syn300013

only consider the primary solid tumor samples and make use of three different molecular data types that can directly be mapped to pathways: somatic mutations, transcriptomics, and proteomics data. The transcriptomic data include the RNAseq gene expression levels, while protein expression is quantified through Reverse Phase Protein Array (RPPA). The exact data files are listed in Table 3.1 and the number of genes (or proteins) in these data types are provided in Table 3.3.

Table 3.3 Number of unique genes in omics

	Gene expression	**Protein Expression**	**Somatic Mutation**
Number of Genes	17,682	131	13,417

3.5.3 Assigning node labels based on molecular alterations

In the case of mutations, the patient node label is assigned as a binary label based on the presence or absence of the mutation. In the expression datasets, the gene and protein expression values are normalized and converted to z-scores relative to other patients. For each data type, if z-score of a gene (a protein) is greater than 1.96 (which stands for 95% confidence), the gene (the protein) is considered overexpressed in that patient with respect to the other patients while the genes (the proteins) with z-score lower than -1.96 is considered underexpressed. For the graphs generated for the overexpression alteration, we use the z value as the node attribute for the genes where the z-score is larger than 1.96, and for other genes, a node label of 0 is assigned. Similarly, for the underexpression alteration graph, if the z-values are less than -1.96, the z-score is used as the node label and otherwise, a value of 0 is assigned. This threshold label assignment of z-scores considers the extent of change when the gene is over or under expressed. We also considered alternative strategies of defining alteration for the expression dataset; as discussed in the results section, this representation yields better results. Finally, from the three different omic data sources, five different types of alterations are defined: somatic mutation in a gene, over and underexpression based on gene expression, over and underexpression based on protein expression.

Chapter 4

RESULTS AND DISCUSSION

In this chapter, we detail the experimental set up the in which we evaluated PAMOGK and present the results attained. We also include our discussions related to these results.

4.1 Experimental Set up

We apply PAMOGK to discover different subgroups of KIRC patients. The dataset contains 361 patients whose molecular profiles come from three separate data types: somatic mutation, gene expression, and protein expression. We define five different molecular alteration types based on these three types of omics data (see Section 3.5.3). We provide the number of genes and proteins and pathway statistics in Table 3.2 and Table 3.3. We compute one kernel matrix for each pathway-molecular alteration type; this results in 825 kernels (165 pathways $\times$ 5 molecular alteration types), each one of which constitutes a distinct view.

Throughout all experiments, we evaluate four different cluster numbers, $k \in \{2, 3, 4, 5\}$. When computing SmSPK, we try 12 different alpha α values (Table 4.1). We conduct experiments by using different multi-view clustering methods. These include average kernel k-means (AKKM), LMKKM, MKKM-MR, SNF-kernel k-means (SNF-KKM) and SNF-Spectral clustering (SNF-Spectral). The parameter λ in MKKM-MR is chosen using grid-search (Table 4.1). In the original LMKK-means iteration count is used for stopping criteria. Here we add one more criterion which causes stops if the change in objective value is less than 1e-16. The maximum number of iterations is set to 50.

If a pathway kernel includes a few or no altered genes, we eliminate it before in-

putting it into multi-view kernel clustering methods to increase time efficiency. The criteria for this is to eliminate those whose nonzero entries constitute at most 1% of all entries.

Table 4.1 Hyperparameters used in different algorithms. RBF values are selected using the median heuristic.

Parameter	Symbol	Used in	Possible value(s)
Number of clusters	k	All clustering methods	$\{2,3,4,5\}$
Smoothing	α	Kernel construction of SmSPK	$\{0,0.01,0.05,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9\}$
Trade-off	λ	MKKM-MR	$2^{\{-15,-12,-9,\cdots,9,12,15\}}$
RBF	γ	Somatic Mutation	6.41e-03
		Gene Expression	8.01e-04
		Protein Expression	1.11e-01
Number of neighbors	**Ks**	SNF	20
Number of iterations	**Ts**	SNF	20
Max. number of iterations	**Tl**	LMKKM	50

We evaluate the clustering results through survival analysis in accordance with the previous work (Jianfang Liu and al., 2018; Liang et al., 2018; Ricketts and al., 2018; Gabasova et al., 2017). We compare the survival distributions of the clusters using Kaplan-Meier (KM) survival curves (Kaplan and Meier, 1958) and log-rank test's p-value (Harrington and Fleming, 1982). In the log-rank test, we test whether there is a statistical difference between the survival times of the clusters. In comparing alternative methods, we use the p-value of this log-rank test as the performance criteria. Specifically, in the figures presented for ease of display, we use $-\log_{10}$(p-value). The $-\log_{10}$(p-value), smaller the p-value and better the clusters are separated.

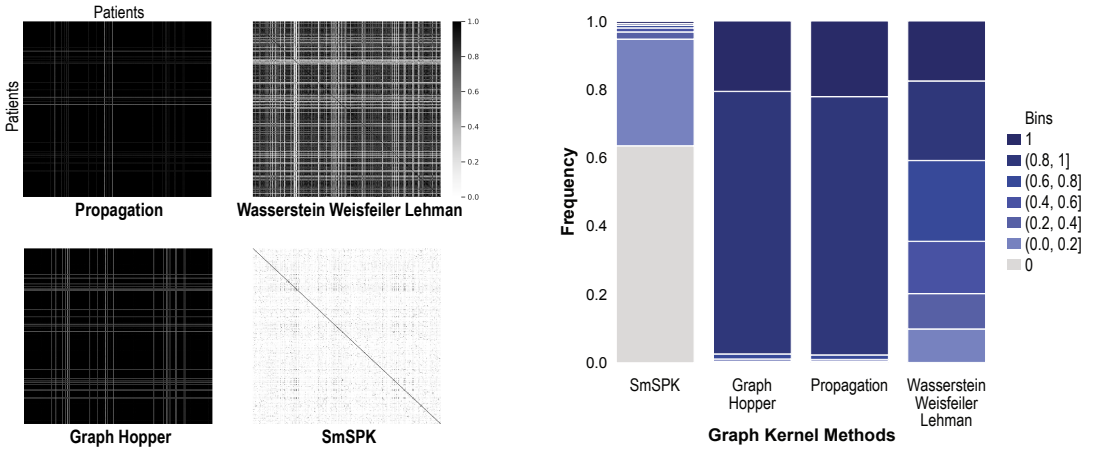
4.2 Assessing the Need of a New Graph Kernel

Constructing kernels that capture the similarity of patients is the crucial step of PAMOGK. First, we would like to understand whether there is any merit in using SmSPK as opposed to deploying an already existing and powerful graph kernel. The motivation behind proposing a new kernel is that the existing graph kernels are designed to capture topological similarities. Since we compare the two patients on the same pathway, the structure of graphs shall always be the same. On the other hand, the node label distribution is different as it is patient specific. Thus, the existing graph kernels computed over the same pathway will consider patients as overly similar and would not serve our purpose.

To check if the intuition above holds, we analyze the distribution of the kernel values computed over all the pathways and the overexpressed molecular alteration type. Since the overexpressed genes are the densest kernels, we choose this data type. We compare SmSPK with kernels that accept continuous node attributes. These

kernels include the propagation kernel (Neumann et al., 2016), Graph Hopper kernel (Feragen et al., 2013) and Wasserstein Weisfeiller Lehman graph kernel (Togninalli et al., 2019b). We also tried to compare our method with node attributed version of shortest path graph kernel (Borgwardt and Kriegel, 2005), but due to slow run time, we abandoned this comparison. We use the implementation provided by the Grakel library (Siglidis et al., 2018) for propagation and graph hopper kernels. We use WWL Python library (Togninalli et al., 2019a) provided by authors for the Wasserstein Weisfeiler Lehman graph kernel.

To assess, whether we need a graph kernel at all, we compare to the case where RBF kernel is used for the kernel computation step. The RBF kernel function: $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|)$, where γ is the kernel parameter and $\mathbf{x}_i$ and $\mathbf{x}_j$ the feature vectors for patients i and j. When RBF kernels are computed, they are directly evaluated on the omic data. Thus, they are computed over all the genes regardless of their participation in a pathway. The gamma values of RBF are determined by the median heuristic (Sejdinovic et al., 2013) (Table 4.1). To make the comparison fair, we run the method with and without smoothing and choose the results with the best smoothing parameter assignment for each method (Table 4.1).



(a) (b)

Figure 4.1 **(a)** Example heatmaps of patient-by-patient kernel matrices calculated by different kernel choices. The kernel functions include the propagation kernel, graph hopper kernel, Wasserstein Weisfeiler Lehman, and SmSPK graph kernel methods. Each kernel belongs to *the direct p53 effectors pathway* and overexpressed gene data type. The color black indicates that the similarity of the two patients is evaluated as 1. **(b)** The frequency of patient similarities for different kernels over all pathways with the overexpression molecular data. For example, the darkest navy indicates the kernel value of 1, and the height of this bar is the proportion of patient-pairs for which the kernel value is evaluated as 1. All the kernels other than SmSPK assign patient similarities of 1 very frequently.

First, we analyze the kernel values computed by each kernel. Figure 4.1a displays the

heatmaps computed by each kernel method on an example pathway (more examples are provided in Figure A.2). The rows and the columns are patients whereas the cell entries are colored as proportional to the kernel value computed for the two patients over a single pathway and data type. Figure 4.1a clearly shows how kernels assign patient similarities of 1 very frequently.

To better analyze this over all pathways, we analyze the distribution of kernel values assigned to patients by each of the different kernels. We bin the kernel matrix entries into groups for each kernel and calculate the frequency of each bin. Next, we calculate the average frequency for each bin across all computed kernel matrices. Figure 4.1b shows, for each graph kernel, how the kernel values are distributed on average. All the kernels other than SmSPK, assign patient similarities of 1 very frequently (the darkest bin). These results confirm our intuition that due to the identical graph structures, the existing graph kernels are unable to distinguish patients with different molecular alterations on the same pathway graph.

Additionally, we compare performances of the kernels based on the lowest p-value attained in the log-rank test on the survival distributions of clusters. In each experiment, each kernel is used with MKKM-MR method, and they are allowed to choose the hyperparameters from a set of predetermined values. These include k for clustering, the smoothing parameter α for SmSPK, λ for MKKM-MR. The best clustering solution obtained for each method is compared in Figure 4.2. We observe that SmSPK outperforms other graph kernels which are compared in this study. This can be explained based on the previous remark that this graph kernel is formulated to distinguish graphs with identical topologies. Additionally, although the use of RBF kernel generally yields good results, the integration of pathway information through SmSPK brings an improvement to the cluster separations in terms of survival.

4.3 Deciding on the Multi-view Kernel Clustering Algorithm to Use in PAMOGK

To determine the multi-view kernel clustering algorithm to be used in PAMOGK, we experiment with different alternatives. The multi-view kernel clustering methods that we analyze include the MKKM-MR (X. Liu, Dou, et al., 2016), AKKM, LMKKM (Gönen and Margolin, 2014) and SNF (B. Wang et al., 2014) with KKM(Kernel K-Means) and spectral clustering (see Section 3.4). For each method, we report the best clustering solution, which is determined based on the lowest

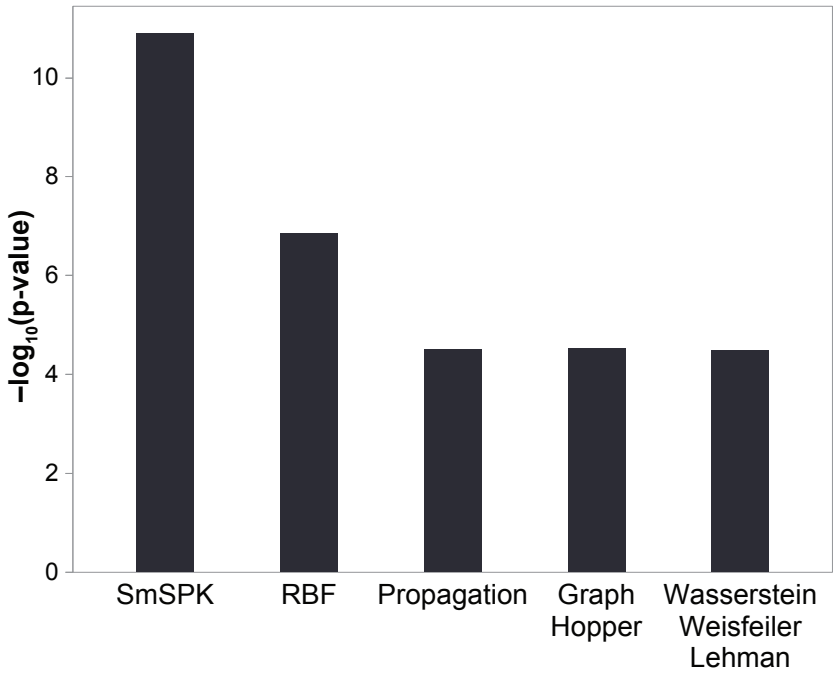


Figure 4.2 The log-rank test p-values obtained with different choices of kernels employed with MKKM-MR multi-view kernel clustering algorithm. Kernel construction methods include SmSPK (our method), propagation graph kernel (Neumann et al., 2016), graph hopper kernel (Feragen et al., 2013), Wasserstein Weisfeiler Lehman graph kernel (Togninalli et al., 2019b) and radial basis function (RBF) kernel.

p-value attained in the log-rank test on the survival distributions of clusters. In each experiment, we allow the methods to choose from a set of predetermined values for each of the hyperparameters. These include k for clustering, the smoothing parameter α for SmSPK, λ for MKKM-MR.

Figure 4.3 summarizes the results in these experiments for the best clustering solution, where $k = 4$. When comparing the three multi-view kernel clustering methods, we observe that MKKM-MR produces the best results. LMKKM, AKKM, and SNF based methods yield similar results with the difference that LMKKM performs slightly better than others. Overall, PAMOGK that uses the MKKM-MR multi-view clustering outperforms all the other clustering alternatives. Thus, we employ MKKM-MR in PAMOGK.

The best clustering solution by PAMOGK is obtained when $k = 4$, smoothing parameter, α is set to 0.3, and λ for MKKM-MR is set to 8. The KM plot of the resulting clustering is provided in Figure 4.4a. The survival distributions significantly differ (log-rank test, p-value $= 1.24e$-11). We should note that the

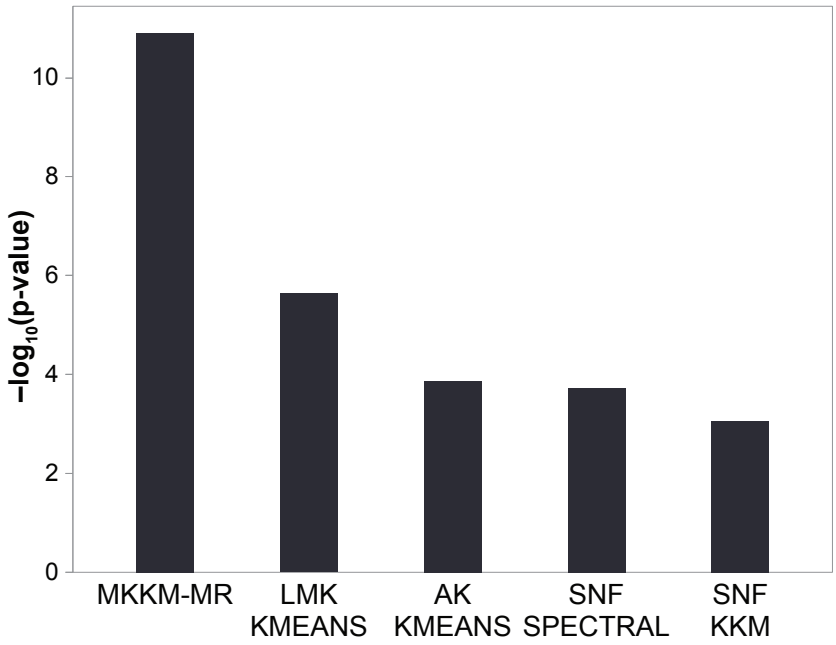


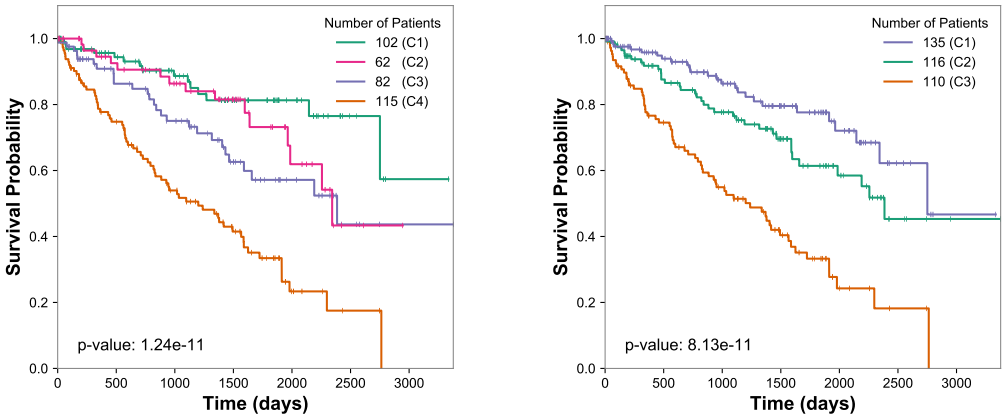
Figure 4.3 The log-rank test p-values obtained with different choices of multi-view kernel clustering methods with SmSPK as the kernel construction method. The clustering methods include average kernel k-means (AKKM), localized multiple kernel k-means (LMKKM) (Gönen and Margolin, 2014), multiple kernel k-means with matrix-induced regularization (MKKM-MR) (X. Liu, Dou, et al., 2016), SNF (B. Wang et al., 2014) with spectral clustering and kernel k-means (KKM).

solution with $k = 3$ is also quite good, p-value $= 8.13e$-11 (Figure 4.4b).

4.4 The Effect of Different Node Label Assignment Strategies for the Expression Data

We checked some of the design choices we made in attaining these results. The first one is how to assign node labels based on alterations in gene or protein expression levels. We compared several alternative strategies,

- **PAMOGK-Disc**: We dichotomize the alterations based on normalized gene expression values, z-scores, and use binary labels. Specifically, we construct two different graphs per one expression data type for a patient, one for overexpression and one for underexpression for each pathway and compute two kernel matrices for patients. In the overexpression graphs, if the gene's z-



(a) (b)

Figure 4.4 **(a)** Kaplan-Meier survival curves of the best clustering solution for KIRC. Result obtained with smoothing parameter $\alpha = 0.3$. The p-value was obtained from a log-rank test between the groups. **(b)** Kaplan-Meier survival curves of the second best clustering solution for KIRC. Result obtained with a smoothing parameter $\alpha = 0.3$. The p-value was obtained from a log-rank test between the groups.

value is larger than 1.96, it receives node label as 1 and 0 otherwise. Similarly, in the underexpression graphs, the z-values < -1.96, receives node label 1, and others are assigned zero.

- **PAMOGK-Cont**: Two graphs per pathway and patient are constructed for the expression data. A gene's node label is set to its z-score. In the first one, genes with positive z-score have these numerical scores as node labels, whereas the other genes are labeled zero. In the second one, only genes with negative z-score have the z-score as node labels, whereas the other genes receive zero labels.

- **PAMOGK-ACont**: One graph per pathway and patient is computed for expression data, as opposed to the separation of underexpression and overexpression. The absolute value of the normalized expression value of the z-score is used.

- **PAMOGK-TCont**: In this one, we include the degree of overexpression and underexpression if the gene is over or underexpressed. In the overexpression graphs, the gene is assigned the node label z if $z > 1.96$, and otherwise, it receives the label 0. Similarly, in the underexpression graphs, the node with $z < -1.96$ receives the label z, and otherwise, it receives the label 0. PAMOGK-TCont considers the extent of the change only if the gene is overexpressed or underexpressed.

- **PAMOGK-BTCont**: Instead of 1.96 threshold, we use the adjusted threshold of Bonferroni(applying Bonferroni correction) by taking into account the number of genes tested. Overexpression and underexpression kernels are constructed as in PAMOGK-TCont, only the threshold is different.

Table 4.2 The different labeling strategies for assigning node labels for the expression graphs.

Method	**Pathway-patient graph labeling**	**Kernel Construction Description for a Single Pathway**	**Best p-value**
PAMOGK-Disc	Label = 1 if $\|z\| > 1.96$; 0 otherwise	Two separate kernels for overexpressed values and underexpressed values	7.47e-10
PAMOGK-Cont	Label = z	Two separate kernels for positive and negative expression	4.17e-09
PAMOGK-ACont	Label = |z|	Single kernel	9.40e-05
PAMOGK-TCont	Label = z if $\|z\| > 1.96$; 0 otherwise	Two separate kernels for overexpressed values and underexpressed values	**1.24e-11**
PAMOGK-BTCont	Bonferroni Correction Label = z if $\|z\|$ > threshold; 0 otherwise	Two separate kernels for overexpressed values and underexpressed values	2.96e-03

The Table 4.2 summarizes the results obtained by each node label assignment strategy. We observe that using the z-value as the node label without considering the extent of over and underexpression does not yield better results than our previous strategy of binary labels PAMOGK-Disc, as seen in the Table 4.2 for the PAMOGK-Cont and PAMOGK-ACont rows. We suspect this is because when the labels are propagated on the graph, for the central nodes, even small z-values accumulate, leading to overly similar patients results. However, when we threshold the values (PAMOGK-TCont) and take the extent of over or underexpression for those that are changed more drastically, it leads to a better separation. Thus, the extent of change might be more meaningful for extreme values. This scheme leads to slightly better log-rank test p-values. We also experiment with the Bonferroni corrected version of this strategy. However, it led to inferior results. Because there are many genes, hardly any gene can pass the threshold; thus, alterations become too sparse.

We further evaluate the three strategies that yield superior results: PAMOGK-Disc, PAMOGK-Cont, PAMOGK-TCont. We conducted an experiment where we bootstrapped the KIRC patient samples, found the best clustering for that set of patients, and compare the log-rank test p-values (Figure 4.5). The higher the values, the better the clusters are separated in terms of survival distributions in this figure. We observe that PAMOGK-TCont yields significantly better results than PAMOGK-Cont at 0.05 significance level (One-tailed Wilcoxon signed-rank test, p-

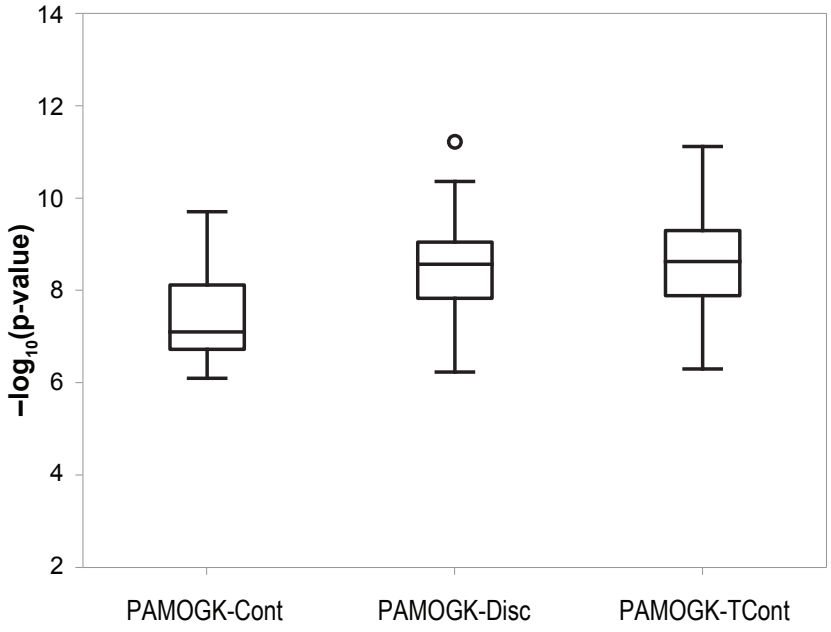


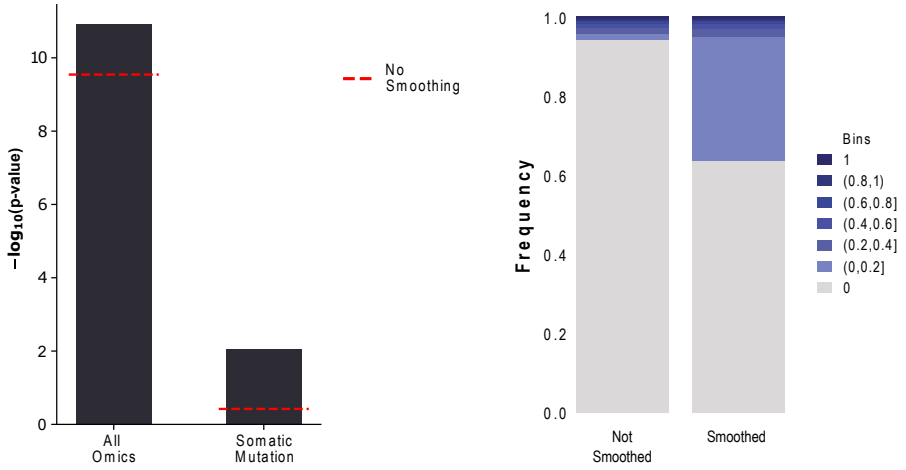
Figure 4.5 Comparison of different node labeling techniques for expression data over 10 different bootstrap samples of KIRC patients. The boxplot shows the -log (p-values) of the log-rank tests conducted on the survival distributions of the clusters attained on each sample. See Section 4.4 and Table 4.2 for a detailed description of each of these labeling strategies.

value 0.00499). When we compare PAMOGK-TCont to PAMOGK-Disc, although it is not statistically significant at 0.05 significance level (One-tailed Wilcoxon signed-rank test, p-value= 0.175), it produces similar or better results in most cases. We conclude that both PAMOGK-Disc and PAMOGK-TCont are good strategies. In this work, we use the PAMOGK-TCont due to its slightly better performance.

4.5 The Effect of Smoothing

Another design choice is whether to use smoothing or not. In selecting the best result for a method, we allowed all methods to choose from a set of alpha values (Table 4.1), and 0 was among them, which corresponds to the case where no smoothing is applied. We observe that the smoothed versions performed better for all methods than the non-smoothed versions of the input graphs. Specifically, to check the effect of smoothing on PAMOGK performance, we examine the results obtained in smoothed and non-smoothed versions. As seen in Figure 4.6a, the smoothed versions achieve better results. We also run the same experiment with only mutation data as com-

mon mutations among patients are rare. With smoothing the best results achieve a p-value $= 8.67e$-03 whereas without smoothing p-value $= 3.79e$-01 is attained. These results show that smoothing is indeed useful in integrating biological knowledge into the framework.



(a) (b)

Figure 4.6 **(a)** The log-rank test p-values obtained for multi-omics data and single-omic data (somatic mutation) with and without smoothing. **(b)** The frequency of patient similarities for SmSPK over all pathways with the overexpression molecular data. For example, the darkest navy indicates the kernel value of 1, and the height of this bar is the proportion of patient-pairs for which the kernel value is 1. When no smoothing is used, more than 90% of the values are evaluated to have zero similarity.

We should note that most patient pairs are evaluated to have zero similarity when there is no smoothing. As an example, Figure 4.6b shows the frequency of patient similarities for different kernels computed over all pathways with the overexpressed data for the smoothing parameter $\alpha = 0.3$. Overexpression is the least scarce alteration type; even in this case, when no smoothing is applied, most patients are evaluated to be dissimilar to each other under different pathways when they do not share any common alteration. In summary, we conclude that smoothing is useful in the evaluation of patients using molecular interactions as context.

4.6 Comparison with the State-of-the Art Multi-Omics Methods

4.6.1 Performance comparison

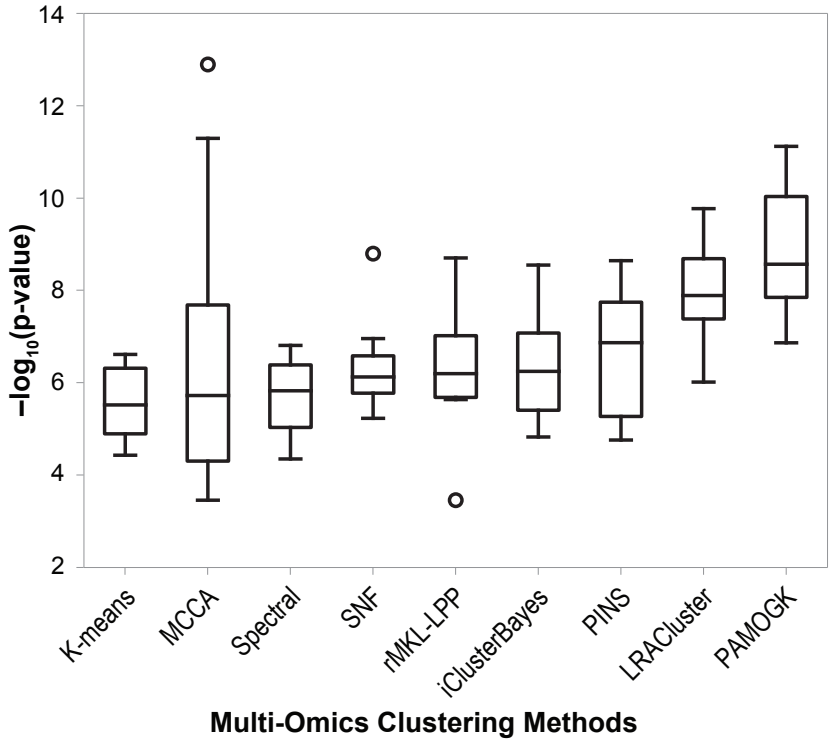


Figure 4.7 Comparison of PAMOGK with the multi-omics clustering methods over 10 different trials. Each trial contains a random subsample of KIRC patients. The boxplot shows the -log (p-values) of the log-rank tests conducted on survival distributions of these clusters. The higher the values, the better the clusters are separated in terms of survival distributions. (Note that PINS method results are over 9 experiments since in one of trial, it did not return a result.)

We compare PAMOGK with eight other multi-omics methods. These include k-means (Lloyd, 1982), MCCA (Witten and Tibshirani, 2009), LRACluster (Wu et al., 2015), rMKL-LPP (Speicher and Pfeifer, 2015), iClusterBayes (Mo, R. Shen, et al., 2017), PINS (Nguyen et al., 2017), SNF (B. Wang et al., 2014), and finally Spectral Clustering (D. Zhou and Burges, 2007). When applying K-means and Spectral Clustering algorithms, the early integration strategy is used, in which the features from different data types are concatenated and then the concatenated data is fed into these methods. These methods cover all methods that are included in a recent comparative benchmark study by Rappoport and Shamir, 2018b with the exception of multiNMF (Jialu Liu et al., 2013). We were not able to to run the source code of this work successfully.

In running these algorithms, we set the maximum number of clusters to five and choose the other parameter configurations for each algorithm exactly as in the benchmark study (Rappoport and Shamir, 2018b). To assess the performance of different methods, we repeatedly subsample the original patient set, and for each subsample, run the algorithms to find the patient clusters. Each subsample contained 300 pa-

Table 4.3 The runtimes in seconds for clustering 361 KIRC patients with the three types of omic data for different methods and PAMOGK.

Method	PAMOGK	LRACluster	PINS	SNF	rMKL-LPP	iClusterBayes	Spectral	K-means	MCCA
Time	352	289	56	7	109	10,898	3	47	6

tients. Due to prohibiting runtime of iClusterBayes, we were able to conduct this experiment 10 times.

The distribution of log-rank test p-values attained by each method is displayed in Figure 4.7. The comparison over ten runs shows that PAMOGK is the best performer among the nine methods. Not only the median performance is high, but even the 90-th percentile of the trials is superior to almost all methods. It also displays low variance across different runs. For all methods, for all trials, the resulting clusters are balanced in terms of the number of patients participating in the clusters except two trials of MCCA. The log-rank test is known to result in unrealistically low p-values when one of cluster size is small (Vandin et al., 2015). In those two trials, MCCA's extremely low p-values are due to clusters with 9 and 14 members.

4.6.2 Runtime comparisons

We conduct a runtime comparison of the algorithms for clustering all the KIRC patients using the three different data types. PAMOGK demands more time to run in comparison to the other methods, with the exception of iClusterBayes Table 4.3. This is because it calculates many more views of the data based on pathways. A second time limiting step is the weight optimization of the kernels in the MKKM-MR algorithm. Despite these additional requirements, the runtime is within reasonable limits, and a typical run takes less than 10 minutes without any parallelization. Replacing the multi-view clustering step with a less demanding algorithm and parallelization could reduce the runtime. Note that the runtime reported in Table 4.3 excludes the time that takes to mapping the alterations on the graphs, which is conducted for once at the beginning for a set of experiments, and takes 8,342 seconds. This costly step can be reduced by using different techniques such as caching. Experiments are conducted on the following system configuration: CPU: Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz CPU. Memory: 256Gb. Operating system: Ubuntu 16.04.4 LTS.

4.7 Detailed Analysis of KIRC Subgroups

Discovered by PAMOGK

In this section, we provide a more detailed analysis of the identified KIRC patient subgroups. For the staging information we use the TNM stating

Table 4.4 Summary of statistical analyses of clinical variables for KIRC subgroups.

Clinical Parameter	Test	p-value
Age	One-way ANOVA	$2.200e$-01
Gender	χ^2	$4.080e$-01
Stage	χ^2	$3.476e$-08
Primary Tumor Pathologic Spread	χ^2	$1.349e$-07
Distant Metastasis Pathologic Spread	χ^2	$3.766e$-04
Neoplasm Histologic Grade	χ^2	$2.104e$-09

4.7.1 KIRC Subgroups' Associations with Other Clinical Parameters

We analyze the association of clinical parameters of the discovered subgroups other than survival. The parameters include age, gender, tumor stage, primary tumor pathological spread, distant metastasis pathological spread, and neoplasm histological grade. The associations of categorical variables are determined using χ^2 test while the continuous variables are tested with one-way ANOVA. We find no statistically significant difference in terms of age (p-value $= 0.220$ in Figure 4.8) and gender (p-value $= 0.408$ Table 4.5). All the other clinical parameters differ across groups at a statistically significant level (see Table 4.4). The distributions of these variables across groups are shown in Figure 4.9 and detailed information is provided in Section A.

Table 4.5 Contingency table for gender vs KIRC clusters. The chi-squared test results in $\chi^2 = 2.893$, $p = 0.408$, df$= 3$

Gender	**Female**	**Male**	**All**
Cluster No			
1	39	63	102
2	22	40	62
3	26	56	82
4	32	83	115
ALL	119	242	361

The best prognosis group is cluster 1, and the worst prognosis group is cluster 4 (Figure 4.4a). There are clear differences between these two groups in terms of these additional clinical parameters. More specifically, 53.9% of the patients in cluster 1 are in stage I, whereas 67.8% of the patients in Cluster 4 are either in

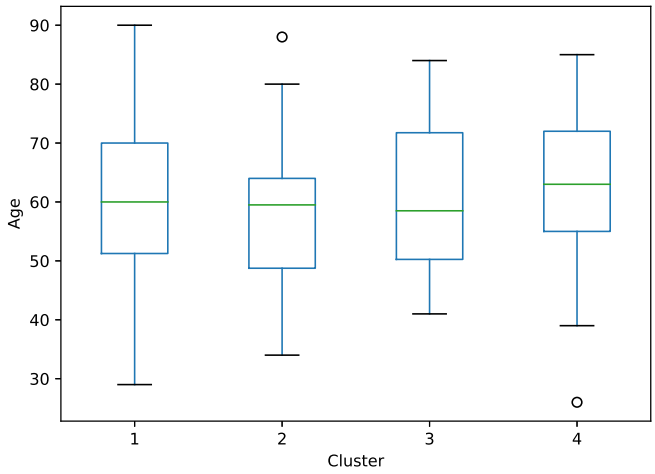


Figure 4.8 Age distribution of patients in each identified RCC cluster. No statistical significance across groups is detected via one-Way ANOVA test (p-value = 0.143)

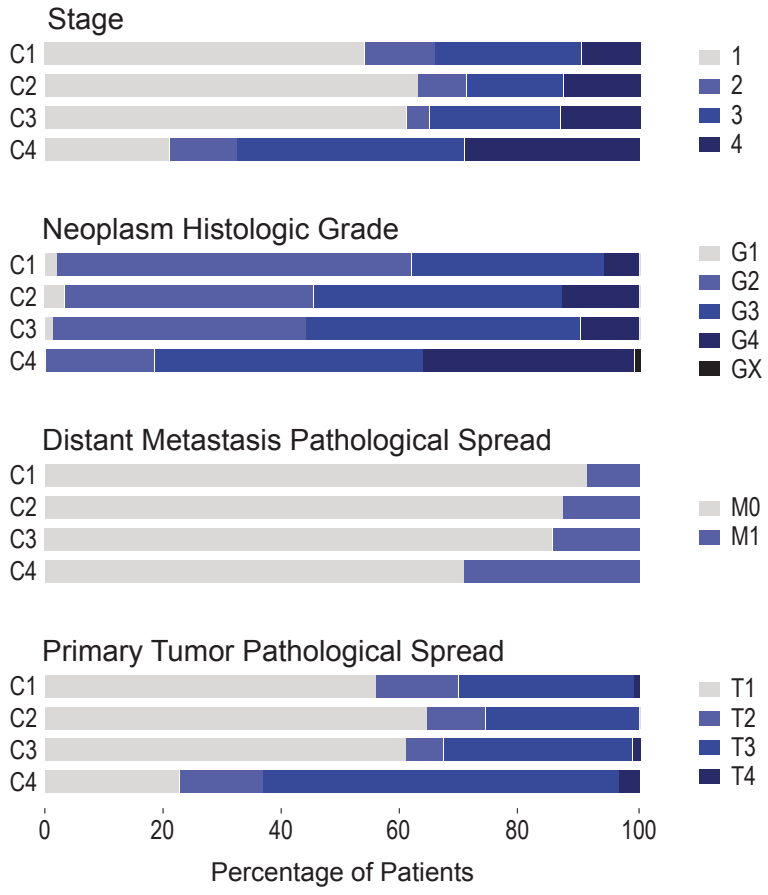


Figure 4.9 The distribution of tumor-related clinical attributes among KIRC clusters.

stage III or Stage IV (Table A.2). Also, nearly half of the patients in cluster 1 have primary tumor T1, whereas 60% of the patients in cluster 4 have primary

tumor T3 (see Table A.3). While only 8.82% of the patients of cluster 1 have distant metastasis, this ratio is 29.6% for cluster 4 patients (Table A.4). Finally, the fraction of cluster 1 patients with histologic grade G1 and G2 is 61.7%, and those with G4 is 5.9%. For cluster 4, the percentage for G1 and G2 drops to 20.5% and G4 increases to 35.7%. (Table A.5). For all prognostic tumor-related features, cluster 1 always has more patients with a lower degree stage and grade, whereas cluster 4 always has more patients with a higher degree stage and grade. Overall, this analysis provides additional evidence that PAMOGK partitions KIRC patients into clinically meaningful subgroups.

4.7.2 Influential pathways and data types

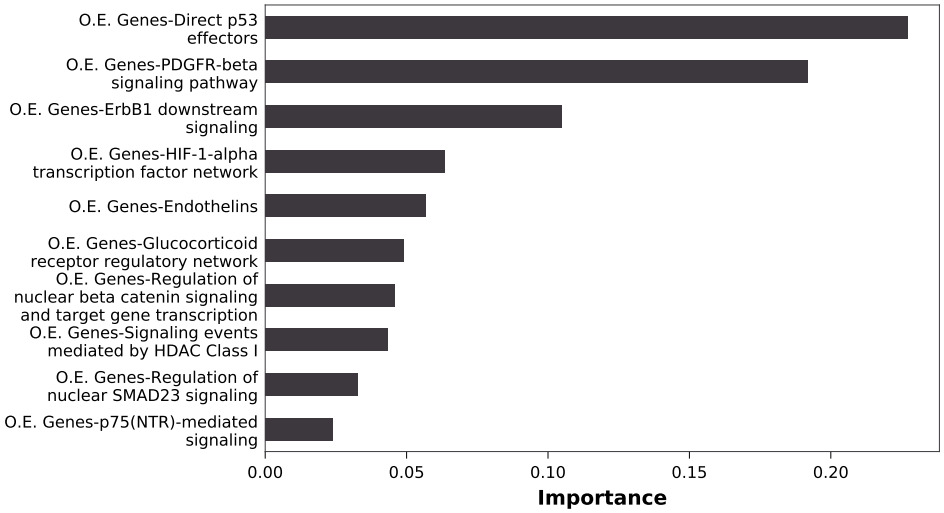


Figure 4.10 Top 10 most influential pathway-alteration type pairs for KIRC. O.E. stands for overexpressed and U.E. stands for under-expression. The relative importance is calculated based on the weights assigned to each kernel matrix of the associated pair by the MKKM-MR algorithm. The results are obtained for the best clustering solution, where the number of cluster is 4, kernel matrices are calculated using SmSPK with smoothing parameter $\alpha = 0.3$.

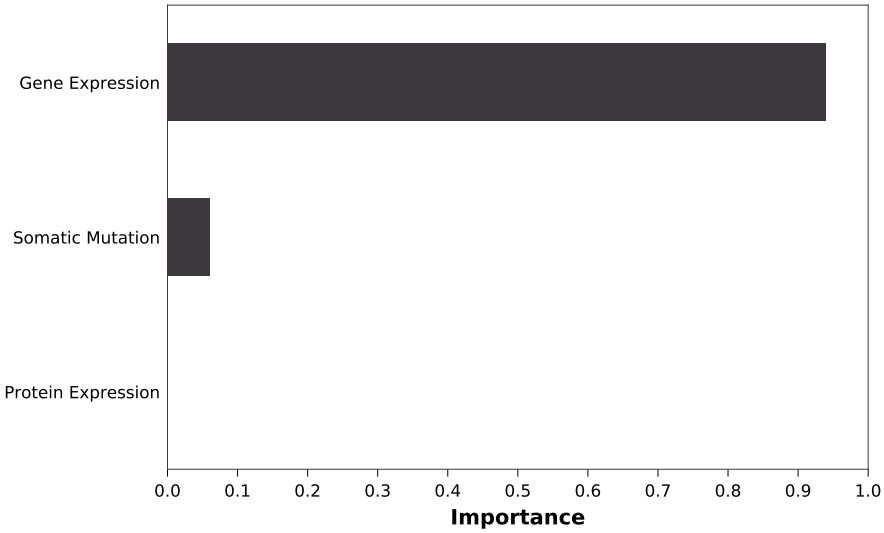


Figure 4.11 Relative importance of the three omic data types for KIRC. One data type weight was calculated by summing up the kernel weights that is available for molecular alteration type and pathway pair. The results are obtained for the best clustering solution, where the number of clusters is 4, and the kernel matrices are calculated by SmSPK with smoothing parameter $\alpha = 0.3$.

By inspecting the assigned kernel weights, we can quantify the relative importance of pathways and molecular data types. For KIRC ($k = 4$), the *direct p53 effectors* pathway and gene overexpression kernel emerge as the most important pathway-molecular alteration pair (see Figure 4.10 for the top 10 pairs). By averaging the weights associated with each omic data type, we find that the gene expression is the top important data type, while protein expression data have almost no effect on the clustering (Figure 4.11). This could be arising from the fact that the protein expression data covers only a small number of proteins.

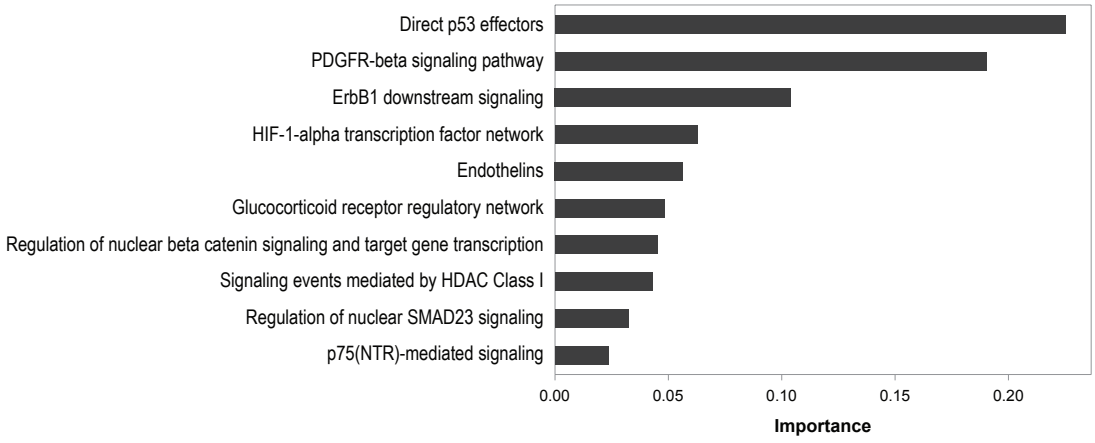


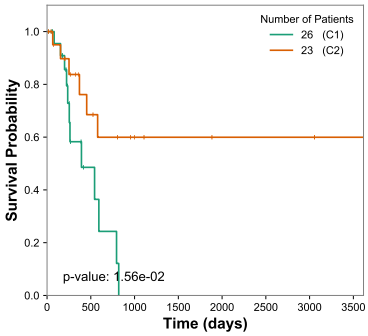
Figure 4.12 The top 10 pathways, which have the highest relative importance in clustering for KIRC patients. One pathway weight is calculated by summing the kernel weights which are calculated using that specific pathway and different omics.

The top relevant pathway in clustering the patients into subgroups emerges as the *the direct p53 effectors pathway* (Figure 4.12). p53 is a tumor suppressor transcription factor that regulates the cell division to prevent uncontrolled growth of cells (Vogelstein et al., 2000). Similarly, the second and the third pathways, which are *PDGFR-beta signaling* and *ErbB1 downstream signaling pathway* (better known as EGFR), respectively, are critical signaling pathways for cancer (Smith et al., 2005). The fourth pathway is related to hypoxia-inducible factors (HIFs), which regulate the expressions of many genes that are related to tumorigenesis (Banumathy and Cairns, 2010). Additionally, C. Shen et al., 2011 shows that HIF1α is a target of 14q loss, which is commonly associated with poor prognosis in kidney cancer. The fifth pathway is the Endothelin pathway and earlier results report that Endothelin-1 promotes cell survival in renal cell carcinoma (Pflug et al., 2007).

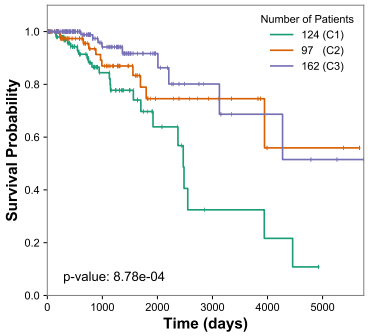
4.8 Application to Other Cancers

We apply the PAMOGK framework to other cancer types. Out of 12 cancer types from TCGA PanCancer study, we exclude 3 types of cancer: Acute myeloid leukemia (LAML) due to lack of primary tumors, rectum adenocarcinoma (READ) since there is only one patient with all the three omic data type available and colon adenocarcinoma (COAD) since none of the patients have passed away(all data is censored). The remaining cancer types include bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC). We apply the PAMOGK framework with SmSPK graph kernel and MKKM-MR as the multi-view kernel clustering algorithm. To compare, we also apply the RBF kernel and MKKM-MR combination.

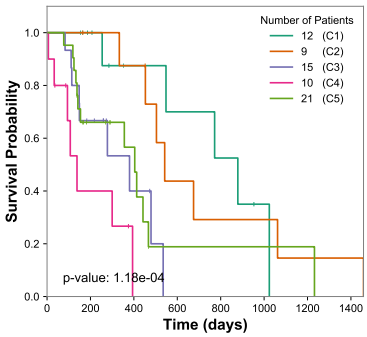
Figure 4.13 shows the KM plots and and the log-rank test results' p-values of the clusters obtained for the different cancer types. We observe that the clusters are well separated in terms of the patient survival distribution (Log-rank test, $p < 0.05$). However, the p-values are not as small as KIRC in these cancers. The log-rank test is known to be inaccurate when sample size small or unbalanced (Latta, 1981). Although BLCA clusters are well separated in terms of survival (Figure 4.13a), the p-value of the log-rank is not that small p-value($1.56e$-02). This could be attributed to the small number of patients in the clusters (26 and 23). Another problem of log-rank test occurs when the number of censored patients are small or unbalanced across patient groups(Latta, 1981). In the case of OV clusters (Figure 4.13g), pa-



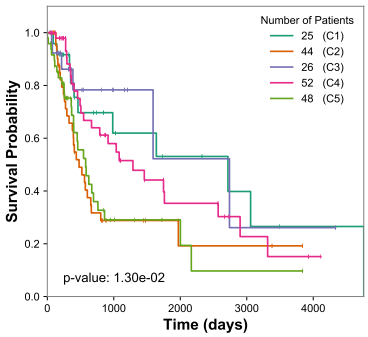
(a) BLCA (k=2 α=0.9)



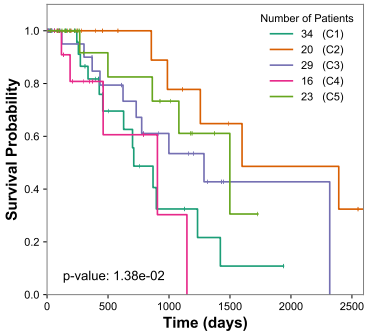
(b) BRCA (k=3 α=0.3)



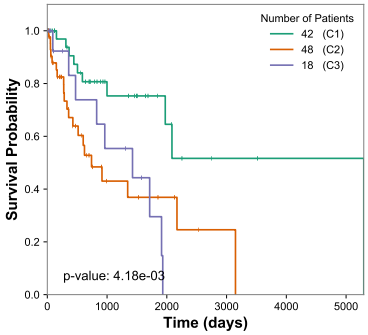
(c) GBM (k=5 α=0.01)



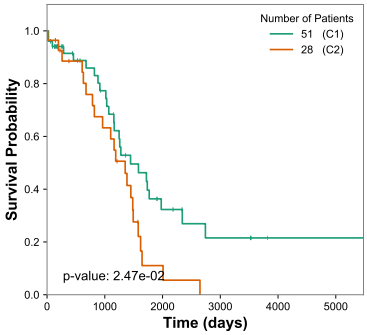
(d) HNSC (k=5 α=0.5)



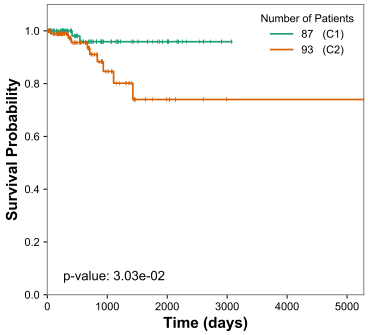
(e) LUAD (k=5, α=0.9).



(f) LUSC (k=3 α=0.01)



(g) OV (k=2 α=0)



(h) UCEC (k=2 α=0.8)

Figure 4.13 Kaplan-Meir plots for best clustering solution for each cancer type. The number of clusters (k) and the smoothing parameter value (α) that leads to these results are provided under each subplot. The log-rank test p-values are shown in the KM curves.

tients are well separated; however, 78.57%(22 over 28) of the patients in the second cluster is censored. This could be the reason why the log rank test yields a poor p-value(2.47e-02). Similarly, almost 94% of the UCEC patients are censored. That results in two clusters with high numbers of censored patients(85 out of 87 and 84 out of 93). Although the separation is evident in the Kaplan-Meir curves of the two clusters, the log-rank test p-value(3.03e-02) does not reflect this strong separation.

One other observation is that some of the clusters of HNSC overlaps (Figure 4.13d) indicating that a smaller number of clusters could be alternative solution. Thus, we examine the clusters with $k = 3$ and $k = 4$. Similar to $k = 5$, HNSC clusters are well separated in terms of patient survival for both $k = 3$ (p-value $= 3.15e$-02) and $k = 4$ (p-value $= 1.92e$-02).

When we replace SmSPK with RBF and compare the results, we observe that the clusters of BLCA, BRCA, GBM, LUAD, LUSC, OV, UCEC from PAMOGK is separated more successfully than the clusters from RBF in terms of survival(Figure 4.14).

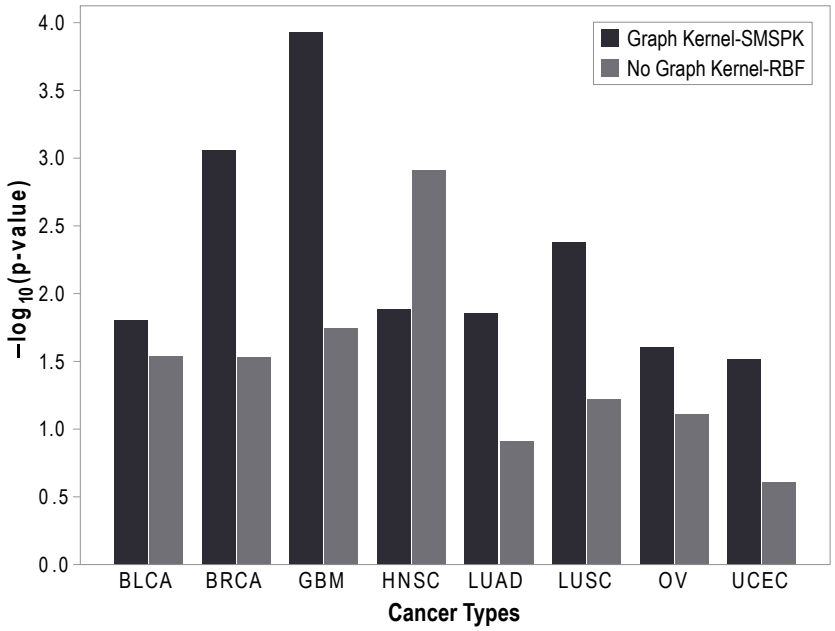


Figure 4.14 The log-rank test p-values obtained on different cancers when two clustering methods with two different kernel choice is applied: PAMOGK with SmSPK kernel using pathway graphs and multi-view clustering with RBF kernel without pathway information.

Although the clusters are well separated in terms of survival distribution, other clinical paremeters do not differ across these clusters Table 4.6. The ones that differ are as follows: Neoplasm lymph node stage differs (p-value $= 3.31e$-03) across BRCA patient subgroups at a statistically significant level(Table A.6). Similarly for HNSC patients subgroups, neoplasm histologic grade(p-value $= 7.70e$-04) (Table A.7), clin-

Table 4.6 Statistical analysis of clinical parameters of other cancer types.

Cancer	Clinical Parameter	Test	p-value
BLCA	**Age**	One-way ANOVA	9.221e-01
	Gender	χ^2	6.051e-01
	Diagnosis subtype	χ^2	8.905e-01
	Tumor Stage	χ^2	5.872e-01
	Metastasis stage	χ^2	5.592e-01
	Neoplasm disease stage	χ^2	6.454e-01
	Neoplasm lymph node stage	χ^2	1.541e-01
BRCA	**Age**	One-way ANOVA	4.864e-01
	Gender	χ^2	4.170e-01
	Tumor Stage	χ^2	3.448e-01
	Metastasis stage	χ^2	7.554e-01
	Neoplasm disease stage	χ^2	2.597e-01
	Neoplasm lymph node stage	χ^2	**3.310e-03**
GBM	**Age**	One-way ANOVA	7.204e-02
	Gender	χ^2	6.632e-01
HNSC	**Age**	One-way ANOVA	**2.000e-05**
	Gender	χ^2	**4.645e-02**
	Tumor Stage	χ^2	3.596e-01
	Distant metastasis pathologic spread	χ^2	5.318e-01
	Neoplasm histologic grade	χ^2	**7.700e-04**
	Primary tumor n stage	χ^2	3.180e-01
	Primary tumor m stage	χ^2	2.277e-01
	Primary tumor t stage	χ^2	**4.351e-02**
	Primary tumor pathologic spread	χ^2	6.025e-02
	Lymph node pathologic spread	χ^2	5.567e-01
	Clinical group stage	χ^2	**2.606e-02**
LUAD	**Age**	One-way ANOVA	6.158e-01
	Gender	χ^2	2.565e-01
	Tumor Stage	χ^2	1.357e-01
	Distant metastasis pathologic spread	χ^2	3.661e-01
	Primary tumor pathologic spread	χ^2	2.773e-01
	Lymph node pathologic spread	χ^2	9.164e-02
LUSC	**Age**	One-way ANOVA	4.453e-01
	Gender	χ^2	5.871e-02
	Tumor Stage	χ^2	8.900e-01
	Distant metastasis pathologic spread	χ^2	5.208e-01
	Primary tumor pathologic spread	χ^2	5.348e-01
	Lymph node pathologic spread	χ^2	4.006e-01
OV	**Age**	One-way ANOVA	4.838e-01
	Gender	χ^2	1.000e-00
	Tumor Stage	χ^2	1.431e-01
	Neoplasm histologic grade	χ^2	3.739e-01
UCEC	**Age**	One-way ANOVA	7.069e-01
	Gender	χ^2	1.000e-00
	Tumor Stage	χ^2	5.000e-01

ical group stage(p-value $= 2.61e$-02)(Table A.8), and clinical primary tumor stage(p-value $= 4.35e$-02) (Table A.9) distributed significantly differently. However, for these parameters, the patients with low degree or grade parameters does not necessarily

belong to best prognosis cluster. Similarly, patients with high degree parameters are not distributed highly into the worst prognosis group. Thus, the analysis of tumor related-clinical parameters for cancer subgroups other than KIRC is inconclusive. In these cancer types, it could be the other omic data that could be reporter of the clusters.

4.8.1 Influential pathways for other cancer types

Similar to the pathway analysis with KIRC patient, we also analyze influential pathways of clustering application on other cancer patients. We examine the important pathways that resulted in the best clustering in terms of survival. Since our model assigns nearly the same weight to each pathway, we couldn't analyze the pathways for BLCA and LUSC. For the rest of the cancer types, the most important 10 pathways are shown in Figure 4.15-4.20. The *direct p53 effectors pathway* which we found as important in the KIRC clustering always ranks as the first or the second important pathways for all cancer types. p53 is a tumor suppressor transcription factor that regulates the cell division to prevent uncontrolled growth of cells (Vogelstein et al., 2000) and it is important for all cancer types. Similarly, *PDGFR-beta signaling* pathway is always among the top two pathways except for the LUAD patients where it is still in the top 10 most important pathways. Another pathway that is important for KIRC clusters is *ErbB1 downstream signaling pathway* (better known as EGFR) and it is always among the top 10 most important pathways for all cancer types. These are critical pathways for various cancer types (Smith et al., 2005). P73 is one of the tumor suppressors of the p53 transcription factors family. It is observed in many studies that p73 is overexpressed in many different types of cancers (DeYoung and Ellisen, 2007). Parallel to these studies, PAMOGK finds *p73 transcription factor network* pathway is found as important for BRCA, OV, LUAD ve UCEC cancer types. $p75^{\mathrm{NTR}}$ is a nerve growth factor receptor and its relation with other cancer types, especially BRCA, has been shown in many studies(Molloy et al., 2011). Although there is no conclusive connection for OV, GBM, HNSC, UCEC; *$p75^{NTR}$-mediated signaling* pathway is found among important pathways. The effect of Histone deacetylases(HDACs), which regulates the activation of many proteins, on cancer is analyzed and studied before (Glozak and Seto, 2007). Especially the HDAC1 group is found as effective on the kidney, ovarian, breast, and colorectal cancers, and PAMOGK also finds *Signaling events mediated by HDAC Class 1* pathway as important for these cancer types. Lastly, the *Regulation of nuclear beta catenin signaling and target gene transcription* pathway which is among the top 10 influential pathways for BRCA, KIRC, GBM, HNSC, LUAD, UCEC is important for many cancer types (Shang et al., 2017).

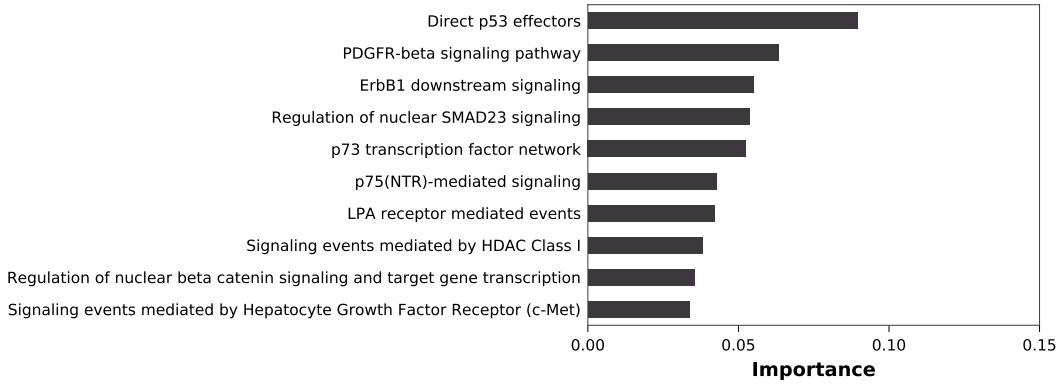


Figure 4.15 The top 10 pathways, which have the highest relative importance in clustering BRCA.

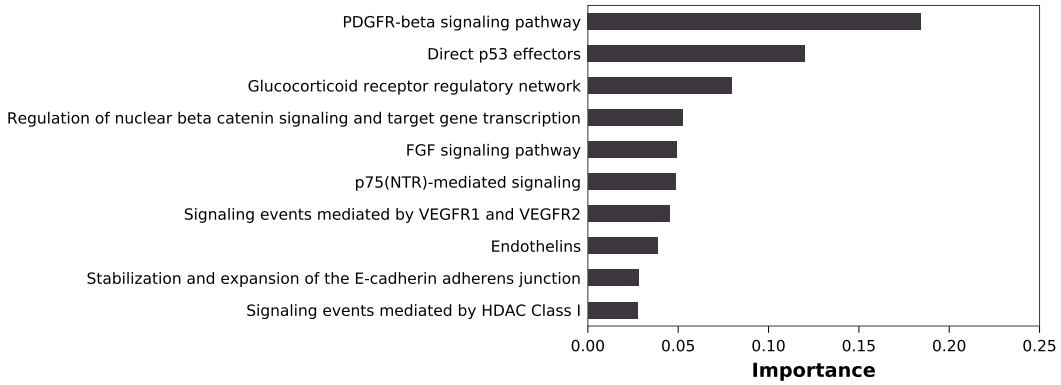


Figure 4.16 The top 10 pathways, which have the highest relative importance in clustering GBM.

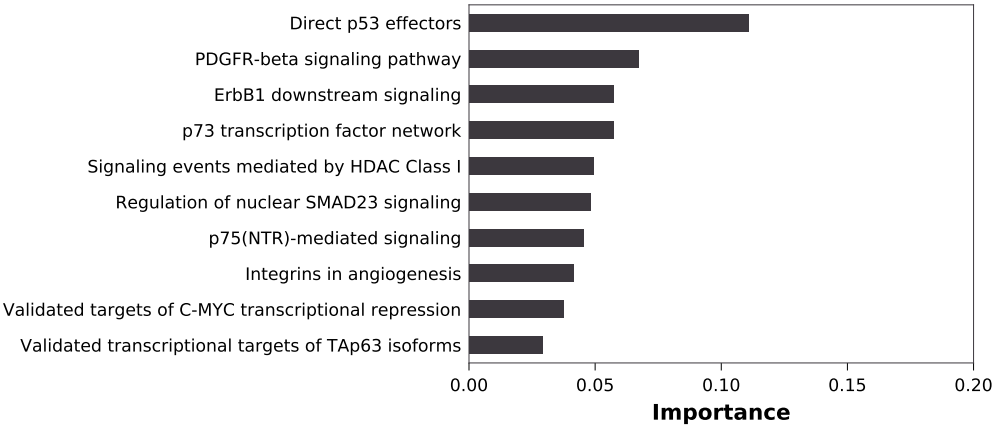


Figure 4.17 The top 10 pathways, which have the highest relative importance in clustering OV.

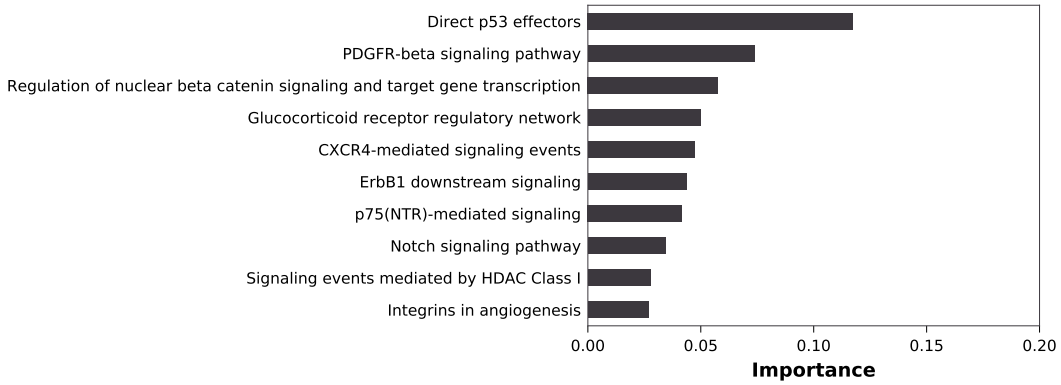


Figure 4.18 The top 10 pathways, which have the highest relative importance in clustering HNSC.

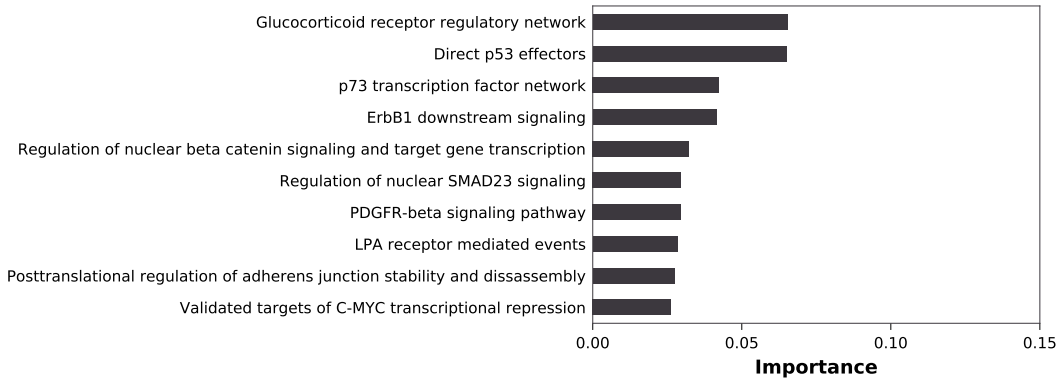


Figure 4.19 The top 10 pathways, which have the highest relative importance in clustering LUAD.

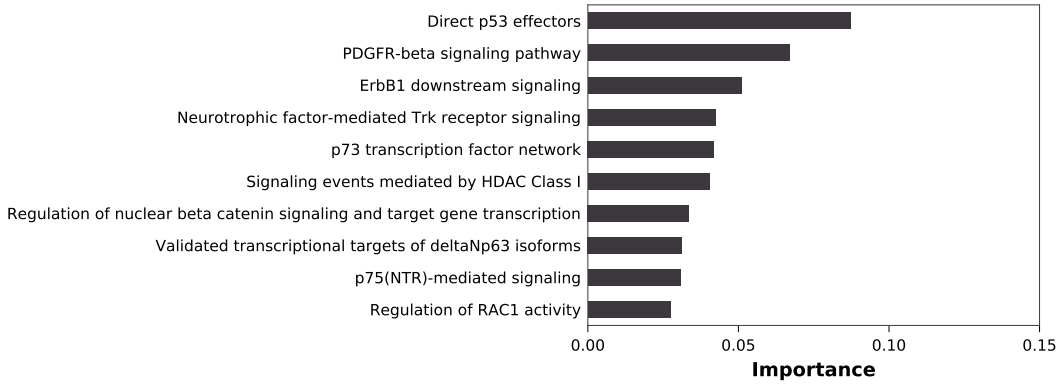


Figure 4.20 The top 10 pathways, which have the highest relative importance in clustering UCEC.

Chapter 5

CONCLUSION

The heterogeneity of cancer due to genetic and non-genetic factors causes variations in cancer cells within a cancer type. The advances in next-generation sequencing technologies and other high-throughput assays, allow characterizing the genome, proteome, and the transcriptome for large cohorts of patients. This multi-omics characterization brings up possibilities to refine these subtypes on a molecular level. These subtypes allow us to design better treatment strategies, make a more accurate diagnosis, and gain insight into different molecular mechanisms.

While different omic types such as mutations, gene expressions, and protein expressions allow us to analyze patients from different data views, another view is provided with the known molecular interactions among proteins. Integrating networks also help overcome the problem that few alterations are shared among cancer patients.

We present PAMOGK for discovering subgroups of patients, which not only operates by integrating different omics data sets derived from patients but also incorporates existing knowledge on biological pathways. To corroborate these data sources, we use a novel graph kernel, SmSPK, that evaluates patient similarities based on their molecular alterations in known pathways. We employ a multi-view kernel clustering technique to leverage views constructed by different molecular alteration types and pathways. The proposed methodology also provides quantitative evidence for the decisive role of known driver pathways on the clustering process.

We evaluated the different aspects of the framework and the choices we made within the framework. First of all, our evaluation of the available kernels that could be used in PAMOGK showed that the proposed graph kernel SmSPK performs better than other state-of-the-art graph kernels in this task and also outperforms the RBF kernels. Secondly, we observe that among many multi-view kernel approaches we use to integrate the multiple graph kernels, MKKM-MR outperforms different multi-

view kernel clustering algorithms. Thirdly, we evaluate whether specific steps in such as smoothing of the labels on the graph is necessary. We observe that smoothing helps to find similarities between patients that do not have a commonly altered gene. We conclude that especially for sparse data types such as mutation, smoothing helps fuse information on the pathway. We also evaluated several techniques for representing dysregulation-related expression patterns on the pathway graphs as node labels.

When applied to KIRC, PAMOGK results in patient clusters that differ significantly in their survival distributions and other clinical parameters. We also show that PAMOGK performs better compared to the state-of-the-art multi-omics approaches.

After all evaluation of PAMOGK on KIRC patients, we also apply the framework to 8 other available cancer types. We find significantly different patient subgroups in terms of survival for other cancer types. Moreover, we compare the log-rank test p-values attained by clustering with PAMOGK to p-values attained by clustering using RBF kernels for each omic in a multi-view kernel type separately. In 7 out of 8 cancer types, PAMOGK gives more successful results. However, the clinical attributes are not distributed differently among the clusters except for BRCA and HNSC. On the other hand, the critical pathways for these cancer types are similar to those we found with KIRC. Most of these pathways are crucial for cancer-related processes.

Some of the earlier work of this thesis has been presented before in the thesis Unal (2019). This thesis, as a contribution, updates the representation of the expression data type on graph with a continuous labeling techniques and different techniques are evaluated during the update process. Since pathway data of earlier work (Unal, 2019) has some problems, pathway data is updated and another database is used. Additionally, each step of the framework has been evaluated separately and alternative strategies are compared to justify the decisions. While earlier work only compares the proposed graph kernel SmSPK with radial basis function kernel, this work also use well-known graph kernel methods in comparison. It also expands the set of compared multi-view kernels with methods that have high-performance in literature. Moreover, as a new evaluation, the framework PAMOGK is evaluated with different state-of-the-art multi-omics methods for cancer subtyping. The smoothing effect, labeling techniques, clinical parameter distributions among clusters are analyzed and discussed which was not done by Unal (2019). Also code environment is moved to python with the collaboration of authors of the paper of this thesis. Finally, the performance in terms of well-separated patient clusters is increased with these updates.

One limitation of the current work is that we use the bulk expression results provided by the TCGA project. However, it is known that there could be a high level of intra-tumor heterogeneity (Dagogo-Jack and Shaw, 2018), and the bulk tumor might include a diverse collection of cells harboring distinct molecular signatures. Future work would be to adapt PAMOGK framework to single-cell measurements as they become available for large cohorts of patients.

One of the most important aspects of the framework PAMOGK is that it is applicable in different fields and not specific to cancer subtyping. PAMOGK can easily be adapted to a clustering problem with given knowledge graphs and datasets that includes feature vectors of samples to label these graphs.

The work can be extended in several directions. In this current work, we use omic datasets containing somatic mutations, gene, and protein expression as they give more direct information on the alterations in the pathways. In the future work, one can map the copy number variations and methylation levels as well to the genes and use them as additional views.

The proposed kernel matrix characterizes the similarities of patients based on the shortest path on the graphs. Other graph kernels can be devised to capture patient similarities on the graphs using other topological features of the graphs.

Furthermore, in the present study, we ignore the direction and label or type of edges in the pathways. A kernel that explicitly accounts for edge directions can be more devised. Also, edge types of interactions could lead to a more expressive representation of the molecular interactions, which we shall investigate in the future. In place or addition to the pathways, protein-protein interaction networks, or super pathways where pathway graphs are combined into one graph can be used. However, it is essential to note that when we do not use separate pathways, we cannot extract the importance associated with the clustering to each pathway.

Lastly, the multi-view kernel clustering algorithms in the literature are not mostly designed for a small number of kernels. For our problem, 3 data types and 165 pathways form a large number of kernels, and the number is likely to increase as these pathways are refined, and more data types are integrated into our framework. Thus, these multi-view kernel clustering methods might not suffice when the number of pathways and the data types are increased. The current multi-view kernel methods usually assign zero-weight to sparse kernels. A new multi-view kernel clustering method that can combine many kernels without losing the signals in sparse kernels can be studied as future work.

BIBLIOGRAPHY

Aiolli, Fabio and Michele Donini (Dec. 2015). "EasyMKL: a scalable multiple kernel learning algorithm". In: *Neurocomputing* 169, pp. 215–224. DOI: 10.1016/j.neucom.2014.11.078. URL: https://doi.org/10.1016/j.neucom.2014.11.078.

Amin, Mahul B. et al. (Jan. 2017). "The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging". In: *CA: A Cancer Journal for Clinicians* 67.2, pp. 93–99. DOI: 10.3322/caac.21388. URL: https://doi.org/10.3322/caac.21388.

Banumathy, Gowrishankar and Paul Cairns (Oct. 2010). "Signaling pathways in renal cell carcinoma". In: *Cancer Biology & Therapy* 10.7, pp. 658–664. DOI: 10.4161/cbt.10.7.13247. URL: https://doi.org/10.4161/cbt.10.7.13247.

Bertucci, François et al. (Mar. 2005). "Gene Expression Profiling Identifies Molecular Subtypes of Inflammatory Breast Cancer". In: *Cancer Research* 65.6, pp. 2170–2178. DOI: 10.1158/0008-5472.can-04-4115. URL: https://doi.org/10.1158/0008-5472.can-04-4115.

Borgwardt, Karsten M and Hans-Peter Kriegel (2005). "Shortest-path kernels on graphs". In: *Data Mining, Fifth IEEE International Conference on*. IEEE, 8–pp.

Brunet, J.-P., P. Tamayo, T. R. Golub, and J. P. Mesirov (Mar. 2004). "Metagenes and molecular pattern discovery using matrix factorization". In: *Proceedings of the National Academy of Sciences* 101.12, pp. 4164–4169. DOI: 10.1073/pnas.0308531101. URL: https://doi.org/10.1073/pnas.0308531101.

Cowen, Lenore, Trey Ideker, Benjamin J Raphael, and Roded Sharan (2017). "Network propagation: a universal amplifier of genetic associations". In: *Nature Reviews Genetics* 18.9, p. 551.

Creighton, Chad J. et al. (July 2013). "Comprehensive molecular characterization of clear cell renal cell carcinoma". In: *Nature* 499.7456, pp. 43–49. ISSN: 1476-4687. DOI: 10.1038/nature12222. URL: https://doi.org/10.1038/nature12222.

Curtis, Christina et al. (2012). "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups". In: *Nature* 486.7403, p. 346.

Dagogo-Jack, Ibiayi and Alice T Shaw (2018). "Tumour heterogeneity and resistance to cancer therapies". In: *Nature reviews Clinical oncology* 15.2, p. 81.

DeYoung, M P and L W Ellisen (Mar. 2007). "p63 and p73 in human cancer: defining the network". In: *Oncogene* 26.36, pp. 5169–5183. DOI: 10.1038/sj.onc.1210337. URL: https://doi.org/10.1038/sj.onc.1210337.

Eason, Katherine, Gift Nyamundanda, and Anguraj Sadanandam (May 2018). "polyClustR: defining communities of reconciled cancer subtypes with biological and prognostic significance". In: *BMC Bioinformatics* 19.1. DOI: 10.1186/s12859-018-2204-4. URL: https://doi.org/10.1186/s12859-018-2204-4.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (Dec. 1998). "Cluster analysis and display of genome-wide expression patterns". In: *Proceedings of the National Academy of Sciences* 95.25, pp. 14863–14868. DOI: 10.1073/pnas.95.25.14863. URL: https://doi.org/10.1073/pnas.95.25.14863.

Feragen, Aasa, Niklas Kasenburg, Jens Petersen, Marleen de Bruijne, and Karsten Borgwardt (2013). "Scalable kernels for graphs with continuous attributes". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Curran Associates, Inc., pp. 216–224. URL: http://papers.nips.cc/paper/5155-scalable-kernels-for-graphs-with-continuous-attributes.pdf.

Frigyesi, Attila and Mattias Höglund (Jan. 2008). "Non-Negative Matrix Factorization for the Analysis of Complex Gene Expression Data: Identification of Clinically Relevant Tumor Subtypes". In: *Cancer Informatics* 6, CIN.S606. DOI: 10.4137/cin.s606. URL: https://doi.org/10.4137/cin.s606.

Gabasova, Evelina, John Reid, and Lorenz Wernisch (2017). "Clusternomics: Integrative context-dependent clustering for heterogeneous datasets". In: *PLoS computational biology* 13.10, e1005781.

Gan, Yanglan, Ning Li, Guobing Zou, Yongchang Xin, and Jihong Guan (Dec. 2018). "Identification of cancer subtypes from single-cell RNA-seq data using a consensus clustering method". In: *BMC Medical Genomics* 11.S6. DOI: 10.1186/s12920-018-0433-z. URL: https://doi.org/10.1186/s12920-018-0433-z.

Glozak, M A and E Seto (Aug. 2007). "Histone deacetylases and cancer". In: *Oncogene* 26.37, pp. 5420–5432. DOI: 10.1038/sj.onc.1210610. URL: https://doi.org/10.1038/sj.onc.1210610.

Gönen, Mehmet and Adam A. Margolin (2014). "Localized Data Fusion for Kernel K-means Clustering with Application to Cancer Biology". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'14. Montreal, Canada: MIT Press, pp. 1305–1313. URL: http://dl.acm.org/citation.cfm?id=2968826.2968972.

Handhayani, Teny and Lely Hiryanto (2015). "Intelligent Kernel K-Means for Clustering Gene Expression". In: *Procedia Computer Science* 59. International Conference on Computer Science and Computational Intelligence (ICCSCI 2015), pp. 171–177. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.07.544. URL: http://www.sciencedirect.com/science/article/pii/S1877050915020736.

Harrington, David P. and Thomas R. Fleming (1982). "A class of rank test procedures for censored survival data". In: *Biometrika* 69.3, pp. 553–566. DOI: 10.1093/biomet/69.3.553. URL: https://doi.org/10.1093/biomet/69.3.553.

Hayes, D Neil et al. (2006). "Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts". In: *Journal of Clinical Oncology* 24.31, pp. 5079–5090.

Hotelling, Harold (Dec. 1936). "Relations Between Two Sets Of Variates". In: *Biometrika* 28.3-4, pp. 321–377. ISSN: 0006-3444. DOI: 10.1093/biomet/28.3-4.321. eprint: https://academic.oup.com/biomet/article-pdf/28/3-4/321/586830/28-3-4-321.pdf. URL: https://doi.org/10.1093/biomet/28.3-4.321.

Jain, Anil K. and Richard C. Dubes (1988). *Algorithms for Clustering Data*. USA: Prentice-Hall, Inc. ISBN: 013022278X.

Jiang, Limin, Yongkang Xiao, Yijie Ding, Jijun Tang, and Fei Guo (Feb. 2019). "Discovering Cancer Subtypes via an Accurate Fusion Strategy on Multiple Pro-

file Data". In: *Frontiers in Genetics* 10. DOI: 10.3389/fgene.2019.00020. URL: https://doi.org/10.3389/fgene.2019.00020.

John, Christopher R, David Watson, Michael R Barnes, Costantino Pitzalis, and Myles J Lewis (Sept. 2019). "Spectrum: fast density-aware spectral clustering for single and multi-omic data". In: *Bioinformatics*. Ed. by Lenore Cowen. DOI: 10.1093/bioinformatics/btz704. URL: https://doi.org/10.1093/bioinformatics/btz704.

Kannan, S. R., R. Devi, S. Ramathilagam, and T. P Hong (May 2016). "Effective fuzzy possibilistic c-means: an analyzing cancer medical database". In: *Soft Computing* 21.11, pp. 2835–2845. DOI: 10.1007/s00500-016-2198-7. URL: https://doi.org/10.1007/s00500-016-2198-7.

Kaplan, E. L. and Paul Meier (June 1958). "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282, pp. 457–481. DOI: 10.1080/01621459.1958.10501452. URL: https://doi.org/10.1080/01621459.1958.10501452.

Kaufman, Leonard and Peter J. Rousseeuw, eds. (Mar. 1990). *Finding Groups in Data*. John Wiley & Sons, Inc. DOI: 10.1002/9780470316801. URL: https://doi.org/10.1002/9780470316801.

Lapointe, Jacques et al. (2004). "Gene expression profiling identifies clinically relevant subtypes of prostate cancer". In: *Proceedings of the National Academy of Sciences* 101.3, pp. 811–816. ISSN: 0027-8424. DOI: 10.1073/pnas.0304146101. eprint: https://www.pnas.org/content/101/3/811.full.pdf. URL: https://www.pnas.org/content/101/3/811.

Latta, Robert B (1981). "A Monte Carlo study of some two-sample rank tests with censored data". In: *Journal of the American Statistical Association* 76.375, pp. 713–719.

Lee, Daniel D. and H. Sebastian Seung (Oct. 1999). "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755, pp. 788–791. DOI: 10.1038/44565. URL: https://doi.org/10.1038/44565.

Liang, Ruqing et al. (July 2018). "A comprehensive analysis of prognosis prediction models based on pathway-level, gene-level and clinical information for glioblastoma". In: *International Journal of Molecular Medicine*. DOI: 10.3892/ijmm.2018.3765. URL: https://doi.org/10.3892/ijmm.2018.3765.

Liu, Jialu, Chi Wang, Jing Gao, and Jiawei Han (May 2013). "Multi-View Clustering via Joint Nonnegative Matrix Factorization". In: *Proceedings of the 2013 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611972832.28. URL: https://doi.org/10.1137/1.9781611972832.28.

Liu, Jianfang and et al. (Apr. 2018). "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics". In: *Cell* 173.2, 400–416.e11. DOI: 10.1016/j.cell.2018.02.052. URL: https://doi.org/10.1016/j.cell.2018.02.052.

Liu, Xinwang, Yong Dou, Jianping Yin, Lei Wang, and En Zhu (2016). "Multiple Kernel K-means Clustering with Matrix-induced Regularization". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI'16. Phoenix, Arizona: AAAI Press, pp. 1888–1894. URL: http://dl.acm.org/citation.cfm?id=3016100.3016163.

Liu, Xinwang, Sihang Zhou, et al. (2017). "Optimal Neighborhood Kernel Clustering with Multiple Kernels". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, pp. 2266–2272. URL: http://dl.acm.org/citation.cfm?id=3298483.3298566.

Lloyd, S. (Mar. 1982). "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137. DOI: 10.1109/tit.1982.1056489. URL: https://doi.org/10.1109/tit.1982.1056489.

Ma, Xiaohua et al. (May 2019). "Identification of a molecular subtyping system associated with the prognosis of Asian hepatocellular carcinoma patients receiving liver resection". In: *Scientific Reports* 9.1. DOI: 10.1038/s41598-019-43548-1. URL: https://doi.org/10.1038/s41598-019-43548-1.

Manica, Matteo, Joris Cadow, Roland Mathis, and Marıa Rodrıguez Martınez (2019). "PIMKL: Pathway-Induced Multiple Kernel Learning". In: *npj Systems Biology and Applications* 5.1, pp. 1–8.

Mo, Qianxing, Ronglai Shen, et al. (May 2017). "A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data". In: *Biostatistics* 19.1, pp. 71–86. DOI: 10.1093/biostatistics/kxx017. URL: https://doi.org/10.1093/biostatistics/kxx017.

Mo, Qianxing, Sijian Wang, et al. (Feb. 2013). "Pattern discovery and cancer gene identification in integrated cancer genomic data". In: *Proceedings of the National Academy of Sciences* 110.11, pp. 4245–4250. DOI: 10.1073/pnas.1208949110. URL: https://doi.org/10.1073/pnas.1208949110.

Molloy, Niamh, Danielle Read, and Adrienne Gorman (Feb. 2011). "Nerve Growth Factor in Cancer Cell Death and Survival". In: *Cancers* 3.1, pp. 510–530. DOI: 10.3390/cancers3010510. URL: https://doi.org/10.3390/cancers3010510.

Monti, Stefano, Pablo Tamayo, Jill Mesirov, and Todd Golub (2003a). "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data". In: *Machine learning* 52.1-2, pp. 91–118.

— (2003b). "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data". In: *Machine learning* 52.1-2, pp. 91–118.

Müller, Mike F, Ashraf EK Ibrahim, and Mark J Arends (2016). "Molecular pathological classification of colorectal cancer". In: *Virchows Archiv* 469.2, pp. 125–134.

Neumann, Marion, Roman Garnett, Christian Bauckhage, and Kristian Kersting (Feb. 2016). "Propagation Kernels: Efficient Graph Kernels from Propagated Information". In: *Mach. Learn.* 102.2, pp. 209–245. ISSN: 0885-6125. DOI: 10.1007/s10994-015-5517-9. URL: http://dx.doi.org/10.1007/s10994-015-5517-9.

Nguyen, Tin, Rebecca Tagett, Diana Diaz, and Sorin Draghici (2017). "A novel approach for data integration and disease subtyping". In: *Genome research* 27.12, pp. 2025–2039.

Nidheesh, N., K.A. Abdul Nazeer, and P.M. Ameer (Dec. 2017). "An Enhanced Deterministic K-Means Clustering Algorithm for Cancer Subtype Prediction from Gene Expression Data". In: *Comput. Biol. Med.* 91.C, pp. 213–221. ISSN: 0010-4825.

Pflug, Beth R et al. (2007). "Endothelin-1 promotes cell survival in renal cell carcinoma through the ETA receptor". In: *Cancer letters* 246.1-2, pp. 139–148.

Prasad, Vinay, Tito Fojo, and Michael Brada (2016). "Precision oncology: origins, optimism, and potential". In: *The Lancet Oncology* 17.2, e81–e86.

Rappoport, Nimrod and Ron Shamir (2018a). "Multi-omic and multi-view clustering algorithms: review and cancer benchmark". In: *Nucleic acids research* 46.20, pp. 10546–10562.

— (Oct. 2018b). "Multi-omic and multi-view clustering algorithms: review and cancer benchmark". In: *Nucleic Acids Research* 46.20, pp. 10546–10562. ISSN: 0305-1048. DOI: 10.1093/nar/gky889. eprint: http://oup.prod.sis.lan/nar/article-pdf/46/20/10546/26817354/gky889.pdf. URL: https://doi.org/10.1093/nar/gky889.

Ren, Zhonglu, Wenhui Wang, and Jinming Li (Nov. 2015). "Identifying molecular subtypes in human colon cancer using gene expression and DNA methylation microarray data". In: *International Journal of Oncology* 48.2, pp. 690–702. DOI: 10.3892/ijo.2015.3263. URL: https://doi.org/10.3892/ijo.2015.3263.

Ricketts, Christopher J. and et al. (Apr. 2018). "The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma". In: *Cell Reports* 23.1, 313–326.e5. DOI: 10.1016/j.celrep.2018.03.075. URL: https://doi.org/10.1016/j.celrep.2018.03.075.

Rousseeuw, Peter J. (Nov. 1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7. URL: https://doi.org/10.1016/0377-0427(87)90125-7.

Schaefer, Carl F. et al. (Oct. 2008). "PID: the Pathway Interaction Database". In: *Nucleic Acids Research* 37.suppl_1, pp. D674–D679. DOI: 10.1093/nar/gkn653. URL: https://doi.org/10.1093/nar/gkn653.

Schölkopf, Bernhard, Alexander Smola, and Klaus-Robert Müller (July 1998). "Nonlinear Component Analysis as a Kernel Eigenvalue Problem". In: *Neural Computation* 10.5, pp. 1299–1319. DOI: 10.1162/089976698300017467. URL: https://doi.org/10.1162/089976698300017467.

Sejdinovic, Dino, Arthur Gretton, and Wicher Bergsma (2013). *A Kernel Test for Three-Variable Interactions*. arXiv: 1306.2281 [stat.ME].

Shang, Shuang, Fang Hua, and Zhuo-Wei Hu (Feb. 2017). "The regulation of Beta-catenin activity and function in cancer: therapeutic opportunities". In: *Oncotarget* 8.20, pp. 33972–33989. DOI: 10.18632/oncotarget.15687. URL: https://doi.org/10.18632/oncotarget.15687.

Shen, C. et al. (June 2011). "Genetic and Functional Studies Implicate HIF1 as a 14q Kidney Cancer Suppressor Gene". In: *Cancer Discovery* 1.3, pp. 222–235. DOI: 10.1158/2159-8290.cd-11-0098. URL: https://doi.org/10.1158/2159-8290.cd-11-0098.

Shen, Ronglai, Adam B. Olshen, and Marc Ladanyi (Sept. 2009). "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis". In: *Bioinformatics* 25.22, pp. 2906–2912. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp543. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/25/22/2906/16891356/btp543.pdf. URL: https://doi.org/10.1093/bioinformatics/btp543.

Shervashidze, Nino, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt (2011). "Weisfeiler-lehman graph kernels". In: *Journal of Machine Learning Research* 12.Sep, pp. 2539–2561.

Shi, Mingguang and Guofu Xu (July 2017). "Spectral clustering using Nyström approximation for the accurate identification of cancer molecular subtypes". In: *Scientific Reports* 7.1. DOI: 10.1038/s41598-017-05275-3. URL: https://doi.org/10.1038/s41598-017-05275-3.

Siglidis, Giannis et al. (2018). "GraKeL: A Graph Kernel Library in Python". In: *arXiv preprint arXiv:1806.02193*.

Smith, Karlene et al. (June 2005). "Silencing of Epidermal Growth Factor Receptor Suppresses Hypoxia-Inducible Factor-2–DrivenVHL-/-Renal Cancer". In: *Cancer Research* 65.12, pp. 5221–5230. DOI: 10.1158/0008-5472.can-05-0169. URL: https://doi.org/10.1158/0008-5472.can-05-0169.

Sotiriou, Christos et al. (2003). "Breast cancer classification and prognosis based on gene expression profiles from a population-based study". In: *Proceedings of the National Academy of Sciences* 100.18, pp. 10393–10398. ISSN: 0027-8424. DOI: 10.1073/pnas.1732912100. eprint: https://www.pnas.org/content/100/18/10393.full.pdf. URL: https://www.pnas.org/content/100/18/10393.

Speicher, Nora K and Nico Pfeifer (2015). "Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery". In: *Bioinformatics* 31.12, pp. i268–i275.

Therneau, Terry M and Patricia M Grambsch (2000). "The Cox model". In: *Modeling survival data: extending the Cox model*. Springer, pp. 39–77.

Togninalli, Matteo, Elisabetta Ghisu, Felipe Llinares-López, Bastian Rieck, and Karsten Borgwardt (2019a). "Wasserstein Weisfeiler–Lehman Graph Kernels". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc. arXiv: 1906.01277 [cs.LG]. Forthcoming.

— (2019b). "Wasserstein Weisfeiler-Lehman Graph Kernels". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pp. 6439–6449. URL: http://papers.nips.cc/paper/8872-wasserstein-weisfeiler-lehman-graph-kernels.pdf.

Toss, Angela and Massimo Cristofanilli (2015). "Molecular characterization and targeted therapeutic approaches in breast cancer". In: *Breast Cancer Research* 17.1, p. 60.

Unal, Ali Burak (July 2019). "Identification of cancer patient subgroups via pathway based multi-view graph kernel clustering". Master's thesis. Bilkent University.

Vandin, Fabio, Alexandra Papoutsaki, Benjamin J. Raphael, and Eli Upfal (May 2015). "Accurate Computation of Survival Statistics in Genome-Wide Studies". In: *PLOS Computational Biology* 11.5. Ed. by Paul Christopher Boutros, e1004071. DOI: 10.1371/journal.pcbi.1004071. URL: https://doi.org/10.1371/journal.pcbi.1004071.

Vaske, Charles J. et al. (June 2010). "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM". In: *Bioinformatics* 26.12, pp. i237–i245. DOI: 10.1093/bioinformatics/btq182. URL: https://doi.org/10.1093/bioinformatics/btq182.

Verhaak, Roel G.W. et al. (2010). "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1". In: *Cancer Cell* 17.1, pp. 98–110.

Vishwanathan, S. V. N., Karsten M. Borgwardt, Imre Risi Kondor, and Nicol N. Schraudolph (2008). "Graph Kernels". In: *CoRR* abs/0807.0093. arXiv: 0807.0093. URL: http://arxiv.org/abs/0807.0093.

Vogelstein, Bert, David Lane, and Arnold J. Levine (Nov. 2000). "Surfing the p53 network". In: *Nature* 408.6810, pp. 307–310. DOI: 10.1038/35042675. URL: https://doi.org/10.1038/35042675.

Wang, Bo et al. (2014). "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature methods* 11.3, p. 333.

Weinstein, John N et al. (Sept. 2013). "The Cancer Genome Atlas PanCancer analysis project". In: *Nature Genetics* 45.10, pp. 1113–1120. DOI: 10.1038/ng.2764. URL: https://doi.org/10.1038/ng.2764.

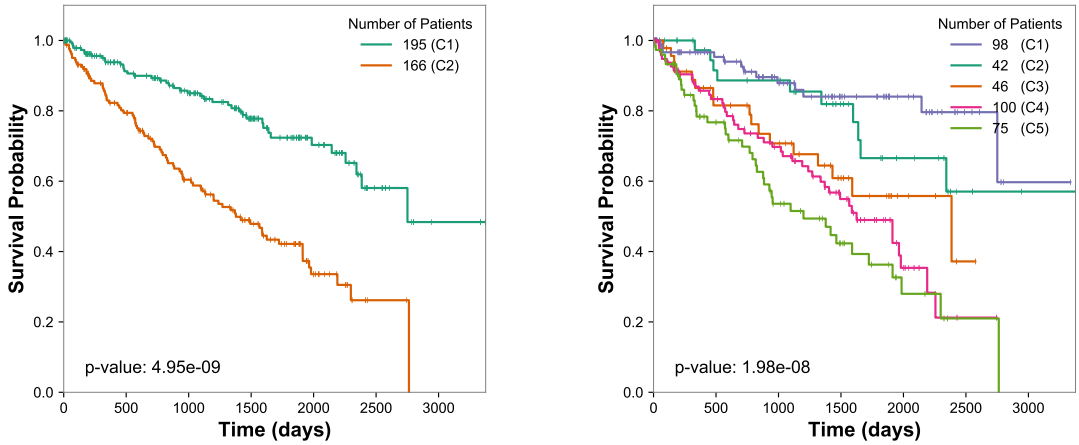
Witten, Daniela M and Robert J. Tibshirani (Jan. 2009). "Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data". In: *Statistical Applications in Genetics and Molecular Biology* 8.1, pp. 1–27. DOI: 10.2202/1544-6115.1470. URL: https://doi.org/10.2202/1544-6115.1470.

Wu, Dingming, Dongfang Wang, Michael Q. Zhang, and Jin Gu (Dec. 2015). "Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification". In: *BMC Genomics* 16.1. DOI: 10.1186/s12864-015-2223-8. URL: https://doi.org/10.1186/s12864-015-2223-8.

Zhou, Dengyong and Christopher JC Burges (2007). "Spectral clustering and transductive learning with multiple views". In: *Proceedings of the 24th international conference on Machine learning*. ACM, pp. 1159–1166.

APPENDIX A

Kaplan Meier Curves for Different k Values



(a) k=2 (b) k=5

Figure A.1 Kaplan-Meier survival curves of the best clustering solutions for KIRC for different number of clusters k = {2,5}. Results obtained with smoothing parameter $\alpha = 0.2$, $\alpha = 0.3$ for k=2(a), k=5(b), respectively. The p-value is the a log-rank test on the survival distributions of between the groups.

Statistical Association of the KIRC Clusters with the Clinical Parameters

Table A.1 Summary of TNM staging according to AJCC Amin et al., 2017. The "X" stands for the degree of parameter that cannot be assessed.

Category	Definition
T1	Tumor ≤ 7 cm in greatest dimension, limited to the kidney
T2	Tumor ≥ 7 cm in greatest dimension, limited to the kidney
N2	Tumor extends into major veins or perinephric tissues but not into the ipsilateral adrenal gland and not beyond Gerota's fascia
T4	Tumor invades beyond Gerota's fascia
N0	No lymph node metastasis
N1	Metastasis in single lymph node, ≤ 3 cm in greatest dimension.
N2	Metastasis in single lymph node, between 3 and 6 cm in greatest dimension or Metastasis in multiple lymph node, ≤ 6 cm in greatest dimension
N3	Metastasis in lymph node, ≥ 6 cm in greatest dimension
M0	No distant metastasis
M1	Distant metastasis
Stage I	T1 - M0
Stage II	T2 - M0
Stage III	T1 or T2 - M0(Additionally metastasis in lymph node) T3 - M0
Stage IV	T4 - M1 Any T - M0
GX	Grade cannot be assessed - Tumor cell and tissue is close to normal
G1	Well differentiated - Tends to grow slowly
G2	Moderately differentiated - Tends to grow rapidly and faster
G3	Poorly differentiated - Tends to grow rapidly and faster
G4	Undifferentiated

Note that for the KIRC subgroups, the Cluster 1 is the patient subgroup with the best prognosis and the Cluster 4 is the worst prognosis. Refer to Figure 4.4a for the cluster ids and their survival distributions and please refer to Supplementary Table A.1 for definition of clinical terms.

Table A.2 Contingency table for tumor stage vs KIRC clusters. The chi-squared test results in $\chi^2 = 52.603$, $p = 3.476e\text{-}08$, df $= 9$

Stage	**I**	**II**	**III**	**IV**	**All**
Cluster No					
1	55	12	25	10	102
2	39	5	10	8	62
3	50	3	18	11	82
4	24	13	44	34	115
ALL	168	33	97	63	361

Table A.3 Contingency table for primary tumor pathological spread vs KIRC cluster. Chi-squared test results in $\chi^2 = 49.479$, $p = 1.349e\text{-}07$, df $= 9$

Pathologic Spread	**T1**	**T2**	**T3**	**T4**	**All**
Cluster No					
1	57	14	30	1	102
2	40	6	16	0	62
3	50	5	26	1	82
4	26	16	69	4	115
ALL	173	41	141	6	361

Table A.4 Contingency table of distant metastasis pathological spread vs KIRC cluster. The chi-squared test results in $\chi^2 = 18.327$, $p = 3.766e\text{-}04$, df $= 3$

Pathologic Spread	**M0**	**M1**	**All**
Cluster No			
1	93	9	102
2	54	8	62
3	70	12	82
4	81	34	115
ALL	298	63	361

Table A.5 Contingency table for neoplasm histological grade vs KIRC clusters. The chi-squared test results in $\chi^2 = 65.608$, $p = 2.104e\text{-}09$, df $= 12$

Histologic grade	**G1**	**G2**	**G3**	**G4**	**GX**	**All**
Cluster No						
1	2	61	33	6	0	102
2	2	26	26	8	0	62
3	1	35	38	8	0	82
4	0	21	52	41	1	115
ALL	5	143	149	63	1	361

Statistical Association of the Other Cancer Clusters with the Clinical Parameters

Table A.6 Contingency table for lymph node stage vs BRCA clusters. The chi-squared test results in $\chi^2 = 23.037$, $p = 3.31e-03$, df $= 8$. While clusters 2&3 is the best prognosis group, the cluster 1 is the worst prognosis group.

Stage	**N0**	**N1**	**N2**	**N3**	**NX**	**All**
Cluster No						
1	41	53	16	9	5	124
2	59	23	10	4	1	97
3	75	49	25	12	1	162
ALL	175	125	51	25	7	383

Table A.7 Contingency table for histologic grade vs HNSC clusters. The chi-squared test results in $\chi^2 = 39.999$, $p = 7.7e-04$, df $= 16$. While clusters 1&3 is the best prognosis group, the cluster 5 is the worst prognosis group.

Stage	**G1**	**G2**	**G3**	**G4**	**GX**	**All**
Cluster No						
1	8	10	7	0	0	25
2	4	31	9	0	0	44
3	0	19	7	0	0	26
4	2	31	15	1	3	52
5	0	35	11	0	2	48
ALL	14	126	49	1	5	195

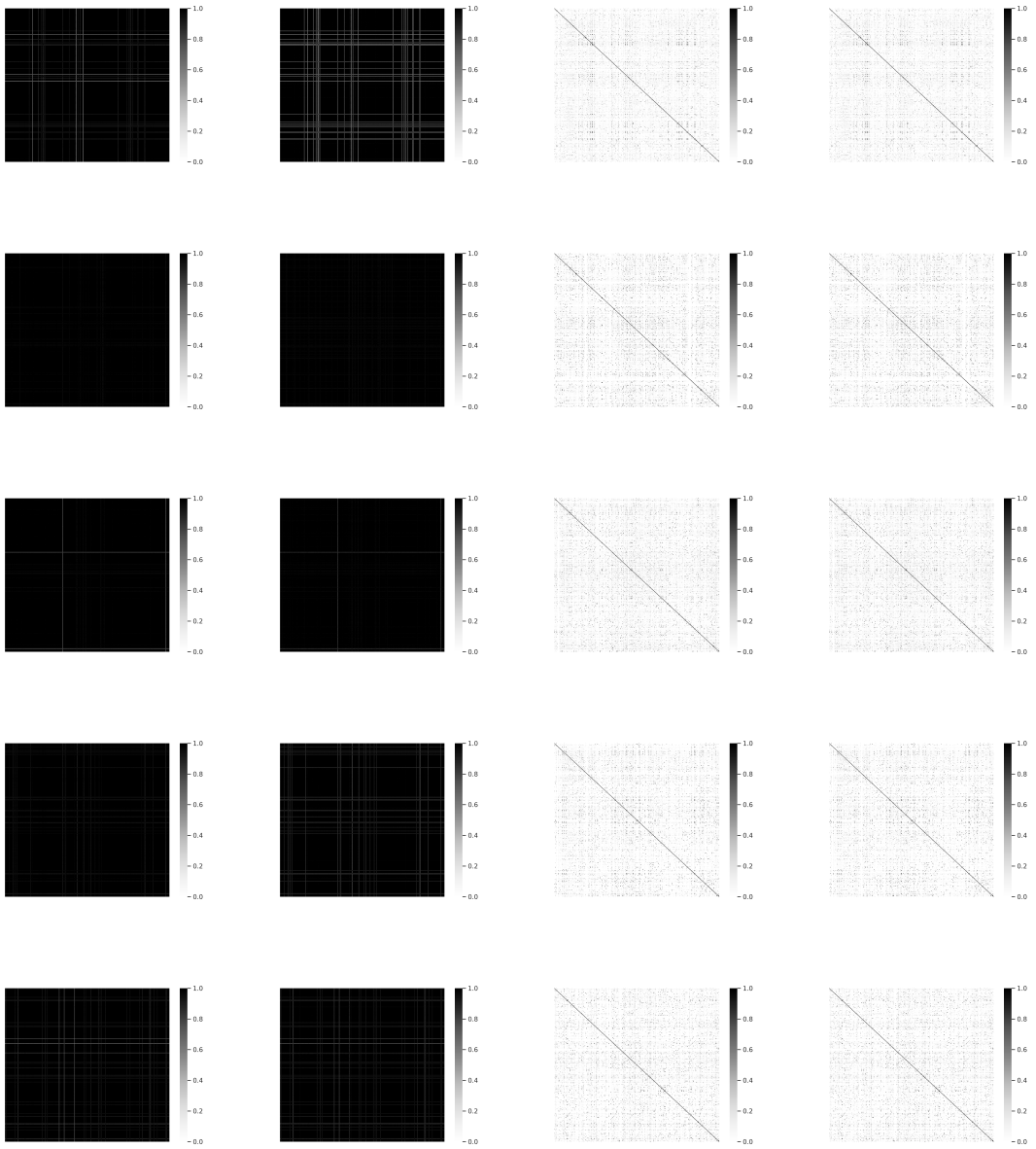
Table A.8 Contingency table for clinic group grade vs HNSC clusters. The chi-squared test results in $\chi^2 = 26.696$, $p = 2.606e-02$, df $= 16$. While clusters 1&3 is the best prognosis group, the cluster 5 is the worst prognosis group.

Stage	**1**	**2**	**3**	**4**	**X**	**All**
Cluster No						
1	8	10	7	0	0	25
2	4	31	9	0	0	44
3	0	19	7	0	0	26
4	2	31	15	1	3	52
5	0	35	11	0	2	48
ALL	14	126	49	1	5	195

Table A.9 Contingency table for primary tumor t stage vs HNSC clusters. The chi-squared test results in $\chi^2 = 26.821$, $p = 4.351e-02$, df $= 16$. While clusters 1&3 is the best prognosis group, the cluster 5 is the worst prognosis group.

Stage	**T1**	**T2**	**T3**	**T4**	**TX**	**All**
Cluster No						
1	1	10	4	5	5	25
2	5	13	8	10	8	44
3	1	7	7	7	4	26
4	0	12	15	21	4	52
5	0	7	15	14	12	48
ALL	7	49	49	57	33	195

Heatmaps of Kernel Examples



(a) Propagation Kernel

(b) Graph Hopper Kernel

(c) Wasserstein Weisfeiler Lehman Kernel

(d) Smoothed Shortest Path Kernel

Figure A.2 Patient-by-patient kernel matrices calculated by different kernel choices for KIRC patients. The kernel functions include the propagation kernel, graph hopper kernel, wasserstein weisfeiler lehman, and SmSPK graph kernel methods. Each row corresponds to a randomly chosen pathway and molecular interaction data type. A color black indicates that the two patient similarity is evaluated as 1.