IDENTIFICATION OF CANCER PATIENT SUBGROUPS VIA PATHWAY BASED MULTI-VIEW GRAPH KERNEL CLUSTERING

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Ali Burak Ünal

July 2017

Identification of Cancer Patient Subgroups via Pathway Based Multi-view Graph Kernel Clustering
By Ali Burak Ünal
July 2017

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Öznur Taştan Okan(Advisor)

Nurcan Tunçbağ

Ercüment Çiçek

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

IDENTIFICATION OF CANCER PATIENT SUBGROUPS VIA PATHWAY BASED MULTI-VIEW GRAPH KERNEL CLUSTERING

Ali Burak Ünal
M.S. in Computer Engineering
Advisor: Öznur Taştan Okan
July 2017

Characterizing patient genomic alterations through next-generation sequencing technologies opens up new opportunities for refining cancer subtypes. Different omics data provide different views into the molecular biology of the tumors. However, tumor cells exhibit high levels of heterogeneity, and different patients harbor different combinations of molecular alterations. On the other hand, different alterations may perturb the same biological pathways. In this work, we propose a novel clustering procedure that quantifies the similarities of patients from their alteration profiles on pathways via a novel graph kernel. For each pathway and patient pair, a vertex labeled undirected graph is constructed based on the patient molecular alterations and the pathway interactions. The proposed smoothed shortest path graph kernel (smSPK) assesses similarities of pair of patients with respect to a pathway by comparing their vertex labeled graphs. Our clustering procedure involves two steps. In the first step, the smSPK kernel matrices for each pathway and data type are computed for patient pairs to construct multiple kernel matrices and in the ensuing step, these kernel matrices are input to a multi-view kernel clustering algorithm to stratify patients. We apply our methodology to 361 renal cell carcinoma patients, using somatic mutations, gene and protein expressions data. This approach yields subgroup of patients that differ significantly in their survival times (p-value $\leq 1.5 \times 10^{-8}$). The proposed methodology allows integrating other type of omics data and provides insight into disrupted pathways in each patient subgroup.

Keywords: Cancer subtype, graph kernel, multi-view kernel clustering, kernel functions, clustering.

ÖZET

KANSER HASTA ALT GRUPLARININ YOLAK ESASLI ÇOK BAKIŞLI ÇIZGE ÇEKİRDEĞİ GRUPLAMASI İLE BELİRLENMESİ

Ali Burak Ünal
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Öznur Taştan Okan
Temmuz 2017

Yeni nesil dizi analizi teknolojisi ile hasta genomik değişimlerin nitelendirilmesi kanser alt tiplerinin belirlenmesinde yeni olanaklar ortaya çıkarıyor. Farklı omik verileri, tümörlerin moleküler biyolojilerine farklı bakış açıları sağlar; bununla birlikte, tümör hücreleri yüksek seviyede heterojenlik sergiler, ve farklı hastalar farklı kombinasyonlarda moleküler değişikliklere sahiptir. Öte yandan, farklı değişiklikler aynı biyolojik yolakları bozabilir. Bu çalışmada, yeni bir çizge çekirdeği aracılığıyla hastaların alterasyon profillerinden yolaklar üzerindeki benzerliklerini nicelleştiren yeni bir kümeleme prosedürü öneriyoruz. Her bir yol ve hasta çifti için, hastanın moleküler değişiklikleri ve yolak etkileşimlerine dayanarak düğümleri etiketlenmiş yönsüz bir çizge oluşturulur. Önerilen dağıtılmış en kısa yol çizge çekirdeği (smSPK), bir yolağa göre hasta çiftlerinin düğümleri etiketli çizgelerini karşılaştırarak benzerliklerini değerlendirir. Gruplama prosedürümüz iki adımdan oluşur. İlk adımda, her yol ve veri tipi için smSPK çekirdek matrisleri, hasta çiftleri için birden çok çekirdek matrisi oluşturmak üzere hesaplanır ve sonraki adımda, bu çekirdek matrisleri, hastaları katmanlaştırmak için çok bakışlı çekirdek gruplandırma yaklaşımına girdi olarak verilir. Metodolojimizi 361 renal hücreli karsinoma hastasında somatik mutasyonlar, gen ve protein ifadeleri verileri kullanarak uyguluyoruz. Bu yaklaşım, hayatta kalma sürelerinde önemli farklılık gösteren hasta alt gruplarını ortaya çıkarıyor (p-değeri $\leq 1.5 \times 10^{-8}$). Önerilen yöntem, diğer omik verilerin entegrasyonuna izin verir ve her hasta alt grubundaki bozuk yolaklarla ilgili fikir verir.

Anahtar sözcükler: Kanser alt tipleri, çizge çekirdeği, çok bakışlı çizge gruplaması, çekirdek fonksiyonları, gruplama.

Acknowledgement

First of all, I would like to thank my supervisor Asst. Prof. Dr. Öznur Taştan Okan. Without her motivation, support and especially understanding, this thesis would not be possible. I also thank Asst. Prof. Dr. Nurcan Tunçbağ and Asst. Prof. Dr. Ercüment Çiçek for their helpful and constructive feedbacks.

Moreover, I want to thank people from my graduate study; Onur Taşar, Iman, Bülent, Noushin, Caner, Yarkın, Gencer, Troya, Simge, İstemi, Necmi, Fuat, Arif Usta, Can Fahrettin, Gülden and others and specially Ebru Ateş. Since I cannot finalize this acknowledgement without expressing my gratitude to people who have been in my life since my undergraduate study, I want to thank Ahmet Küçük, Kamil, Şaban, Alper, Raşit, Hüseyin, Ahmet (Kaptan), Metin, Abdul, Fatma, Nihan, Betül, Cansu, Serdar (Aslan Kral) and many others.

I also thank Seher, Jonathon, Başak and Sayın Güvercin. We have shared so many beautiful memories which I cannot forget. They, especially Seher, have given so many constructive and useful advices.

I owe Emin Yıldırım, Onur Barut and Hüseyin Kılıç so many things I have today. They are among the most influencial people in my life along with my brother.

Most importantly, I would like to express my gratitude to my beloved mother Ayşe, father İsmet, brother Çağrı and sister Duygu. I am truly thankful for your support and love. I also want to thank Kenan, Onur and Emre whom I consider them from the family.

Finally, I want to thank Elif Eser with whom I share so many things. She has a huge share in whatever I have achieved in my graduate study. I am doubtlessly grateful to her.

At last, I thank TÜBİTAK for supporting me in my graduate study through BİDEB scholarship program.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is a molecularly diverse disease; within the same cancer type, tumors exhibit distinct pathological features and bear different molecular alterations. This heterogeneity exhibits itself as variation in the patient clinical trajectories; although classified in the same cancer type, some patients respond well to therapy and have longer survival rates, whereas others succumb to the disease. Accurate patient stratification based on molecular profiles of patients is therefore essential for better diagnosis of patients, understanding molecular mechanisms that drive these different cancer subtypes to carcinogenesis and for developing subtype targeted treatment strategies.

The utility of characterizing molecular subtypes based on molecular profiles has been exemplified in the case of breast cancer, where molecular subtypes are defined based on gene expression profiles [1, 2, 3] and subtype specific treatment strategies have been developed based on these subtypes [4]. On the other hand for many other cancer types molecular subtypes are yet to be defined. One such cancer is renal cell carcinoma (RCC). RCC is a cancer type that originates from the renal epithelium and is the major class of kidney cancer accounting for $70 - 80\%$ of cancers in the kidney [5]. In this thesis, we focus on identifying patient subroups of RCC.

With the advancement of sequencing technologies, patient-specific molecular alteration datasets are now available for a large number of cancer patients and we are now able to sequence DNA and RNA more accurately and faster. In addition to protein expressions among cancer patients, their somatic mutations and gene expressions are more precisely available for large cohorts [6]. By setting a different *view* into the molecular biology of the tumor, such large catalogues of molecular alterations open up new opportunities to refine the subtypes of many cancer types.

Although molecular alterations are useful, the heterogeneity of the molecular alterations still exhibits a challenge. For example, there are very few gene mutations shared among the patients and most patients harbor rare mutations [7]. Even if different genes were to be affected in each patient, they might serve to the same end result of affecting the same cellular mechanisms. In other words, the disturbance of the same process in two patients do not necessitate the alteration of the same genes. Taking this fact into account, mapping mutations on biological networks by making use of the current knowledge of interactions have been proposed in the literature. This was shown to be a more effective strategy for interpreting genetic alterations compared to strategies that would center on individual genes [7]. One such class of biological knowledge is pathways. Pathways are diagrams that summarize interactions of well-studied cellular processes. They consist of genes, orthologs, varying chemical components, and their interactions with each other. These interactions represent regulation and signaling events, and biochemical reactions.

In this work, we develop a computational approach that provides a flexible way of integrating different molecular alteration data and incorporating prior biological knowledge on cellular interactions. To achieve this, we assess patient similarities based on their molecular alterations on pathways through a novel graph kernel. In our framework, in order to assess patient similarities, graph kernels have been proposed in literature for measuring the similarity of pairs of graphs and are shown to be successful in comparing structured objects [8, 9]. The suggested graph kernels; however, are designed for comparing graphs with different topological structure and focus on finding similar subgraphs. In our

problem, though, for each pathway we derive a single undirected graph, where the vertices represent genes and the edges represent interactions between genes. For each patient the vertices of the graph are labeled based on the set of molecular alterations the patient harbors. Thus, the graphs that we want to compare are topologically identical but the vertex label distributions are different. To this end, we propose a smoothed shortest path graph kernel (smSPK), which compares two graphs based on their vertex label distributions while accounting for the underlying graph topology. To our knowledge, this is the first study that makes use of graph kernels for cancer subtype identification.

Our clustering procedure involves comparing patients based on their molecular alterations on KEGG pathways using smSPK. Each type of alterations is mapped on KEGG pathways separately and a kernel matrix is calculated separately for each pathway. These kernel matrices are input to multi-view kernel clustering approach proposed by Tzortzis et al. [10] in order to retrieve the clusters of patients. The stratification of patients this way enables us to incorporate prior biological knowledge from different pathways and molecular alterations, and help us capture patients similarities that stem from the dysregulation of similar processes in the pathways. The method also offers additional insights by showing how informative each pathway is to the clustering process.

In this study, we apply this framework to elucidate the intrinsic subtypes of RCC. We utilize patient somatic mutations, gene expression levels and protein expression levels. The framework is flexible enough to integrate other available alteration data (copy number variation, methylation, miRNA expression, etc.). We also assess whether the newly identified clusters correlate well with the clinical outcomes through survival analysis of patient subgroups. We apply this framework to elucidate the intrinsic subtypes of RCC. We characterize the key biological features of the core subtypes for RCC and arrive at clusters that differ in their survival distribution.

The outline of the thesis is as follows: In Chapter 2, we present a literature review on the commonly used clustering approaches for cancer subtype identification. We also review multi-view kernel clustering approaches and graph kernels.

In Chapter 3, we introduce our framework. We explain the data sets utilized in the experiments. Next, we formulate smSPK and describe the multi-view graph kernel clustering step. Finally, we detail the clustering evaluation techniques that are employed. Chapter 4 describes the setup of experiments and presents original results obtained by applying our methodology to RCC and the results of compared approaches. Once the results are given, we discuss these results and evaluate the performance of our methodology. Chapter 5 concludes and lists possible future directions for the work presented in this thesis.

Chapter 2

Related Work

In this chapter, we review the related work under different headings. First, we will cover clustering techniques that are widely applied in cancer subtype identification. Clustering analysis refers to a broad set of techniques that seeks to partition the observations such that observations that are assigned to the same group are similar while those in different groups are dissimilar and are formulated with a vast array of approaches. We will not attempt to exhaustively cover all possible strategies but instead we will limit our discussion to the widely applied approaches for finding cancer subtypes. We will dedicate a section for multi-view clustering as it is utilized in this study. We will also review related recent work in network and pathway assisted approaches that make use of known functional relationships to integrated different views. Finally, will review the existing graph kernel methods in the literature in terms of their expressiveness and potential applicability to this problem.

2.1 Traditional Clustering Approaches for Cancer Subtype Identification

The three methods that are most widely used in cancer subtype identification include hierarchical clustering, non-negative matrix factorization and consensus clustering. Below we briefly introduce and discuss these methodologies.

2.1.1 Hierarchical Clustering

Hierarchical clustering groups examples into partitions in a hierarchical manner and produces a nested set of clusterings [11]. Hierarchical clustering can be conducted in an agglomerative (bottom-up) or a divisive (top-down) fashion. The agglomerative approach finds more frequent use in cancer subtype identification. Here, each sample initially belongs to a separate cluster and at each step of the algorithm; clusters that are most similar are merged. The iteration is repeated until all the samples are finally in the same cluster. In the divisive approach, the clustering starts with all examples in the same cluster and partition clusters further at each iterative step.

In addition to the similarity of the samples, hierarchical clustering needs a definition of similarity of clusters, referred to as linkage methods. In single linkage, the distance between the clusters are defined as the distance between the closest examples points of two clusters; while in complete linkage it is defined as the distance between the farthest points in the two clusters. Average linkage, on the other hand, considers the average of the distances between the members of the two clusters.

Hierarchical clustering used in many of the seminal work in cancer subgroup discovery [1]. Verhaak et al. used hierarchical clustering with consensus clustering to discover Glioblastoma Muliformae (GBM) subtypes [12]. Distinct types of B-cell lymphoma is identified using hierarchical clustering of gene expression data [13]. In other cancer types too, hierarchical clustering has been widely used

segregate tumors into distinct molecular subtypes [14, 15, 16, 17].

2.1.2 Consensus Clustering

Another widely adapted methodology is consensus clustering [18]. Consensus clustering is a meta-algorithm that can be used in conjunction with different clustering algorithms. It aims at finding stable clusters that are robust to variation in samples. To achieve this, several bootstrap examples are selected and the clustering algorithm is run multiple times with each of the bootstrap samples as input. In the ensuing step, the cluster assignments obtained at each clustering run are combined in a consensus matrix, entries of which report the frequency common cluster assignments for pairs of items.

Consensus clustering is widely used in cancer subgroup identification. Verhaak et al. [12] utilized consensus clustering with hierarchical clustering to find GBM subtypes. TCGA Network group used NMF consensus clustering to find subgroups in breast cancer patients [19] while Brannon et al. used a similar methodology to find subtypes in clear cell renal cell carcinoma [20].

2.1.3 Non-Negative Matrix Factorization

The non-negative matrix factorization(NMF) formulates the clustering task as a matrix factorization problem. Consider a $n \times m$ matrix, $\boldsymbol{V}$, where n is the number of samples and m is the number of features derived from molecular profiles and each element $V_{i,j} \geq 0$. Given desired rank $k < min\{m, n\}$, which is the number of clusters, NMF decomposes $\boldsymbol{V}$ into two non-negative matrices $\boldsymbol{W}$ $(n \times k)$ and $\boldsymbol{H}$ $(k \times m)$ such that

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}$$

This formulation explains each example by an additive linear combination of

few basis components that is defined by columns of $\boldsymbol{H}$. If there are few basis vectors that captures the clustering structure in the data, a good approximation to the original matrix $\boldsymbol{V}$ can be obtained. The goodness of the approximation is assessed with Frobenius norm and the factorization is achieved through solving the following optimization problem through iterative algorithms [21].

$$\min_{\boldsymbol{W}\geq 0, \boldsymbol{H}\geq 0} f(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{2} \|\boldsymbol{A} - \boldsymbol{W}\boldsymbol{H}\|_F^2$$

A large number of studies have applied NMF clustering to find cancer subtypes [22, 23].

2.2 Integrating Different Data Sources

The simplest way of integrating different data sources is to concatenate normalized measurements obtained from different sources such as mRNA expression, somatic mutations, etc.

Speicher et al. [24] introduced a regularized unsupervised multiple kernel learning which integrates different data types to identify cancer subtypes. The method, called regularized Multiple Kernel Learning for dimensionality reduction with Locality Preserving Projections (rMKL-LPP), is applied on five cancer data sets to determine the subtypes of these cancer types. The idea behind the proposed method is to combine kernel matrices obtained from each data type on which the dimensionality reduction that conserves the distances of samples to its k nearest neighbors is applied. The projection vector lies in the span of the samples. Thus, the projection vector $\boldsymbol{v}$ is written as $\boldsymbol{v} = \sum_{i=1}^{N} \alpha_i \phi(x_i)$ where N is the number of samples, α_i is the weight of i^{th} mapped feature. The kernelized version of the original optimization problem for solving the optimal projection vector is formulated

as follows:

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

where $\boldsymbol{M}$ is the number of kernel matrices, C is a real constant, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]$ are the weights of mapped feature vectors in the projection vector, $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_M]$ are the weights of kernels, w_{ij} is an entry of a similarity matrix $\boldsymbol{W}$, d_{ij} is an entry of a constraint matrix $\boldsymbol{D}$ to prevent trivial solution. w_{ij} is 1 if sample i and sample j are in the k-nearest-neighborhood of each other. Otherwise, it is 0. d_{ij} is defined as $\sum_{n=1}^{N} w_{in}$ if $i = j$. Otherwise, it is 0. The purpose of the constraint $\|\boldsymbol{\beta}\|_1 = 1$ is included to prevent overfitting. $\boldsymbol{\mathcal{K}}$ is defined as follows:

$$\boldsymbol{\mathcal{K}}^i = \begin{pmatrix} \boldsymbol{K}_1(1,i) & \cdots & \boldsymbol{K}_M(1,i) \\ \vdots & \ddots & \vdots \\ \boldsymbol{K}_1(N,i) & \cdots & \boldsymbol{K}_M(N,i) \end{pmatrix}$$

Since the proposed optimization problem is hard to solve, an alternating optimization technique is employed. In this technique, one of the parameters over which the objective function is optimized is fixed and the objective function is optimized over the free parameter. Then, the previous free parameter is fixed to the optimum value found in the previous step and the objective function is optimized over the previously fixed parameter. This process continues until the objective value converges. The proposed method, rMKL-LPP, is experimented on five different cancer types, namely glioblastoma multiforme, breast invasive carcinoma, kidney renal clear cell carcinoma, lung squamous cell carcinoma and colon adenocarcinoma. It generates 6, 7, 14, 6 and 6 clusters for each cancer type respectively. The resulting clusters of each cancer type are evaluated by survival analysis and the analysis yielded statistically significant difference among the survival time of clusters. In the interpretation step of the results, one must need to analyze the statistics of common mutations or different mutations which possibly

yield the separation of patients. Even if this analysis is made, it is not guaranteed that the analysis yields meaningful information of underlying structure of the clusters.

2.3 Multi-View Clustering Approaches

In the literature, there are number of studies on multi-view kernel clustering. Optimized kernel k-means clustering (OKKC) proposed by Yu et al. [25] is one of the well-known approaches. The objective of OKKC is to find the optimal kernel weights for combining kernels jointly with the optimal cluster assignment. The corresponding optimization problem is formulated as follows:

$$\begin{aligned} \underset{\boldsymbol{A},\boldsymbol{\Theta}}{\text{maximize}} \quad & \text{Tr}(\boldsymbol{A}^T\boldsymbol{\Omega}\boldsymbol{\Omega}\boldsymbol{A}) \\ \text{subject to} \quad & \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_k \\ & \boldsymbol{\Omega} = \sum_{i=1}^{p} \theta_i \boldsymbol{G}_i \\ & \theta_i \geq 0, \quad i = 1, 2, \dots p \\ & \sum_{i=1}^{p} \theta_i^{\delta} = 1 \end{aligned}$$

where N is the number of samples, p is the number of views (i.e. kernels), k is the number of clusters in k-means, $\boldsymbol{G}_i$ is the centered kernel matrix for i^{th} view, θ_i is the assigned weight of i^{th} kernel matrix, $\boldsymbol{\Theta}$ is the vector of all weights, δ is a parameter to adjust the sparsity of assigned weights, $\boldsymbol{A}$ is defined as follows:

$$\boldsymbol{A}_{ij} = \begin{cases} \frac{1}{\sqrt{n_j}} & \text{if sample } i \text{ in cluster } j \\ 0 & otherwise \end{cases}$$

where n_j is the number of samples in cluster j. Since the solution for the formulated optimization problem is hard, they resort to an alternating optimization scheme. First the objective function is optimized over the cluster assignment matrix $\boldsymbol{A}$ by fixing the weight vector $\boldsymbol{\Theta}$. Then, $\boldsymbol{A}$ is fixed to the optimal value obtained in the previous step and the objective function is optimized over $\boldsymbol{\Theta}$.

This iterative process continues until the objective value converges. The time complexity of the proposed algorithm is $O(\gamma[N^3 + \nu(N^2 + p^3)] + lkN^2)$ where γ is the number of OKKC iterations for finding the optimal value of the objective function, ν is the number of semi-infinite programming iterations, l is the fixed number of k-means clustering iteration, k is the number of clusters. They experimented OKKC on five data sets from UCI machine learning repository, namely Iris, Wine, Yeast, Satimage and Pen digit recognition, a data set from bioinformatics for clustering genes which are disease relevant and a data set from scientometrics for clustering journal publications. The comparison of the results of these experiments with single best kernel matrix experiment and non-linear adaptive metric learning algorithm [26] revealed that OKKC outperforms competing strategies in many data sets.

Tzortzis et al. [10] propose multi-view kernel k-means (MVKKM). MVKKM aims to combine the information coming from different representation of the same sample set in order to achieve a better clustering result. The proposed algorithm works on kernel matrices each of which corresponds to a different representation, which is called as "view". The method seeks for the best kernel matrix weights that yield the best clustering result. In order to achieve this, the objective function is defined as follows:

$$\begin{aligned} \underset{\boldsymbol{Y},\boldsymbol{W}}{\text{minimize}} \quad & \sum_{v=1}^{V} w_v^p (\text{Tr}(\boldsymbol{K}^{(v)}) - \text{Tr}(\boldsymbol{Y}^T \boldsymbol{K}^{(v)} \boldsymbol{Y})) \\ \text{subject to} \quad & w_v \geq 0, \quad v = 1, 2, \ldots, V \\ & \sum_{v=1}^{V} w_v = 1 \\ & p \geq 1 \end{aligned}$$

where V is the number of views, w_v is the weight of v^{th} view, $\boldsymbol{W}$ is the vector of all weights of views, p is the parameter adjusting the sparsity of weights of views, $\boldsymbol{K}^{(v)}$ is the kernel matrix of v^{th} view, $\boldsymbol{Y}$ is a cluster indicator matrix with $\boldsymbol{Y}_{ik} = \frac{\delta_{ik}}{\sqrt{\sum_{j=1}^{N} \delta_{jk}}}$ where δ_{ik} is 1 if i^{th} sample is in k^{th} cluster. Otherwise, it is 0. Again, an alternating optimization strategy is employed to attain the desired solution. At first, the cluster assignments are updated by fixing the weights of

kernel matrices. All input kernel matrices are summed by using the fixed weights found in the previous iteration. The resulting kernel matrix is input to kernel k-means clustering to obtain the cluster assignment of samples. When the clusters are determined, the weights of kernel matrices are updated for each view v as follows:

$$w_v = \frac{1}{\sum_{v'=1}^{V} \left(\frac{D_v}{D_{v'}}\right)^{(\frac{1}{p-1})}} \quad \text{if } p > 1$$

where D_v is defined as:

$$D_v = \sum_{i=1}^{N} \sum_{k=1}^{M} \delta_{ik} \left\| \Phi^{(v)}(\boldsymbol{x}_i^{(v)}) - \boldsymbol{m}_k^{(v)} \right\|^2$$

where N is the number of samples, M is the number of clusters, $\Phi^{(v)}$ is the feature mapping function for view v, $\boldsymbol{m}_k^{(v)}$ is the cluster centroid of k^{th} cluster in view v. The update of weights when $p = 1$ is as follows:

$$w_v = \begin{cases} 1, & v = \underset{v'}{\operatorname{argmin}} \quad D_{v'} \\ 0 & otherwise \end{cases}$$

They experimented MVKKM on synthetic data and a number of real data sets, namely the multiple feature data sets (i.e. handwritten data sets) and Corel images data set. The performance of MVKKM is compared with multi-view spectral clustering, correlational spectral clustering [27] and weighted multi-view convex mixture models [28]. The results of experiments showed that MVKKM outperforms other algorithms in all data sets. However; when some views have irrelevant or noisy information, the performance of MVKKM decreases. Furthermore, MVKKM is tested on small number of views. When relatively high number of views are input to MVKKM, MVKKM do not perform well.

In addition to the kernel weighted approaches, Gönen et al. [29] proposed a multi-view kernel k-means clustering approach, called localized multiple kernel k-means (LMKKM), in which each sample in each view (i.e. kernel) has a special weight. In this method, combination of kernels is achieved by a sample weighted summation. In order to determine the weights of samples in each kernel matrix,

they formulate an optimization problem as follows:

$$\begin{aligned} \underset{H,\Theta}{\text{maximize}} \quad & \text{Tr}(\boldsymbol{H}^T \boldsymbol{K}_\Theta \boldsymbol{H} - \boldsymbol{K}_\Theta) \\ \text{subject to} \quad & \boldsymbol{H}^T \boldsymbol{H} = \boldsymbol{I}_k \\ & \boldsymbol{\Theta} 1_p = 1_n \end{aligned}$$

where p is the number of kernels, $\boldsymbol{K}_\Theta = \sum_{i=1}^{V} (\boldsymbol{\theta}_i \boldsymbol{\theta}_i^T) \circ \boldsymbol{K}_i$ where $\boldsymbol{\theta}_i$ is the vector of weights of samples in kernel i, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2\ \ldots\ \boldsymbol{\theta}_p]$ is the matrix of sample weights of all kernels, $\boldsymbol{H}$ is an orthogonal matrix with arbitrary real values. As it is the case in OKKC and MVKKM, optimizing over two variables at the same time is hard. Therefore, alternating optimization technique is utilized. $\boldsymbol{H}$ is optimized assuming $\boldsymbol{\Theta}$ is given. The next step is to fix the value of H to the optimal value found in the previous step and optimize over $\boldsymbol{\Theta}$. The proposed algorithm is experimented on human colon and rectal cancer data set from The Cancer Genome Atlas (TCGA). DNA copy number, gene expression and DNA methylation data of patients in TCGA are utilized. The analysis of results reveals that LMKKM outperforms and identified patient groups demonstrate clinically distinct characteristics.

2.4 Network and Pathway Assisted Approaches

There are also methods in the literature that make use of existing biological networks and pathways.

Hofree et al. [22] present network-based stratification (NBS) network propagation technique. The method employs a protein-protein interaction network on which somatic mutation profiles of patients are mapped as node labels. If the gene has a mutation, which could be a point mutation, an insertion or a deletion, then the label of the corresponding node is assigned 1. Otherwise, it is assigned 0. Next, the effect of mutations are spread to the neighboring genes using a network propagation technique proposed by Vanunu et al. [30]. The underlying idea is that patients having similar clinical outcomes may have too few common mutations [19], [23], [31]. By spreading the effect of mutation on a gene, the similarity

of patients who have different but close-by mutated genes can be identified. Each patient is described based on the mutations propagated on the network and NMF is run to obtain the clusters of patients. Hofree et al. applied this technique to identify the subtypes of ovarian, uterine and lung cancer.

Wang et al. [32] proposed similarity network fusion (SNF) to stratify patients into clinically meaningful subtypes. They utilize a separate network for each data type of each patient and these networks are fused into a single network which covers the information of all data types. The underlying idea is to make use of the complementarity of data types.

Liu et al. [33] introduced another network assisted approach called network-assisted co-clustering for the identification of cancer subtypes (NCIS). The algorithm integrates the information of gene network in order to cluster samples such that the clinical trajectories of each group is different. NCIS assigns a weight to each gene considering the impact of it in the network. Next, weighted co-clustering algorithm making use of semi-nonnegative matrix tri-factorization is utilized to obtain the groups.

Besides the network assisted approaches, Kim et al. [34] present three different pathway assisted methods. These methods benefit from gene set enrichment algorithms (GSEA) to perform clustering. GSEA-based Leading Edge Gene feature (GLEG) is one of them and it uses GSEA leading edge genes as features to stratify patients. The second one is GSEA Pathway Feature (GPF), which utilizes GSEA-enriched pathways as features. SVM-based Pathway Feature (SPF) is the third one and it employs pathway features which are decided by SVM. All these methods are experimented on ovarian and breast cancer so as to demonstrate the performance.

2.5 Graph Kernels

Graph kernels are specialized kernels to compare graphs [38]. There is a substantial literature that describes proposed graph kernels utilizing different characteristics of graphs.

2.5.1 Random Walk Kernel

Random walk graph kernels measure the similarity of graphs based on the number of common random walks in the compared graphs [35]. Consider $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ where V_1 and V_2 are the vertex sets and E_1 and E_2 are the edge sets of corresponding graphs. To compute random walks a graph product is used. Performing a random walk on the direct product graph is equivalent to performing a simultaneous random walk on G_1 and G_2 [36]. The direct graph product of G_1 and G_2 is denoted as G_x and its vertex and edge sets are defined as follows:

$$\begin{aligned} V(G_1 \times G_2) =& \{(v_1, v_2) \in V_1 \times V_2 : (label(v_1) = label(v_2)\} \\ E(G_1 \times G_2) =& \{((u_1, u_2), (v_1, v_2)) \in V^2(G_1 \times G_2) : \\ & (u_1, v_1) \in E_1 \wedge (u_2, v_2) \in E_2 \wedge (label(u_1, v_1) = label(u_2, v_2))\} \end{aligned}$$

In the direct product, an edge exists if and only if corresponding nodes are adjacent in both G_1 and G_2. The random walk kernel function between graphs G_1 and G_2 is defined as follows:

$$k_x(G_1, G_2) = \sum_{i,j=1}^{|V_x|} \left[\sum_{n=0}^{\infty} \lambda_n \boldsymbol{E}_{\boldsymbol{x}}^n \right]_{ij}$$

where $\boldsymbol{E_x}$ is the adjacency matrix of direct graph product G_x.

Random walk suffers from *tottering*. In a walk it is possible to visit same nodes and edges repeatedly which yields to artificially high similarity scores. In [37], the idea of adding additional node labels is presented as a remedy to reduce the

number of matching nodes. Another issue with random walk graph kernel is efficiency, which is ($O(n^6)$ [38] in a nave implementation. Three techniques are used to reduce runtime of the algorithm: Sylvester equations, conjugate gradients, and fixed-point iterations. Although the worst-case complexity of Sylvester equation method is $O(n^3)$, conjugate gradients and fixed-point methods perform faster in the experiments of [38].

2.5.2 Shortest-Path Graph Kernel

Another path-based graph kernel is proposed by Borgwardt et al. [9]. The shortest path graph kernel utilizes the shortest paths of nodes in graphs to calculate the similarity of input graphs. In calculating the shortest path kernel, the graph is converted into shortest-paths graph in which nodes that are connected by a shortest path in the original graph are connected with an edge. The edge weights are the length of the shortest path between the nodes of the original graph. To be able to construct such a graph, one first needs to find the shortest paths between all pairs of nodes in the original graph. For this purpose, they employ Floyd-Warshall algorithm [39] which finds the all-pairs-shortest-paths in a graph. After constructing shortest-paths graph utilizing Floyd-Warshall algorithm, the graph is input to the shortest-path graph kernel which is defined as follows:

$$K_{sp}(S_1, S_2) = \sum_{e_1 \in E_1} \sum_{e_2 \in E_2} K^{(1)}_{walk}(e_1, e_2)$$

where $K^{(1)}_{walk}$ is a positive definite kernel on edge walks of length 1. This graph kernel is designed assess similarity of graphs with different topologies.

2.5.3 Weisfeiler-Lehman (WL) Graph Kernels

Shervashidze et al. [8] introduces Weisfeiler-Lehman (WL) kernel that contains subtree-based, edge-based graph and path-based graph kernels. The introduced graph kernel framework is based on the WL graph isomorphism test. Underlying idea of this algorithm is to combine the sorted form of labels of neighboring nodes

with the label of current node and compress this augmented node into a short node label representation. This process is repeated until the label set of two graphs differ or the number of iterations reaches to a pre-determined value. WL graph isomorphism test is utilized in WL kernel framework for two graphs G and G' as follows:

$$K_{WL}^{(h)} = K(G_0, G'_0) + K(G_1, G'_1) + \ldots + K(G_h, G'_h)$$

where h is the number of WL graph isomorphism test iteration, G_i and G'_i are respectively the WL graphs of G and G' at i^{th} iteration of WL graph isomorphism test, k is any positive semidefinite graph kernel.

The most distinguishing feature of WL kernel framework is the flexibility to integrate any positive semi-definite graph kernel. They introduce three different graph kernels belonging to WL graph kernel family. The first one is the WL subtree kernel that utilizes the number of matching original and compressed labels of two graphs. The WL subtree graph kernel forms a vector of counts of each label for each graph and takes the dot product of these vectors to calculate the similarity between two graphs. This is repeated for each WL graph in WL graph isomorphism test and the similarity scores are simply summed up to obtain the final similarity score of these graphs. The second proposed graph kernel of WL graph kernel family is WL edge kernel. This graph kernel relies on the number of matching edges in two graphs. An edge is counted as a matching if the nodes enclosing the edge have the same label distribution in both graphs. To calculate the similarity of two graphs in the i^{th} iteration of WL graph isomorphism test, the graph kernel forms a vector of counts of edges for each graph and takes the dot product of these vectors of graphs. The total similarity is calculated by summing up the similarity scores of graphs in each WL graphs. The last introduced graph kernel belonging to WL graph kernel family is WL shortest path kernel. The base kernel here is the shortest-path kernel, which was explained previously in this chapter. In this scenario, the similarity of two graphs is the summation of the similarity scores of each WL graph in WL graph isomorphism test calculated by the shortest-path kernel.

Chapter 3

Methodology

In this section, we present our proposed methodology for patient subgroup identification. We outline the overview of the framework and then detail data curation and processing steps. Next, we introduce the proposed shortest path graph kernel and describe how the multi-view clustering is employed. Finally, we present the survival analysis strategy which is used to assess the survival differences among newly added clusters.

3.1 Framework

Our proposed framework starts with mapping alterations of different data types on pathways separately. Once we have the alteration mapped pathways, we compute the kernel matrix of each pathway of each data type by utilizing smSPK. In order to obtain the subgroups of patients, we input these kernel matrices to MVKKM. Finally, we analyze the identified subgroups in terms of their survival times to evaluate the performance of our framework.

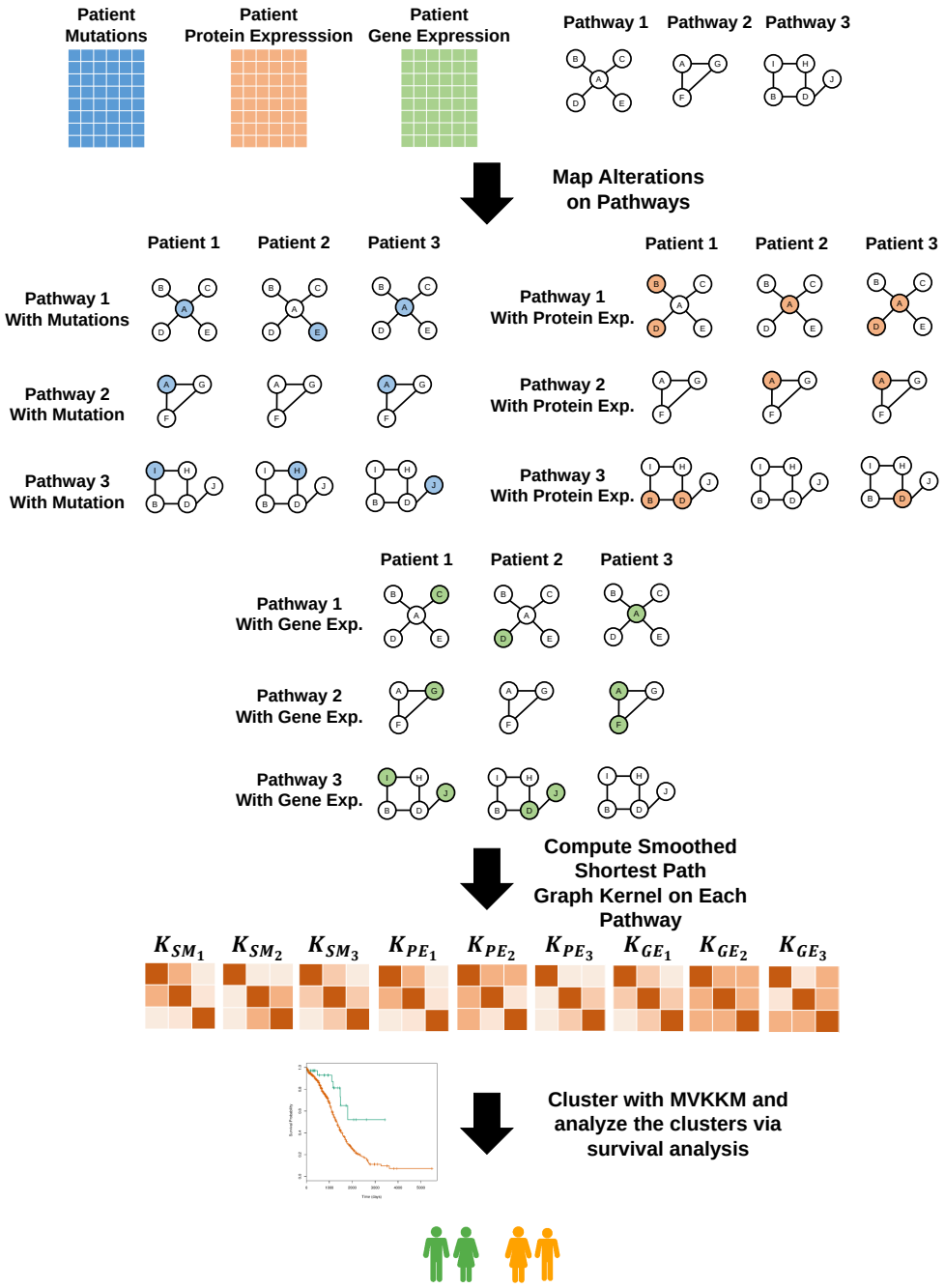


Figure 3.1: The proposed framework to stratify cancer patients into clinically meaningful subgroups. K_{X_i} represents the kernel matrix indicating similarities of patients based on i^{th} pathway on which X data type is mapped. SM stands for somatic mutations, PE stands for protein expression and GE stands for gene expression.

3.2 Data Curation and Processing

3.2.1 Molecular Patient Data

We use the molecular data made available through the TCGA project. The data are obtained through Synapse and is publicly available at `https://www.synapse.org/#!Synapse:syn395608`. Only the patient data for primary solid tumors are used. We utilize three different molecular data, details of which are provided below:

- **Somatic Mutations:** The data file *"KIRC-BCM-BI-UCSC-gapfill-v1.10.whitelist.maf"* is used (last update date: March 7 2013). It includes the information of altered genes of patients and the position of these alterations as well as their type. We exclude the data pertaining to non-solid tumors of patients from the initial set. The working data include 417 patients and 36353 genomic alterations, 11887 of which maps to a gene in KEGG pathways. The total number of mapped unique genes is 4379.

- **RNA Expresion Data:** The data file *"unc.edu_KIRC_IlluminaHiSeq_RNASeqV2.geneExp.whitelist_tumor"* (July 27 2012.) contains expression levels of RNA transcripts in cancer cells quantified using RNAseq. The data are RSEM normalized and they are available for 428 patients and 17682 genes,. Of these genes, 5348 of them are mapped on pathways. We define a gene as differentially expressed if the expression of that gene is one standard deviation higher or lower than the average expression of the gene in the samples.

- **Protein Expression Data:** The data file *"mdanderson.org_KIRC_MDA_RPPA_Core.whitelist_tumor"* is downloaded (latest update date: April 29 2013). Protein abundance is quantified through Reverse Phase Protein Array (RPPA). The data include 423 patients and expression data for 165 proteins and their phosphorylated versions. We ignore the phosphorylated proteins in our analysis, which include 130 proteins. Among these proteins,

117 of them can be mapped to at least one pathway. In a similar manner, we define a protein as differentially expressed if the expression of corresponding protein of that protein is one standard deviation higher or lower than the average expression of the protein in the samples.

3.2.2 Survival Data

To make survival analysis on the clustering result, we use the clinical data provided in the file: *"tcga_KIRC_clinical.aliquot.whitelist_tumor"* (latest update date: July 27 2012). For patients who passed away, their last known alive status are used. For patients with censored survival information, that is the patients are alive or passed away due to another reason during the study, the days to the last follow-up is used. For 457 patients, the survival time data is available.

3.2.3 Final Patient Set

We create the final patient set with patients that have the three types of molecular data and whose survival data are available. The final data set includes 361 kidney RCC cancer patients. 236 of them are right censored. The distribution of number of mutations across these 361 patients is given in Figure 3.2. Similarly, Figure 3.3 and 3.4 provide the distribution of number of alterations for differentially expressed genes and proteins, respectively.

3.2.4 Pathway Data

We download pathways from the KEGG database [40] on February 17 2016. Each data is parsed using KEGGParser [41]. We process each pathway such that only the genes are kept and the nodes that represent entity types such as ortholog, map and others and their relations are discarded. We treat compounds differently. KEGG compounds are collections of molecules such as lipids and sugars that are

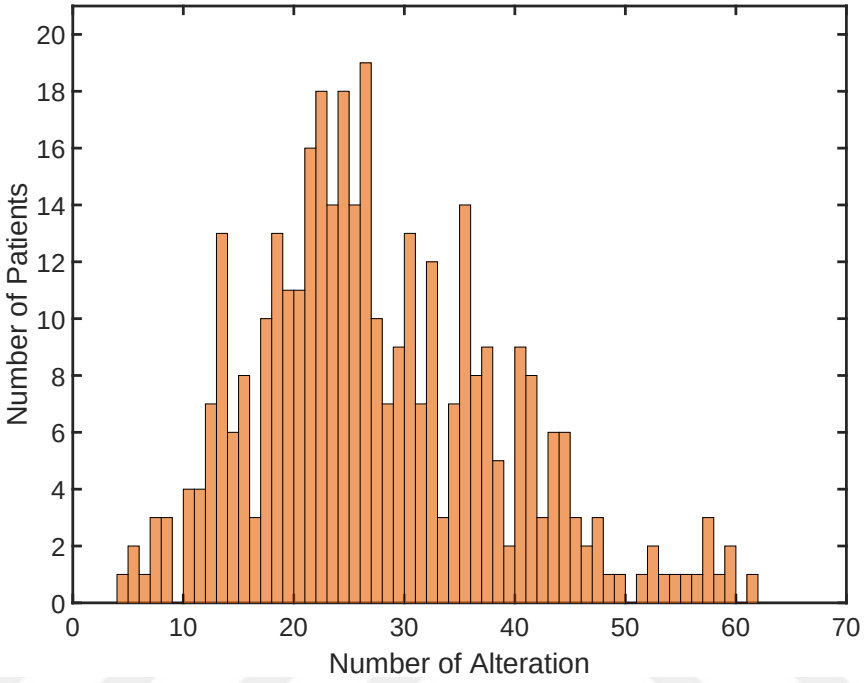


Figure 3.2: Histogram of number of mutated genes. The bin size is 1.

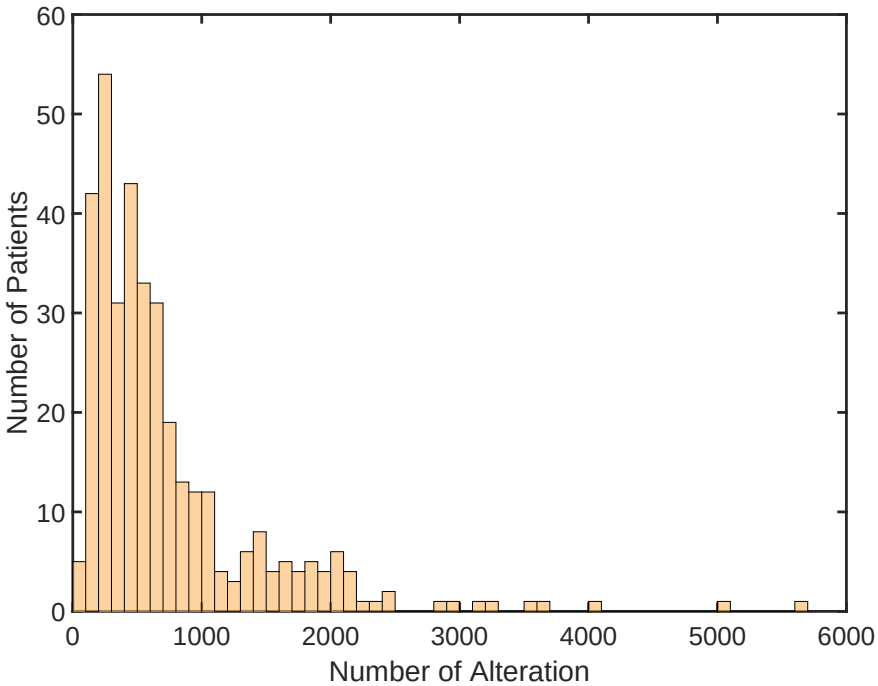


Figure 3.3: Histogram of number of differentially expressed genes. The bin size is 100.

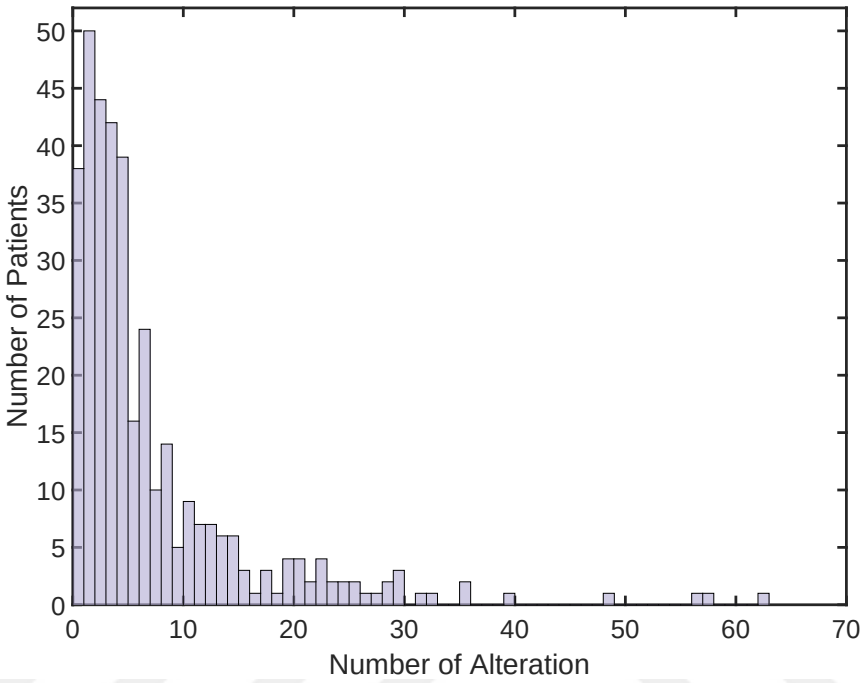


Figure 3.4: Histogram of number of differentially expressed proteins. The bin size is 1.

relevant to biological pathways. Often times if they interact with a gene product, there are edges in and out of the compound. Thus, when removing *compound* type entries, we insert new edges between genes to which the compound is connected so that information flow is preserved.

We eliminate pathways with less than 15 edges as well. Of the 264 pathways, 63 are excluded based on this criterion. Among 201 pathways, the largest pathway in terms of number of genes is the *hsa04060 Cytokine-cytokine receptor interaction* pathway with 295 genes whereas the smallest one is *hsa00591 Linoleic acid metabolism* pathway with 7 genes. In terms of number of edges, the largest pathway is *hsa00230 Purine metabolism* pathway with 257 edges. The smallest pathways are with 15 edges: *hsa00531 Glycosaminoglycan degradation*, *hsa00860 Porphyrin and chlorophyll metabolism*, *hsa04710 Circadian rhythm*, *hsa04970 Salivary secretion*, *hsa04976 Bile secretion*, *hsa05016 Huntington's disease* and *hsa05416 Viral myocarditis*. The mean values of the number of genes and the number of edges are 51.06 and 53.30 respectively. Each pathway is converted to

an undirected graph by ignoring the directionality of the edges.

3.3 Smoothed Shortest Path Graph Kernel

Shortest Path Graph Kernel is built upon the shortest path graph kernel [9] (details in section 2.5.2) .The shortest path graph kernel is designed for assessing similarities and differences of the graphs based on the topology. Here, though for a single pathway, we have identical graphs for each patient, but the node label distributions differ.

We would like the kernel function to compare pairs of graphs with the same topology but with different label distributions. The graph kernel function should reflect the similarities in the label distribution of patients using the graph as the context. For instance, despite the set of nonidentical altered genes, if two patients have alterations in genes in close proximity, they should influence their similarities. We would also want the central nodes to have more influence compared to genes that are in the periphery of the pathway. For example, consider four patients displayed in Figure 3.5. The alteration profile of patient 1 is different from the rest of the patients because she harbors mutations on a different part of the pathway. The other three patients are similar to each other, as they have identical or closeby altered genes. Among patients 2, 3 and 4, patient 2 and patient 4 bear the largest similarity as they are both mutated in a central gene (vertex 5) as opposed to a gene in the periphery of the pathway.

Formally, for each pathway i and patient j we define an undirected vertex labeled graph $G_i^{(j)} = (V_j, E, \ell)$. $V = \{v_1, v_2, \ldots, v_n\}$ is the ordered set of n genes in the pathway and $E \subset V \times V$ is a set of undirected edges between genes. The label set $\ell = \{l_1, l_2, \ldots, l_n\}$ is in the same order of V and represents the corresponding vertex's label. l is assigned based on patient's molecular alteration profile; if the corresponding gene is altered in patient i, label 1 is assigned and set 0 otherwise. Thus, graphs defined on the same pathway have the same topology but different node label distributions. The adjacency matrix for the graph is $n \times n$

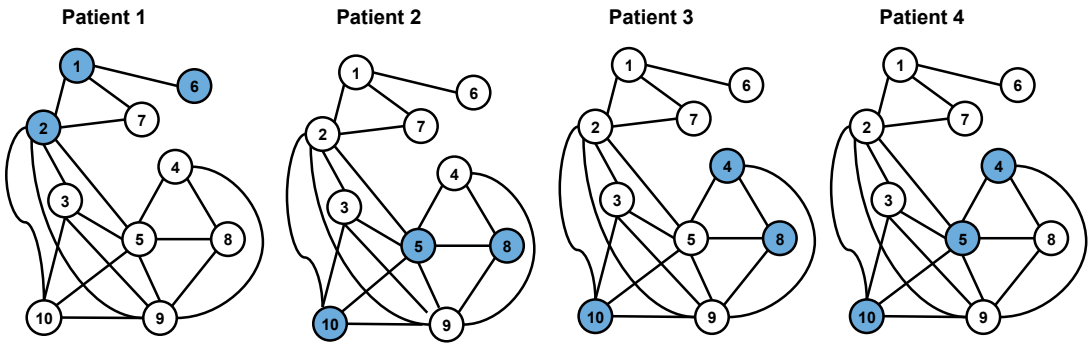


Figure 3.5: Mutational profiles of patients shown on an example undirected graph derived from the same pathway. Blue nodes indicate mutated genes and white nodes indicate unaltered genes.

matrix $\boldsymbol{A}$ with $\boldsymbol{A}_{ij} = 1$ if there is an edge between v_i and v_j, and 0 otherwise.

SMSPK compares patient genomic alterations along all the shortest paths of the undirected graph. For kernel function to consider changes in the neighbors, we first smooth the genomic alterations along the neighboring nodes. The utilized smoothing algorithm is as follows:

$$\boldsymbol{S}_{t+1} = \alpha \boldsymbol{S}_t \boldsymbol{A} + (1 - \alpha)\, \boldsymbol{S}_0 \tag{3.1}$$

where $\boldsymbol{S}_0$ is a patient-by-gene matrix which represents genomic alteration states of the genes in the graph and determined by $\boldsymbol{\ell}$. If there is a genomic alteration in the gene for a given patient, the entry is 1 and if there is no alteration the entry of the matrix is 0. $\boldsymbol{S}_t$ is the same matrix computed at time t. $\mathbf{A}$ is the degree normalized adjacency matrix. $\alpha \in [0, 1]$ is the parameter that defines the degree of smoothing. We iterate over propagation until convergence is attained.

Let $\boldsymbol{s}_p^{(i)}$ be the vector that represents the genomic alteration of patient i on shortest path p after smoothing. Let N be the number of shortest path on a graph g. smSPK for patients i and j for graph g is defined as follows:

$$K_g(i,j) = \sum_{p=1}^{N} \boldsymbol{s}_p^{(i)} . \boldsymbol{s}_p^{(j)T} \tag{3.2}$$

The above function is a valid kernel function, as the dot product is a linear kernel and kernel property is preserved under summation. When we apply this function on the example patients shown in Figure 3.5, there would be no significant similarity between patient 1 with the rest of the patients. Of the other three patients, the highest similarity was among Patient 2 and 4 as we would like to achieve. Applying this function on all patient pairs, we obtain a patient-by-patient kernel matrix $\boldsymbol{K}$ that reflects similarities of the patients on a single pathway.

The smSPK kernel matrices are evaluated for each of the KEGG pathway and for each data type and contains the patients similarities. The kernel matrices are input to multi-view kernel clustering, as described in the next section.

3.4 Multi-view Graph Kernel Clustering

Considering that kernel matrix of each pathway captures the similarity of a subset of patients under a different view, it is natural to use a multi-view kernel clustering approach to integrate the different views. In order to achieve this, we use Multi-view Kernel K-Means (MVKKM) proposed by Tzortzis et al. [10] to cluster patients based on the kernel matrices of pathways computed by smSPK. In the first step, we first compute kernel matrices on each KEGG pathway for each data type. Then we perform multi-view clustering of patients with each of these pathways.

We sum all kernel matrices of pathways by assigning equal weights. Then, we run kernel k-means with 20 random restarts [42] on this kernel matrix to obtain the initial clustering result required by MVKKM. Next, we input the kernel matrices of pathways along with the initial clustering result to MVKKM

in order to attain the final clustering result. We repeat this process 1000 times and we select the best one based on the clustering energy calculated by MVKKM. We perform clustering by varying $k = 2, 3, 4, 5$, and sparsity adjustment parameter p of MVKKM and the smoothing parameter α.

3.5 Survival Analysis of Clusters

To assess if the patients in the arrived clusters have different prognoses, we compare the survival distributions of the clusters using Kaplan-Meier survival curves [43] and log-rank test [44]. We conduct two analyses. In the first one, given all the groups, we test whether there is a statistical difference between the survival times. The null hypothesis states that the hazards are the same at all times for all pairs, while alternative hypothesis state that there is at least one pair of groups that differ. To understand where the difference should be attributed to, we conduct a second analysis, in which we compare one groups survival distribution with the rest of the patients. We utilize R survival package [45, 46] for carrying out both analyses.

Chapter 4

Results and Discussion

In this section, we introduce the set-up of the conducted experiments. Next, we present the results of our experiments with different settings. Once we have the results of the experiments, we discuss these results to analyze the performance of our proposed method for stratification of cancer patients.

4.1 Experimental Setup

In each method we cluster the patients into k groups for all $k \in \{2, 3, 4, 5\}$. We deploy several experimental settings to dissect the performance of the different aspects of the proposed algorithm.

1. **LMKKM with RBF Kernel in All Genes:** In this group of experiments a kernel matrix is computed using RBF kernel for each data type. Computations are carried out over all genes which have measurements in that particular data type. Three kernel matrices for each distinct data type are input to MVKMM.

2. **LMKKM with RBF Kernel with Pathway Genes:** This experiment type is similar to the previous one, with the difference that kernel matrix

for a given data type is only computed using genes that maps to at least one pathway. Three RBF kernel matrices for each distinct data type are input to MVKMM.

3. **MVKKM with RBF Kernel in All Genes:** In this experiment set, a kernel matrix is computed using RBF kernel for each data type. Computations use all the genes that have corresponding measurements in a particular data type. Three kernel matrices each for a different data type is input to LMKMM.

4. **MVKKM with RBF Kernel with Pathway Genes:** This experiment type is similar to the previous one, with the difference that kernel matrix for each data type is only computed using genes that maps to at least one pathway. Again, three RBF kernel matrices each for a different data type is input to LMKMM.

5. **MVKKM with Pathway Based Shortest Path Graph Kernels:** Smoothed shortest path kernel matrices are computed on each pathway and for each data type. There are 201 pathways and three different data types utilized, thus 603 kernel matrices are computed and input to MVKMM to obtain the clustering of patients.

6. **MVKKM with Pathway Based Shortest Path Graph Kernels along with RBF Kernels Using All Genes:** This framework is the same as the previous one, with one difference: RBF Kernels are computed for each data type and along with the 603 pathway kernels; the three resulting RBF kernels are input to MVKMM as well.

In MVKKM experiments, we set p as 1.3. A p value varying between 1.5 and 2 is suggested for five kernel matrices in the original paper, yet our number of kernels are 603 in our case. Therefore, we experiment with $p \in \{1.1, 1.3, 1.5\}$. Since the results do not improve with the choice of p from the specified set, we set it to 1.3. In running MVKKM with RBF kernel we try all $\sigma \in \{2^{-5}, 2^{-4}, \ldots, 2^{9}, 2^{10}\}$ and for all MVKKM experiments, we repeat the run with random restarts for 1000 times and pick the clustering with the lowest objective function. On the other

hand, in running LMKKM, σ values used are $\in \{2^{-5}, 2^{-4}, \ldots, 2^4, 2^5\}$. And we repeat the experiment 100 times with 100 optimization iterations. The reason we limit LMKMM with 100 and smaller set of σ is the long execution time. In smSPK experiments, we vary the smoothing parameter α and obtain results for all values $\in \{0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. For all k values, we select the best result with the minimum objective function value.

Note that we also attempted to use LMKKM as an alternative multi-view kernel clustering approach with the smSPK; however, due to large number of kernel matrices, 603, we were not able to conduct experiments due to prohibitive memory requirements.

4.2 Results

In this section, we demonstrate the results of different settings listed in section 4.1. At first, we present the experiments with LMKKM and RBF kernels computed on the molecular alteration data. These experiments show us the performance of cancer patients stratification when each sample in kernel matrices are weighted separately. A similar experimental layout is adopted for MVKKM. While conducting experiments with MVKKM, we assess how well the patients are clustered when we only use RBF kernels. We experiment with two settings here as well, one with all genes with an alteration and one with the subset of genes that are only part of the pathway. Finally, we present results of our proposed method, where we utilize the smSPK. We follow two different approaches. One strategy is to employ smSPK on each pathway separately and combine the resulting kernel matrices via MVKKM. The second approach is to utilize smSPKM on each pathway along with RBF kernel on all data types. The purpose is to evaluate whether we lose information by only using the pathways, as not every gene participate in a pathway.

4.2.1 LMKKM with RBF Kernel in All Genes

In this experiment group, we stratify patients using only alterations of patients without any pathway information using LMKKM. RBF kernel matrices are calculated for each data type and these kernel matrices are combined to cluster patients via LMKKM. In table 4.1, p-values of survival analysis of each setting are shown. Among all k values, the best p-value is retrieved for $k = 4$ when $\sigma = 4$. Kaplan-Meier plots of best results for each k value are demonstrated in Figure 4.1.

$\sigma \backslash\ k$	2	3	4	5
0.031	7.700e-02 (150-211)	2.539e-01 (214-106-41)	3.061e-01 (312-7-38-4)	4.072e-01 (51-39-129-9-133)
0.063	7.700e-02 (150-211)	2.539e-01 (214-106-41)	3.061e-01 (312-7-38-4)	4.072e-01 (51-39-129-9-133)
0.125	9.995e-01 (273-88)	2.625e-01 (118-83-160)	8.313e-01 (122-3-81-155)	9.779e-01 (80-92-35-88-66)
0.250	9.047e-01 (77-284)	3.923e-01 (13-202-146)	2.750e-01 (197-66-72-26)	7.205e-01 (73-192-31-53-12)
0.500	9.789e-01 (341-20)	2.804e-02 (49-301-11)	1.812e-02 (34-291-14-22)	6.364e-02 (284-11-46-10-10)
1.000	8.211e-02 (345-16)	3.387e-03 (326-19-16)	1.478e-02 (313-17-15-16)	1.629e-02 (11-17-302-16-15)
2.000	1.280e-01 (75-286)	1.214e-01 (257-38-66)	4.670e-04 (237-34-30-60)	1.113e-03 (217-48-34-24-38)
4.000	7.136e-01 (160-201)	9.615e-04 (140-187-34)	**5.126e-05** (120-53-27-161)	**1.222e-04** (22-49-50-140-100)
8.000	4.177e-01 (186-175)	4.325e-01 (108-138-115)	4.080e-01 (65-86-82-128)	5.436e-03 (62-61-73-64-101)
16.000	6.860e-01 (193-168)	1.017e-03 (123-152-86)	2.696e-03 (133-74-80-74)	3.296e-02 (65-63-73-122-38)
32.000	**3.535e-02** (141-220)	**6.311e-05** (102-97-162)	6.788e-02 (57-92-138-74)	2.747e-02 (57-75-114-61-54)

Table 4.1: Overall survival analysis of results of LMKKM on all genes of kidney cancer patients. The numbers in parenthesis indicate the cluster sizes. The bold p-values are the best values obtained for the corresponding k value.

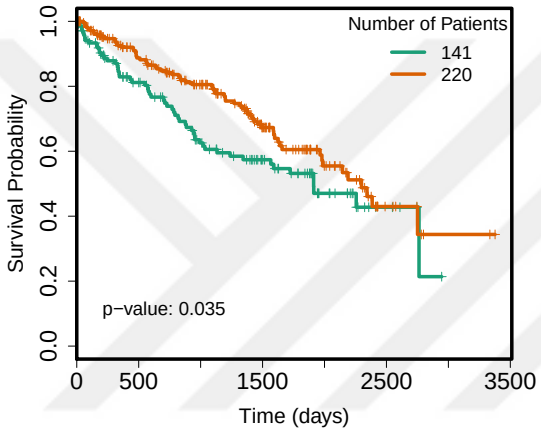
4.2.2 LMKKM with RBF Kernel with Pathway Genes

In this experiment, we exclusively utilize alterations on genes that are part of a pathway. For each selected genomic alterations of each data type, RBF kernel matrices are computed. Next, patients are clustered with LMKMM. In table 4.2, we demonstrate the obtained p-value of survival analysis of each setting. In this experiment, we attain the best p-value when $k = 3$ and $\sigma = 32$. Kaplan-Meier plots of best results for each k value is are given in Figure 4.2.

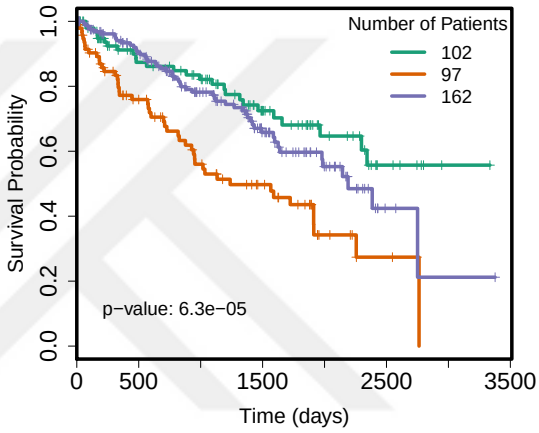
$\sigma \setminus k$	2	3	4	5
0.031	2.027e-01 (262-99)	5.503e-01 (72-51-238)	3.738e-01 (178-11-44-128)	6.895e-01 (116-75-44-103-23)
0.063	2.027e-01 (262-99)	5.503e-01 (72-51-238)	3.738e-01 (178-11-44-128)	6.895e-01 (116-75-44-103-23)
0.125	8.060e-02 (181-180)	5.353e-01 (124-88-149)	7.983e-01 (86-134-3-138)	7.631e-01 (78-92-85-103-3)
0.250	2.389e-01 (222-139)	4.685e-01 (49-229-83)	3.591e-01 (86-212-56-7)	8.513e-02 (205-10-8-69-69)
0.500	4.135e-03 (336-25)	4.487e-03 (315-30-16)	5.124e-01 (306-38-11-6)	5.932e-04 (293-18-16-14-20)
1.000	3.818e-01 (333-28)	4.181e-01 (299-49-13)	5.586e-01 (275-28-46-12)	5.413e-02 (236-35-16-46-28)
2.000	9.319e-01 (244-117)	4.035e-01 (223-107-31)	1.097e-02 (205-74-31-51)	1.705e-02 (186-30-62-44-39)
4.000	2.947e-01 (178-183)	6.235e-01 (122-116-123)	5.510e-02 (106-105-120-30)	4.962e-03 (97-97-20-116-31)
8.000	8.142e-01 (171-190)	1.683e-03 (101-125-135)	1.540e-02 (83-88-76-114)	1.175e-02 (73-81-81-91-35)
16.000	4.542e-01 (161-200)	4.949e-04 (150-108-103)	4.235e-03 (96-78-58-129)	7.017e-02 (102-71-85-42-61)
32.000	**3.431e-03** (230-131)	**1.578e-04** (143-91-127)	**3.395e-04** (44-76-121-120)	**3.886e-04** (54-106-60-106-35)

Table 4.2: Overall survival analysis of results of LMKKM on genes of kidney cancer patients which are in pathways. The numbers in parenthesis indicate the cluster sizes. The bold p-values are the best values obtained for the corresponding k value.

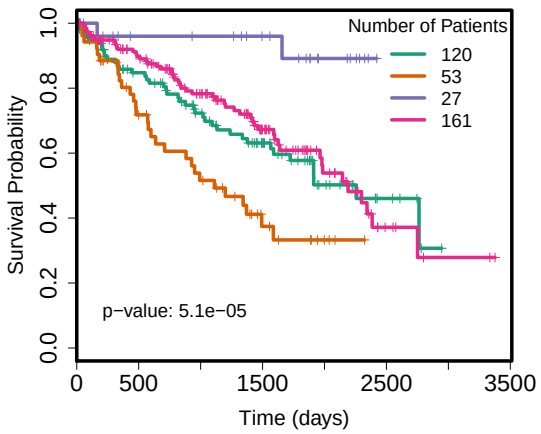
Figure 4.1: Kaplan-Meier plots of experiment utilizing LMKKM with RBF kernel on all genes



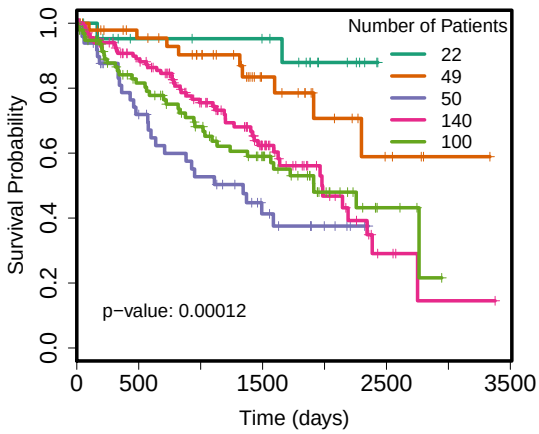
(a) The setting is $k = 2$ and $\sigma = 32$



(b) The setting is $k = 3$ and $\sigma = 32$

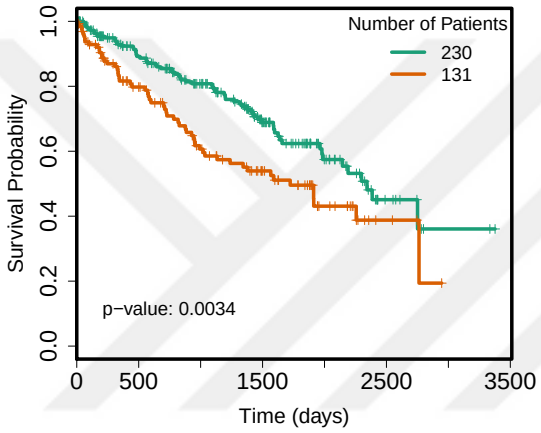


(c) The setting is $k = 4$ and $\sigma = 4$

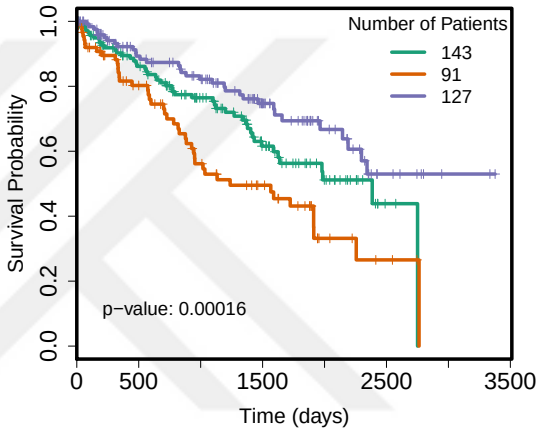


(d) The setting is $k = 5$ and $\sigma = 4$

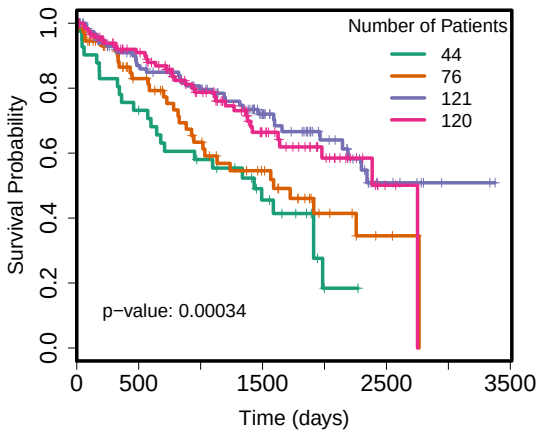
Figure 4.2: Kaplan-Meier plots of experiment utilizing LMKKM with RBF kernel on genes in pathways.



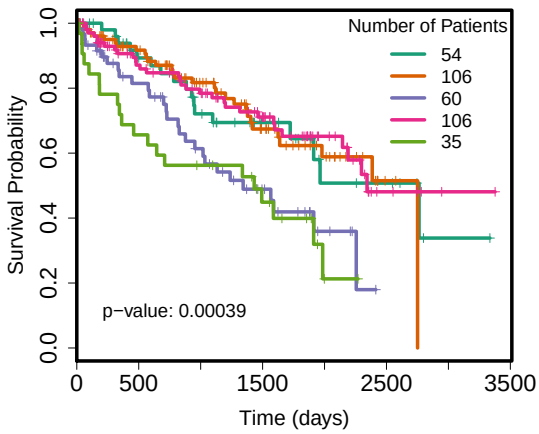
(a) The setting is $k = 2$ and $\sigma = 32$



(b) The setting is $k = 3$ and $\sigma = 32$



(c) The setting is $k = 4$ and $\sigma = 32$



(d) The setting is $k = 5$ and $\sigma = 32$

$\sigma \setminus k$	2	3	4	5
0.031	1.883e-01 (323-38)	3.063e-01 (38-221-102)	2.968e-01 (178-68-38-77)	3.726e-01 (66-38-142-70-45)
0.063	1.883e-01 (38-323)	4.147e-01 (38-230-93)	4.501e-01 (38-90-167-66)	6.666e-01 (38-68-54-153-48)
0.125	1.883e-01 (323-38)	2.610e-01 (231-38-92)	7.703e-03 (83-38-63-177)	3.527e-02 (49-38-61-61-152)
0.250	1.883e-01 (323-38)	3.606e-01 (91-232-38)	3.697e-01 (66-38-76-181)	4.974e-01 (38-44-150-71-58)
0.500	1.883e-01 (323-38)	9.853e-03 (38-109-214)	3.000e-01 (70-171-38-82)	7.163e-02 (134-56-60-73-38)
1.000	9.995e-01 (273-88)	3.861e-01 (38-229-94)	2.604e-01 (175-92-38-56)	3.932e-01 (45-38-52-52-174)
2.000	2.307e-01 (229-132)	1.090e-01 (97-132-132)	1.998e-01 (110-55-132-64)	3.880e-02 (78-12-131-48-92)
4.000	1.688e-02 (293-68)	4.442e-02 (21-268-72)	6.052e-02 (24-251-32-54)	6.643e-02 (70-228-29-13-21)
8.000	1.220e-01 (336-25)	3.839e-02 (18-305-38)	1.290e-01 (293-39-10-19)	1.063e-01 (33-12-20-286-10)
16.000	1.422e-01 (340-21)	5.251e-04 (146-15-200)	5.110e-05 (166-21-45-129)	2.977e-05 (18-128-86-11-118)
32.000	**7.488e-09** (65-296)	**7.922e-07** (14-67-280)	**3.267e-06** (13-66-272-10)	5.802e-04 (45-17-9-87-203)
64.000	2.748e-05 (316-45)	1.575e-03 (17-321-23)	1.069e-03 (30-32-17-282)	3.696e-03 (34-14-283-12-18)
128.000	2.660e-01 (18-343)	2.550e-03 (23-17-321)	6.300e-06 (37-292-12-20)	5.235e-04 (13-34-19-14-281)
256.000	2.660e-01 (18-343)	4.064e-03 (23-322-16)	1.973e-03 (295-31-15-20)	7.445e-05 (26-19-282-12-22)
512.000	2.660e-01 (343-18)	2.550e-03 (321-17-23)	3.028e-04 (34-292-15-20)	8.783e-06 (12-287-11-19-32)
1024.000	2.660e-01 (18-343)	4.064e-03 (322-23-16)	4.133e-04 (292-10-20-39)	**3.042e-06** (272-30-17-14-28)

Table 4.3: Overall survival analysis of results of MVKKM on all genes of kidney cancer patients. The numbers in parenthesis indicate the cluster sizes. The bold p-values are the best values obtained for the corresponding k value.

σ \ CI	1	2
0.03	1.9e-01 (323)	1.9e-01 (38)
0.06	1.9e-01 (38)	1.9e-01 (323)
0.13	1.9e-01 (323)	1.9e-01 (38)
0.25	1.9e-01 (323)	1.9e-01 (38)
0.50	1.9e-01 (323)	1.9e-01 (38)
1.00	1.0e+00 (273)	1.0e+00 (88)
2.00	2.3e-01 (229)	2.3e-01 (132)
4.00	1.7e-02 (293)	1.7e-02 (68)
8.00	1.2e-01 (336)	1.2e-01 (25)
16.00	1.4e-01 (340)	1.4e-01 (21)
32.00	7.5e-09 (65)	7.5e-09 (296)
64.00	2.7e-05 (316)	2.7e-05 (45)
128.00	2.7e-01 (18)	2.7e-01 (343)
256.00	2.7e-01 (18)	2.7e-01 (343)
512.00	2.7e-01 (343)	2.7e-01 (18)
1024.00	2.7e-01 (18)	2.7e-01 (343)

Table 4.4: One-vs-all survival analysis of results of RBF with MVKKM on all genes of kidney cancer patients for $k = 2$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

σ \ CI	1	2	3
0.03	1.9e-01 (38)	9.0e-01 (221)	2.8e-01 (102)
0.06	1.9e-01 (38)	4.4e-01 (230)	9.1e-01 (93)
0.13	8.7e-01 (231)	1.9e-01 (38)	2.2e-01 (92)
0.25	4.1e-01 (91)	8.7e-01 (232)	1.9e-01 (38)
0.50	1.9e-01 (38)	3.1e-03 (109)	7.2e-02 (214)
1.00	1.9e-01 (38)	7.9e-01 (229)	4.9e-01 (94)
2.00	4.6e-02 (97)	7.9e-01 (132)	1.3e-01 (132)
4.00	4.2e-01 (21)	1.3e-02 (268)	2.4e-02 (72)
8.00	2.7e-01 (18)	1.1e-02 (305)	2.6e-02 (38)
16.00	2.6e-04 (146)	4.4e-01 (15)	1.0e-04 (200)
32.00	2.3e-01 (14)	4.3e-07 (67)	1.7e-07 (280)
64.00	2.5e-01 (17)	8.1e-04 (321)	8.1e-04 (23)
128.00	1.4e-03 (23)	2.5e-01 (17)	1.2e-03 (321)
256.00	1.4e-03 (23)	3.0e-03 (322)	4.7e-01 (16)
512.00	1.2e-03 (321)	2.5e-01 (17)	1.4e-03 (23)
1024.00	3.0e-03 (322)	1.4e-03 (23)	4.7e-01 (16)

Table 4.5: One-vs-all survival analysis of results of RBF with MVKKM on all genes of kidney cancer patients for $k = 3$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

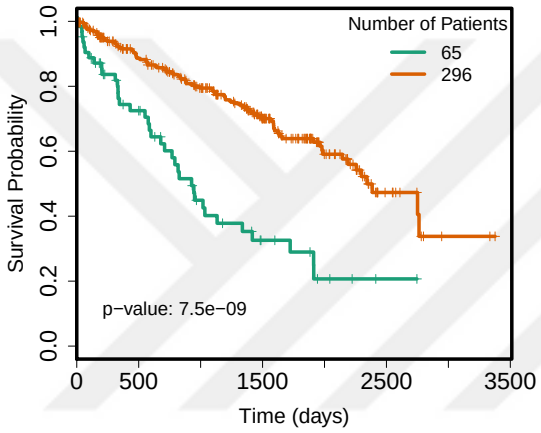
σ \ CI	1	2	3	4
0.03	2.4e-01 (178)	2.6e-01 (68)	1.9e-01 (38)	4.7e-01 (77)
0.06	1.9e-01 (38)	6.1e-01 (90)	3.2e-01 (167)	4.9e-01 (66)
0.13	3.2e-01 (83)	1.9e-01 (38)	3.6e-02 (63)	7.7e-04 (177)
0.25	4.2e-01 (66)	1.9e-01 (38)	9.5e-01 (76)	1.5e-01 (181)
0.50	8.3e-01 (70)	7.4e-01 (171)	1.9e-01 (38)	1.1e-01 (82)
1.00	5.6e-01 (175)	3.8e-01 (92)	1.9e-01 (38)	1.5e-01 (56)
2.00	2.1e-01 (110)	9.4e-01 (55)	7.9e-01 (132)	4.1e-02 (64)
4.00	3.8e-01 (24)	4.2e-02 (251)	1.5e-02 (32)	8.4e-01 (54)
8.00	2.8e-02 (293)	3.2e-02 (39)	4.4e-01 (10)	7.1e-01 (19)
16.00	2.8e-01 (166)	1.8e-02 (21)	1.1e-04 (45)	2.8e-02 (129)
32.00	2.9e-04 (13)	2.9e-04 (66)	7.7e-06 (272)	8.8e-01 (10)
64.00	6.9e-02 (30)	1.9e-03 (32)	2.5e-01 (17)	1.1e-04 (282)
128.00	2.0e-04 (37)	3.0e-07 (292)	1.1e-02 (12)	4.2e-02 (20)
256.00	1.7e-04 (295)	3.0e-03 (31)	1.0e-01 (15)	1.7e-01 (20)
512.00	1.5e-03 (34)	1.6e-05 (292)	9.5e-02 (15)	4.2e-02 (20)
1024.00	2.3e-05 (292)	8.9e-02 (10)	4.2e-02 (20)	2.2e-03 (39)

Table 4.6: One-vs-all survival analysis of results of RBF with MVKKM on all genes of kidney cancer patients for $k = 4$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

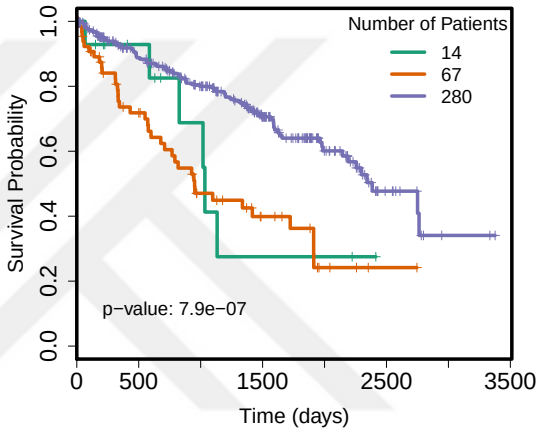
σ \ CI	1	2	3	4	5
0.03	9.5e-01 (66)	1.9e-01 (38)	6.7e-02 (142)	5.8e-01 (70)	5.4e-01 (45)
0.06	1.9e-01 (38)	8.5e-01 (68)	6.8e-01 (54)	2.8e-01 (153)	9.1e-01 (48)
0.13	2.0e-02 (49)	1.9e-01 (38)	5.4e-01 (61)	1.2e-01 (61)	8.3e-02 (152)
0.25	1.9e-01 (38)	4.9e-01 (44)	2.4e-01 (150)	3.9e-01 (71)	9.0e-01 (58)
0.50	6.6e-01 (134)	6.7e-03 (56)	4.4e-01 (60)	9.0e-01 (73)	1.9e-01 (38)
1.00	6.7e-01 (45)	1.9e-01 (38)	3.1e-01 (52)	1.9e-01 (52)	6.9e-01 (174)
2.00	1.2e-02 (78)	3.9e-02 (12)	9.9e-01 (131)	5.3e-01 (48)	3.8e-01 (92)
4.00	6.6e-01 (70)	1.2e-01 (228)	1.2e-02 (29)	1.8e-01 (13)	7.5e-01 (21)
8.00	8.7e-01 (33)	1.4e-02 (12)	4.6e-01 (20)	8.0e-02 (286)	4.4e-01 (10)
16.00	4.1e-01 (18)	2.3e-02 (128)	5.0e-06 (86)	5.7e-02 (11)	6.9e-02 (118)
32.00	2.5e-03 (45)	4.4e-02 (17)	7.8e-01 (9)	1.7e-01 (87)	8.8e-05 (203)
64.00	9.1e-03 (34)	3.9e-02 (14)	1.1e-04 (283)	3.3e-01 (12)	2.0e-01 (18)
128.00	1.4e-02 (13)	1.3e-03 (34)	3.0e-01 (19)	2.3e-01 (14)	1.5e-03 (281)
256.00	2.8e-04 (26)	2.2e-01 (19)	5.2e-04 (282)	3.3e-01 (12)	6.2e-03 (22)
512.00	3.4e-03 (12)	1.0e-04 (287)	5.3e-01 (11)	5.2e-01 (19)	2.0e-05 (32)
1024.00	4.3e-05 (272)	4.9e-06 (30)	2.4e-01 (17)	2.3e-01 (14)	2.0e-02 (28)

Table 4.7: One-vs-all survival analysis of results of RBF with MVKKM on all genes of kidney cancer patients for $k = 5$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

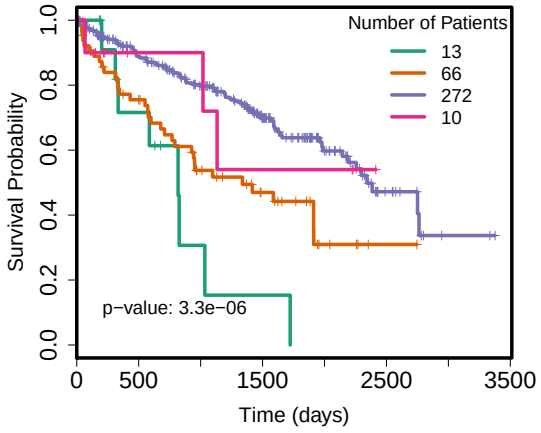
Figure 4.3: Kaplan-Meier plots of experiment utilizing MVKKM with RBF kernel on all genes



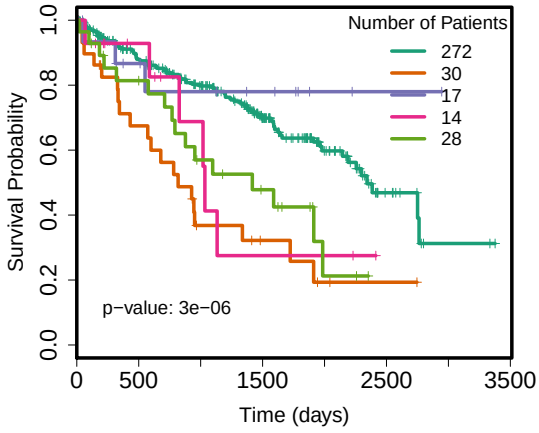
(a) The setting is $k = 2$ and $\sigma = 32$



(b) The setting is $k = 3$ and $\sigma = 32$



(c) The setting is $k = 4$ and $\sigma = 32$



(d) The setting is $k = 5$ and $\sigma = 1024$

4.2.3 MVKKM with RBF Kernel in All Genes

We conduct the experiments by using molecular alteration data from all recorded genes. RBF kernel matrices are calculated for each data type. Then, we input these kernel matrices to MVKKM to obtain the groups of patients. In table 4.3, we provide p-value of survival analysis of each setting. Among all k values, the best p-value is attained for $k = 2$ when $\sigma = 32$. In addition to overall survival analysis, we demonstrate the one-vs-all survival analysis of identified clusters for each setting. We show the one-vs-all survival analysis results for each k value in a separate table. Tables 4.4, 4.5, 4.6 and 4.7 display the results. Kaplan-Meier plots of best results for each k value is given in Figure 4.3.

4.2.4 MVKKM with RBF Kernel with Pathway Genes

RBF kernel matrices are calculated for each data type using only the genes that are part of at least one pathway. The kernel matrices are used in MVKKM to obtain the groups of patients. In table 4.8, we provide p-value of survival analysis of each setting. Among all the k values, the best p-value is retrieved for $k = 4$ when $\sigma = 256$. In addition to overall survival analysis, we demonstrate the one-vs-all survival analysis of identified clusters for each setting. We show the one-vs-all survival analysis results for each k value in a separate table. Tables 4.9, 4.10, 4.11 and 4.12 display the results. Kaplan-Meier plots of best results for each k value is given in Figure 4.4.

$\sigma \backslash\ k$	2	3	4	5
0.031	6.148e-01 (317-44)	2.999e-01 (95-222-44)	3.374e-01 (60-44-170-87)	9.240e-01 (60-44-47-74-136)
0.063	6.148e-01 (44-317)	2.992e-01 (219-44-98)	8.374e-01 (65-178-44-74)	8.757e-01 (146-63-45-44-63)
0.125	6.148e-01 (44-317)	1.978e-01 (98-219-44)	6.714e-01 (44-68-176-73)	7.916e-01 (44-49-144-57-67)
0.250	6.148e-01 (44-317)	5.699e-01 (44-98-219)	7.613e-01 (44-59-82-176)	7.119e-01 (128-74-65-44-50)
0.500	6.148e-01 (44-317)	3.149e-01 (212-105-44)	8.321e-01 (44-79-140-98)	8.901e-01 (44-136-56-72-53)
1.000	6.717e-01 (267-94)	7.182e-01 (211-44-106)	4.818e-01 (44-89-52-176)	7.788e-01 (61-31-62-44-163)
2.000	9.012e-02 (246-115)	1.326e-02 (82-129-150)	3.554e-01 (69-60-111-121)	3.034e-02 (23-139-39-90-70)
4.000	3.241e-03 (298-63)	4.663e-02 (67-275-19)	1.527e-02 (19-40-44-258)	4.139e-02 (34-71-21-220-15)
8.000	**2.952e-06** (265-96)	2.185e-04 (177-19-165)	9.281e-02 (21-11-36-293)	7.846e-02 (13-10-281-40-17)
16.000	1.339e-05 (65-296)	**2.158e-05** (76-12-273)	6.341e-07 (17-90-236-18)	1.195e-06 (13-81-9-226-32)
32.000	3.161e-05 (47-314)	3.203e-03 (13-34-314)	1.365e-04 (33-272-16-40)	3.338e-05 (21-33-18-274-15)
64.000	1.230e-01 (19-342)	2.490e-03 (22-15-324)	4.772e-07 (25-30-287-19)	9.287e-06 (274-18-15-23-31)
128.000	1.230e-01 (342-19)	1.215e-03 (323-15-23)	1.611e-07 (32-19-31-279)	**4.050e-08** (15-273-19-23-31)
256.000	1.230e-01 (342-19)	2.774e-03 (324-13-24)	**2.224e-08** (17-284-26-34)	7.879e-06 (276-27-21-15-22)
512.000	1.230e-01 (19-342)	1.215e-03 (15-323-23)	3.660e-03 (11-11-319-20)	3.137e-06 (279-13-22-28-19)
1024.000	1.230e-01 (19-342)	7.637e-03 (13-25-323)	2.002e-06 (287-30-16-28)	3.398e-05 (16-23-15-274-33)

Table 4.8: Overall survival analysis of results of MVKKM on pathway genes of kidney cancer patients. The numbers in parenthesis indicate the cluster sizes. The bold p-values are the best values obtained for the corresponding k value.

σ \ CI	1	2
0.03	6.1e-01 (317)	6.1e-01 (44)
0.06	6.1e-01 (44)	6.1e-01 (317)
0.13	6.1e-01 (44)	6.1e-01 (317)
0.25	6.1e-01 (44)	6.1e-01 (317)
0.50	6.1e-01 (44)	6.1e-01 (317)
1.00	6.7e-01 (267)	6.7e-01 (94)
2.00	9.0e-02 (246)	9.0e-02 (115)
4.00	3.2e-03 (298)	3.2e-03 (63)
8.00	3.0e-06 (265)	3.0e-06 (96)
16.00	1.3e-05 (65)	1.3e-05 (296)
32.00	3.2e-05 (47)	3.2e-05 (314)
64.00	1.2e-01 (19)	1.2e-01 (342)
128.00	1.2e-01 (342)	1.2e-01 (19)
256.00	1.2e-01 (342)	1.2e-01 (19)
512.00	1.2e-01 (19)	1.2e-01 (342)
1024.00	1.2e-01 (19)	1.2e-01 (342)

Table 4.9: One-vs-all survival analysis of results of RBF with MVKKM on genes of kidney cancer patients which are in pathways for $k = 2$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

σ \ CI	1	2	3
0.03	1.9e-01 (95)	1.3e-01 (222)	6.1e-01 (44)
0.06	1.2e-01 (219)	6.1e-01 (44)	2.0e-01 (98)
0.13	1.2e-01 (98)	7.8e-02 (219)	6.1e-01 (44)
0.25	6.1e-01 (44)	4.4e-01 (98)	2.9e-01 (219)
0.50	1.3e-01 (212)	2.1e-01 (105)	6.1e-01 (44)
1.00	4.3e-01 (211)	6.1e-01 (44)	6.2e-01 (106)
2.00	3.8e-03 (82)	1.2e-01 (129)	4.3e-01 (150)
4.00	3.1e-02 (67)	1.3e-02 (275)	3.2e-01 (19)
8.00	1.1e-04 (177)	3.2e-02 (19)	2.0e-03 (165)
16.00	3.6e-05 (76)	7.3e-02 (12)	3.8e-06 (273)
32.00	2.1e-01 (13)	2.0e-03 (34)	8.3e-04 (314)
64.00	2.9e-03 (22)	1.0e-01 (15)	6.3e-04 (324)
128.00	3.1e-04 (323)	1.0e-01 (15)	1.4e-03 (23)
256.00	9.0e-04 (324)	2.1e-01 (13)	1.7e-03 (24)
512.00	1.0e-01 (15)	3.1e-04 (323)	1.4e-03 (23)
1024.00	2.1e-01 (13)	5.0e-03 (25)	2.2e-03 (323)

Table 4.10: One-vs-all survival analysis of results of RBF with MVKKM on genes of kidney cancer patients which are in pathways for $k = 3$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

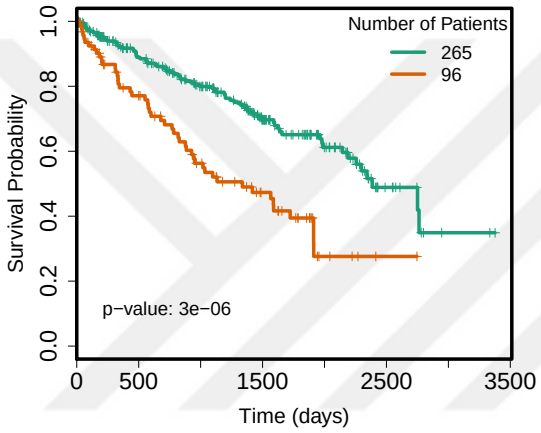
σ \ CI	1	2	3	4
0.03	1.1e-01 (60)	6.1e-01 (44)	1.9e-01 (170)	7.3e-01 (87)
0.06	5.7e-01 (65)	4.0e-01 (178)	6.1e-01 (44)	9.3e-01 (74)
0.13	6.1e-01 (44)	8.1e-01 (68)	2.3e-01 (176)	4.2e-01 (73)
0.25	6.1e-01 (44)	7.5e-01 (59)	3.6e-01 (82)	5.3e-01 (176)
0.50	6.1e-01 (44)	9.0e-01 (79)	7.4e-01 (140)	3.8e-01 (98)
1.00	6.1e-01 (44)	1.9e-01 (89)	7.7e-01 (52)	2.0e-01 (176)
2.00	9.5e-01 (69)	8.1e-02 (60)	4.9e-01 (111)	5.0e-01 (121)
4.00	1.2e-01 (19)	2.4e-02 (40)	2.5e-01 (44)	2.2e-03 (258)
8.00	4.3e-01 (21)	5.3e-01 (11)	2.6e-02 (36)	1.6e-02 (293)
16.00	3.6e-02 (17)	1.2e-02 (90)	2.7e-06 (236)	5.1e-05 (18)
32.00	9.6e-03 (33)	7.3e-06 (272)	2.3e-01 (16)	2.0e-03 (40)
64.00	5.4e-03 (25)	1.4e-05 (30)	3.6e-08 (287)	1.2e-01 (19)
128.00	1.7e-05 (32)	1.2e-01 (19)	2.1e-03 (31)	1.0e-08 (279)
256.00	3.6e-02 (17)	9.5e-10 (284)	4.5e-03 (26)	2.9e-06 (34)
512.00	1.4e-02 (11)	5.3e-01 (11)	6.1e-04 (319)	1.1e-02 (20)
1024.00	1.5e-07 (287)	7.0e-05 (30)	2.3e-01 (16)	2.8e-03 (28)

Table 4.11: One-vs-all survival analysis of results of RBF with MVKKM on genes of kidney cancer patients which are in pathways for $k = 4$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

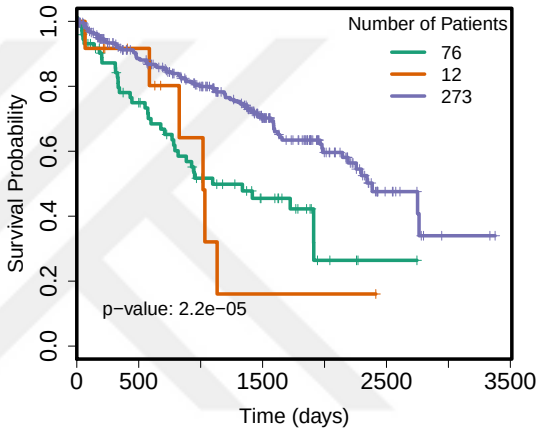
σ \ CI	1	2	3	4	5
0.03	5.5e-01 (60)	6.1e-01 (44)	8.5e-01 (47)	6.7e-01 (74)	5.7e-01 (136)
0.06	6.9e-01 (146)	3.9e-01 (63)	6.7e-01 (45)	6.1e-01 (44)	6.9e-01 (63)
0.13	6.1e-01 (44)	2.3e-01 (49)	9.8e-01 (144)	6.0e-01 (57)	9.6e-01 (67)
0.25	1.7e-01 (128)	7.6e-01 (74)	4.2e-01 (65)	6.1e-01 (44)	8.6e-01 (50)
0.50	6.1e-01 (44)	3.2e-01 (136)	9.7e-01 (56)	6.2e-01 (72)	7.6e-01 (53)
1.00	8.1e-01 (61)	2.7e-01 (31)	9.7e-01 (62)	6.1e-01 (44)	3.9e-01 (163)
2.00	2.8e-01 (23)	3.5e-01 (139)	7.9e-02 (39)	6.3e-01 (90)	6.2e-03 (70)
4.00	5.7e-01 (34)	3.9e-01 (71)	2.3e-01 (21)	2.7e-01 (220)	6.5e-03 (15)
8.00	1.0e-01 (13)	4.4e-01 (10)	2.5e-02 (281)	7.1e-01 (40)	3.9e-02 (17)
16.00	4.5e-02 (13)	1.5e-02 (81)	7.8e-01 (9)	3.9e-07 (226)	3.6e-05 (32)
32.00	6.2e-04 (21)	9.9e-04 (33)	8.6e-01 (18)	6.5e-06 (274)	4.4e-01 (15)
64.00	2.9e-06 (274)	9.0e-01 (18)	4.4e-01 (15)	5.8e-03 (23)	2.6e-05 (31)
128.00	4.4e-01 (15)	3.8e-07 (273)	6.0e-01 (19)	1.3e-03 (23)	3.1e-07 (31)
256.00	1.7e-05 (276)	5.5e-04 (27)	7.8e-01 (21)	4.4e-01 (15)	1.8e-04 (22)
512.00	9.5e-06 (279)	2.1e-01 (13)	2.6e-05 (22)	2.6e-03 (28)	5.3e-01 (19)
1024.00	6.9e-01 (16)	7.3e-05 (23)	4.4e-01 (15)	3.6e-05 (274)	7.3e-03 (33)

Table 4.12: One-vs-all survival analysis of results of RBF with MVKKM on genes of kidney cancer patients which are in pathways for $k = 5$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

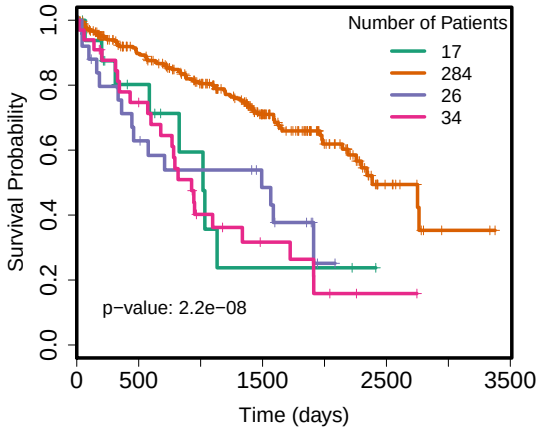
Figure 4.4: Kaplan-Meier plots of experiment utilizing MVKKM with RBF kernel on genes in pathways



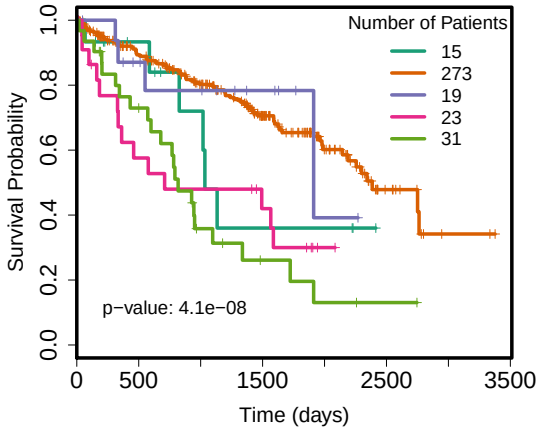
(a) The setting is $k = 2$ and $\sigma = 8$



(b) The setting is $k = 3$ and $\sigma = 16$



(c) The setting is $k = 4$ and $\sigma = 256$



(d) The setting is $k = 5$ and $\sigma = 128$

$\alpha \backslash\ k$	2	3	4	5
0.00	**6.1e-08** (125-236)	4.1e-03 (86-191-84)	8.7e-06 (76-99-71-115)	5.7e-02 (78-164-24-71-24)
0.01	2.1e-07 (129-232)	4.6e-01 (87-156-118)	**2.1e-06** (109-91-80-81)	1.8e-02 (88-68-65-75-65)
0.05	1.1e-04 (230-131)	2.6e-02 (88-187-86)	4.7e-04 (126-56-104-75)	6.3e-06 (71-78-78-50-84)
0.10	4.1e-05 (245-116)	1.6e-05 (179-90-92)	7.6e-06 (124-86-71-80)	6.7e-07 (66-80-87-57-71)
0.20	3.9e-06 (126-235)	1.9e-07 (84-81-196)	1.4e-04 (104-83-87-87)	1.8e-03 (63-204-31-39-24)
0.30	6.4e-08 (237-124)	4.1e-07 (93-159-109)	7.7e-06 (90-95-101-75)	5.7e-05 (125-57-71-64-44)
0.40	1.3e-04 (124-237)	5.9e-06 (148-99-114)	3.6e-05 (130-78-79-74)	1.3e-05 (47-59-127-61-67)
0.50	2.0e-05 (115-246)	1.1e-05 (106-86-169)	6.7e-04 (77-76-132-76)	3.7e-02 (59-68-74-57-103)
0.60	4.6e-05 (259-102)	8.5e-04 (90-164-107)	2.0e-05 (135-90-63-73)	2.1e-04 (64-73-61-65-98)
0.70	2.6e-04 (102-259)	**1.5e-08** (82-88-191)	4.7e-06 (83-72-129-77)	2.0e-04 (77-90-48-72-74)
0.80	1.6e-03 (112-249)	7.3e-06 (83-83-195)	4.9e-04 (79-96-103-83)	1.8e-05 (86-71-80-50-74)
0.90	7.5e-06 (254-107)	6.7e-03 (181-90-90)	6.3e-03 (78-67-129-87)	**5.7e-07** (64-67-75-91-64)

Table 4.13: Overall survival analysis of results of smSPK on all pathways of kidney cancer patients. The numbers in parenthesis indicate the cluster sizes. The bold p-values are the best values obtained for the corresponding k value.

CI \ α	1	2
0.05	1.1e-04 (230)	1.1e-04 (131)
0.10	4.1e-05 (245)	4.1e-05 (116)
0.20	3.9e-06 (126)	3.9e-06 (235)
0.30	6.4e-08 (237)	6.4e-08 (124)
0.40	1.3e-04 (124)	1.3e-04 (237)
0.50	2.0e-05 (115)	2.0e-05 (246)
0.60	4.6e-05 (259)	4.6e-05 (102)
0.70	2.6e-04 (102)	2.6e-04 (259)
0.80	1.6e-03 (112)	1.6e-03 (249)
0.90	7.5e-06 (254)	7.5e-06 (107)

Table 4.14: One-vs-all survival analysis of results of smSPK on all pathways of kidney cancer patients for $k = 2$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

CI \ α	1	2	3
0.05	7.8e-03 (88)	2.7e-01 (187)	2.1e-01 (86)
0.10	1.2e-03 (179)	4.5e-06 (90)	6.4e-01 (92)
0.20	3.3e-08 (84)	4.2e-01 (81)	2.9e-04 (196)
0.30	5.7e-02 (93)	2.7e-03 (159)	5.8e-08 (109)
0.40	3.5e-04 (148)	4.8e-01 (99)	2.0e-06 (114)
0.50	6.1e-06 (106)	9.6e-01 (86)	1.3e-04 (169)
0.60	8.8e-02 (90)	8.1e-02 (164)	1.9e-04 (107)
0.70	3.4e-01 (82)	4.8e-08 (88)	3.1e-07 (191)
0.80	2.2e-06 (83)	1.7e-02 (83)	1.1e-01 (195)
0.90	1.3e-01 (181)	2.0e-01 (90)	1.7e-03 (90)

Table 4.15: One-vs-all survival analysis of results of smSPK on all pathways of kidney cancer patients for $k = 3$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

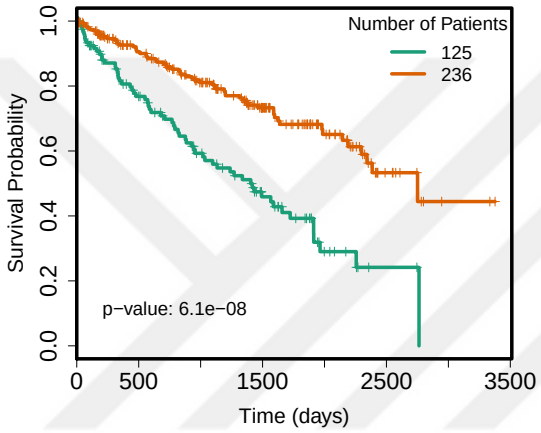
CI \ α	1	2	3	4
0.05	6.1e-01 (126)	9.2e-05 (56)	8.7e-02 (104)	1.1e-01 (75)
0.10	1.5e-01 (124)	2.9e-05 (86)	1.3e-04 (71)	7.6e-01 (80)
0.20	5.7e-01 (104)	1.4e-03 (83)	8.3e-01 (87)	6.2e-05 (87)
0.30	2.2e-02 (90)	5.2e-01 (95)	2.2e-01 (101)	4.9e-07 (75)
0.40	2.9e-01 (130)	1.3e-02 (78)	7.8e-01 (79)	4.3e-06 (74)
0.50	2.7e-01 (77)	4.8e-04 (76)	4.3e-03 (132)	2.1e-01 (76)
0.60	1.2e-01 (135)	6.7e-02 (90)	5.6e-01 (63)	9.2e-07 (73)
0.70	1.8e-07 (83)	2.0e-01 (72)	1.0e-01 (129)	9.9e-02 (77)
0.80	2.0e-02 (79)	1.8e-01 (96)	8.6e-01 (103)	1.0e-04 (83)
0.90	4.8e-04 (78)	5.6e-01 (67)	1.1e-01 (129)	4.4e-01 (87)

Table 4.16: One-vs-all survival analysis of results of smSPK on all pathways of kidney cancer patients for $k = 4$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

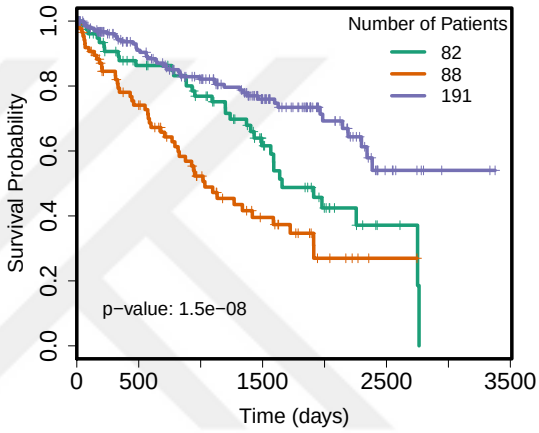
CI \ α	1	2	3	4	5
0.05	8.6e-01 (71)	3.4e-01 (78)	5.7e-01 (78)	1.5e-07 (50)	2.9e-02 (84)
0.10	3.0e-08 (66)	6.2e-02 (80)	2.7e-02 (87)	2.9e-01 (57)	4.5e-01 (71)
0.20	1.4e-01 (63)	1.0e-02 (204)	3.6e-01 (31)	3.4e-04 (39)	8.5e-01 (24)
0.30	4.8e-02 (125)	6.6e-02 (57)	4.2e-06 (71)	3.9e-01 (64)	3.2e-01 (44)
0.40	7.6e-01 (47)	7.5e-07 (59)	2.3e-01 (127)	1.2e-02 (61)	6.9e-01 (67)
0.50	1.9e-01 (59)	8.3e-02 (68)	5.9e-03 (74)	6.6e-01 (57)	9.8e-01 (103)
0.60	8.7e-06 (64)	4.8e-01 (73)	8.8e-01 (61)	8.8e-01 (65)	1.5e-02 (98)
0.70	1.6e-04 (77)	4.8e-04 (90)	2.2e-01 (48)	9.6e-01 (72)	6.5e-01 (74)
0.80	1.1e-01 (86)	3.8e-02 (71)	1.5e-01 (80)	2.3e-01 (50)	1.9e-06 (74)
0.90	8.7e-02 (64)	2.9e-01 (67)	7.4e-08 (75)	1.7e-01 (91)	8.1e-03 (64)

Table 4.17: One-vs-all survival analysis of results of smSPK on all pathways of kidney cancer patients for $k = 5$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

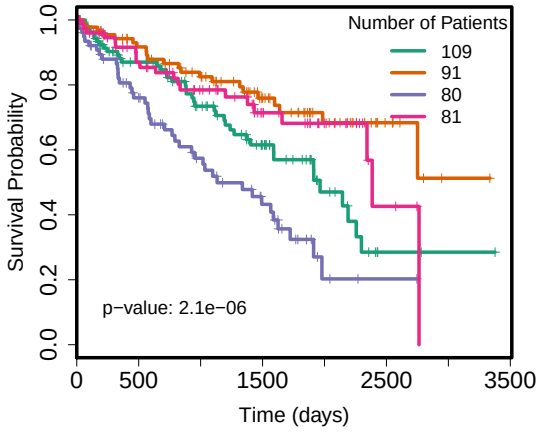
Figure 4.5: Kaplan-Meier plots of experiment utilizing smSPK on all pathways



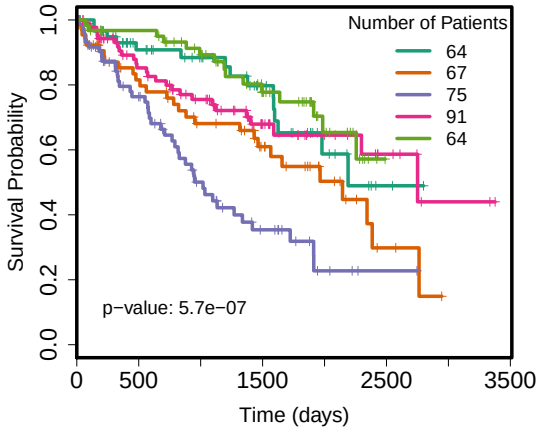
(a) The setting is $k = 2$ and $\alpha = 0$



(b) The setting is $k = 3$ and $\alpha = 0.7$



(c) The setting is $k = 4$ and $\alpha = 0.01$



(d) The setting is $k = 5$ and $\alpha = 0.9$

4.2.5 Multi-view Kernel Clustering with Pathway Based Shortest Path Graph Kernels

The smSPK kernels are computed for each data type and pathway. These 60 kernel matrices are used as input to MVKKM. The results of survival analysis for each setting are given in table 4.13. In addition to the overall survival analysis results, we present one-vs-all survival analysis results for each k value in tables 4.14, 4.15, 4.16 and 4.17. The best p-value is obtained when $k = 3$ and $\alpha = 0.7$. Kaplan-Meier plots of best results for each k value are given in Figure 4.5.

4.2.6 Multi-view Kernel Clustering with Pathway Based Shortest Path Graph Kernels along with RBF Kernels Using All Genes

This experiment is similar to the previous experiment except for the three additional RBF kernels computed in each data type. In computing the kernels, all the genes recorded for the respective data type are utilized. Once we calculate the kernel matrices of pathways, we input them along with RBF kernels on each data type to MVKKM. The result of survival analysis of each setting is given in table 4.18. In addition to the overall survival analysis results, we present one-vs-all survival analysis results for each k value in tables 4.19, 4.20, 4.21 and 4.22. The best p-value is obtained when $k = 3$ and $\alpha = 0.2$. Kaplan-Meier plots of best results for each k value is given in Figure 4.6.

$\alpha \backslash\ k$	2	3	4	5
0.00	1.3e-04 (113-248)	8.4e-03 (85-192-84)	5.9e-05 (73-101-76-111)	8.4e-05 (75-65-98-60-63)
0.01	4.5e-06 (118-243)	6.1e-06 (107-91-163)	3.3e-06 (107-90-80-84)	2.9e-03 (55-54-83-88-81)
0.05	1.5e-06 (253-108)	4.0e-04 (173-99-89)	2.7e-05 (65-127-78-91)	3.8e-04 (90-60-63-97-51)
0.10	9.0e-04 (120-241)	6.6e-05 (91-91-179)	4.4e-04 (80-67-74-140)	**1.6e-06** (52-89-76-92-52)
0.20	7.3e-04 (125-236)	**2.0e-07** (169-94-98)	2.2e-05 (91-78-92-100)	2.4e-05 (88-60-66-79-68)
0.30	1.5e-04 (123-238)	4.0e-06 (85-93-183)	6.7e-06 (79-76-136-70)	2.6e-05 (79-98-64-66-54)
0.40	2.0e-02 (119-242)	9.6e-07 (86-74-201)	1.9e-05 (91-76-117-77)	2.5e-05 (71-53-73-86-78)
0.50	4.0e-04 (119-242)	3.7e-03 (74-96-191)	**6.4e-07** (98-122-68-73)	1.1e-04 (102-55-59-85-60)
0.60	5.5e-05 (95-266)	3.8e-03 (98-101-162)	3.4e-04 (77-95-74-115)	8.1e-05 (73-65-73-84-66)
0.70	**9.9e-07** (259-102)	6.3e-05 (177-86-98)	2.3e-05 (92-150-90-29)	8.8e-05 (83-60-71-82-65)
0.80	1.5e-05 (115-246)	2.3e-06 (84-81-196)	3.8e-04 (103-63-74-121)	1.8e-04 (59-85-70-78-69)
0.90	8.6e-04 (246-115)	7.5e-05 (101-83-177)	1.5e-03 (66-135-82-78)	6.1e-04 (68-84-79-66-64)

Table 4.18: Overall survival analysis of results of smSPK on all pathways and RBF kernel on all genes of kidney cancer patients. The numbers in parenthesis indicate the cluster sizes. The bold p-values are the best values obtained for the corresponding k value.

CI \ α	1	2
0.00	1.3e-04 (113)	1.3e-04 (248)
0.01	4.5e-06 (118)	4.5e-06 (243)
0.05	1.5e-06 (253)	1.5e-06 (108)
0.10	9.0e-04 (120)	9.0e-04 (241)
0.20	7.3e-04 (125)	7.3e-04 (236)
0.30	1.5e-04 (123)	1.5e-04 (238)
0.40	2.0e-02 (119)	2.0e-02 (242)
0.50	4.0e-04 (119)	4.0e-04 (242)
0.60	5.5e-05 (95)	5.5e-05 (266)
0.70	9.9e-07 (259)	9.9e-07 (102)
0.80	1.5e-05 (115)	1.5e-05 (246)
0.90	8.6e-04 (246)	8.6e-04 (115)

Table 4.19: One-vs-all survival analysis of results of smSPK on all pathways and RBF kernel on all genes of kidney cancer patients for $k = 2$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

CI \ α	1	2	3
0.00	2.0e-03 (85)	1.2e-01 (192)	3.0e-01 (84)
0.01	1.8e-06 (107)	6.2e-01 (91)	2.7e-04 (163)
0.05	2.0e-02 (173)	7.7e-05 (99)	2.8e-01 (89)
0.10	5.2e-02 (91)	1.4e-05 (91)	7.4e-02 (179)
0.20	1.6e-05 (169)	9.1e-01 (94)	1.0e-07 (98)
0.30	1.8e-02 (85)	9.6e-07 (93)	7.2e-02 (183)
0.40	1.8e-07 (86)	3.9e-02 (74)	2.2e-02 (201)
0.50	1.0e+00 (74)	1.4e-03 (96)	6.2e-03 (191)
0.60	9.7e-04 (98)	4.8e-01 (101)	2.6e-02 (162)
0.70	3.0e-01 (177)	6.3e-03 (86)	3.3e-05 (98)
0.80	3.5e-07 (84)	1.4e-01 (81)	6.8e-03 (196)
0.90	1.3e-05 (101)	1.2e-01 (83)	2.2e-02 (177)

Table 4.20: One-vs-all survival analysis of results of smSPK on all pathways and RBF kernel on all genes of kidney cancer patients for $k = 3$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

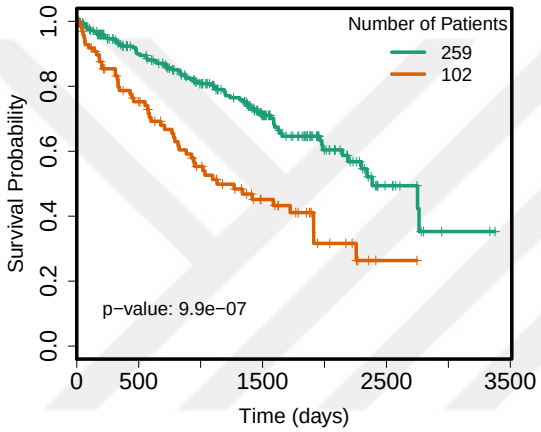
CI \ α	1	2	3	4
0.00	5.7e-01 (73)	4.9e-01 (101)	4.7e-05 (76)	6.3e-04 (111)
0.01	4.8e-01 (107)	3.3e-03 (90)	1.6e-06 (80)	7.2e-02 (84)
0.05	7.6e-01 (65)	1.4e-01 (127)	1.6e-06 (78)	3.8e-02 (91)
0.10	3.7e-02 (80)	7.4e-01 (67)	4.3e-05 (74)	3.2e-01 (140)
0.20	2.4e-01 (91)	1.4e-05 (78)	3.4e-02 (92)	1.2e-02 (100)
0.30	3.2e-07 (79)	2.7e-01 (76)	1.3e-02 (136)	5.9e-01 (70)
0.40	1.7e-01 (91)	7.2e-07 (76)	1.4e-01 (117)	2.3e-01 (77)
0.50	1.2e-02 (98)	4.6e-01 (122)	2.4e-01 (68)	3.8e-08 (73)
0.60	5.4e-02 (77)	8.1e-01 (95)	3.3e-05 (74)	1.9e-01 (115)
0.70	4.5e-06 (92)	3.5e-01 (150)	1.9e-03 (90)	9.0e-01 (29)
0.80	8.8e-01 (103)	6.3e-01 (63)	8.5e-05 (74)	4.2e-03 (121)
0.90	4.0e-01 (66)	3.9e-01 (135)	7.5e-02 (82)	1.2e-04 (78)

Table 4.21: One-vs-all survival analysis of results of smSPK on all pathways and RBF kernel on all genes of kidney cancer patients for $k = 4$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

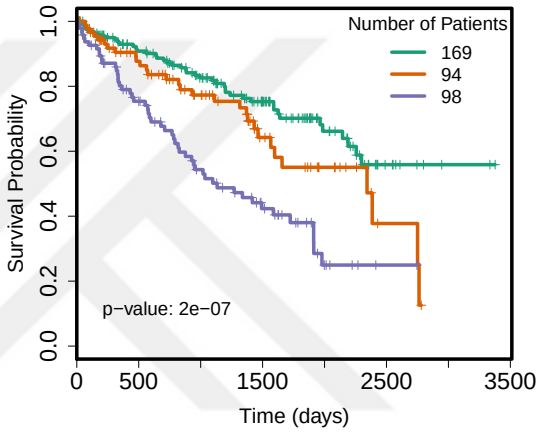
CI \ α	1	2	3	4	5
0.00	2.4e-01 (75)	8.9e-01 (65)	9.5e-03 (98)	6.0e-01 (60)	5.5e-06 (63)
0.01	3.7e-01 (55)	1.2e-04 (54)	1.6e-01 (83)	3.0e-01 (88)	7.4e-01 (81)
0.05	1.6e-02 (90)	5.8e-01 (60)	2.7e-05 (63)	2.3e-01 (97)	8.6e-01 (51)
0.10	7.3e-01 (52)	9.1e-01 (89)	8.8e-08 (76)	1.2e-02 (92)	4.4e-02 (52)
0.20	9.9e-01 (88)	1.4e-01 (60)	5.9e-07 (66)	2.4e-01 (79)	9.3e-02 (68)
0.30	4.0e-02 (79)	5.0e-02 (98)	8.0e-01 (64)	1.5e-06 (66)	5.2e-01 (54)
0.40	4.1e-01 (71)	8.7e-01 (53)	4.2e-06 (73)	3.0e-03 (86)	1.7e-01 (78)
0.50	5.0e-01 (102)	3.0e-06 (55)	4.1e-01 (59)	4.2e-02 (85)	8.9e-01 (60)
0.60	1.3e-01 (73)	9.0e-02 (65)	9.3e-01 (73)	4.0e-01 (84)	2.1e-06 (66)
0.70	6.8e-01 (83)	1.9e-06 (60)	6.9e-01 (71)	2.4e-01 (82)	6.9e-02 (65)
0.80	6.6e-01 (59)	3.7e-01 (85)	2.2e-01 (70)	4.3e-02 (78)	7.2e-06 (69)
0.90	5.3e-01 (68)	3.4e-04 (84)	1.1e-02 (79)	9.8e-02 (66)	1.4e-01 (64)

Table 4.22: One-vs-all survival analysis of results of smSPK on all pathways and RBF kernel on all genes of kidney cancer patients for $k = 5$. The numbers in parenthesis indicate the cluster size that is compared with the rest of the patients.

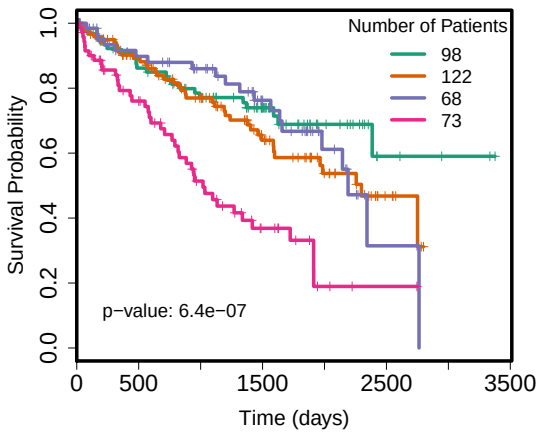
Figure 4.6: Kaplan-Meier plots of experiment utilizing smSPK along with RBF kernel on all genes



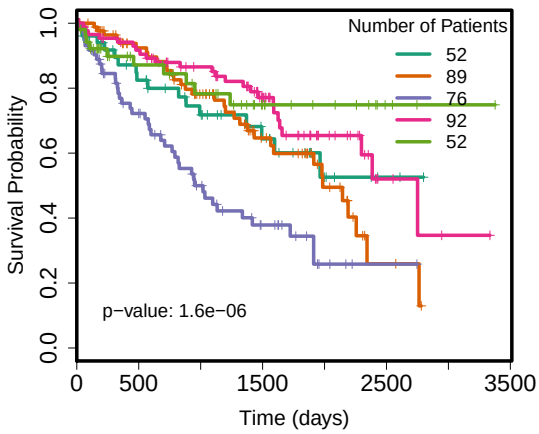
(a) The setting is $k = 2$ and $\alpha = 0$



(b) The setting is $k = 3$ and $\alpha = 0.7$



(c) The setting is $k = 4$ and $\alpha = 0.01$



(d) The setting is $k = 5$ and $\alpha = 0.9$

4.3 Discussion

4.3.1 Performance Comparison

When the survival analysis results of experiments with LMKKM are compared with the survival analysis results of experiments with MVKKM, MVKKM outperforms LMKKM (Compare Table 4.3 and 4.8 results with Table 4.1 and 4.2 results). Kernel-based weighting works better than sample-based weighting when we use only alteration information without mapping on pathway. The summary of the results is given in table 4.23.

Overall, smSPK on all pathways yields good results with $k = 3$ in terms of the log-rank test(p-value 1.5×10^{-8}). However, MVKKM with RBF for $k = 3$ and $k = 5$ also achieve close p-values, 2.2×10^{-8} and 4.1×10^{-8} respectively and MVKKM with all genes yields the best result with p-value 7.5×10^{-9} (65-296). However, as we look at the cluster sizes, the low p-values attained in the RBF case could be attributed to small sample sizes. The problem with log-rank test producing artificially low p-values have been discussed in the literature. This is also apparent in the KM plots. Figure 4.5b shows that smSPK indeed arrives at three well-separated clusters; one vs. all log-rank test results also validate these results in Table 4.15. On the other hand, for RBF with MVKMM there is only one cluster that is well separated; the other small groups help lower the p-value. Therefore, we conclude smSPK with pathway genes provides superior clustering results.

When we compare the results of the two settings, smSPK run with pathways and smSPK with pathway and RBF kernels, in terms of p-value smSPK run only with pathways performs better. This is somewhat surprising; as 33% of the mutated genes, 31% of the RNA expression genes and 90% of the protein expression genes are parts of a pathway. We would expect that the additional RBF kernels would bring more information from the unmapped genes. That is not supported by data. The survival analysis results show that the performance is improved for only $k = 4$. For the rest of k values, the p-value is higher. This

could be attributed to a large number of genes lowering the signal-to-noise ratio; thus kernel matrices do not bring any additional information. Using only pathway genes might be serving as a good feature selection step. The reason for this could also be the number of kernel matrices input to MVKKM.

4.3.2 Important Pathways

One advantage of the smSPK approach is that it provides valuable insight on the pathways that are instrumental in arriving the cluster results. In the study of TCGA [6], Lysine degradation pathway and PI(3)K-Akt pathway are pointed out as potential driver pathways. Furthermore, mutation of the SETD2 gene, which is located in HIF signalling pathway, is considered as a driver mutation. When we analyze our weight distribution of the experiment with smSPK, we observe that HIF signalling pathway is selected for almost all data types over all k values. Lysine degradation pathway is only selected when gene expression data is applied on it. One exception to this is the smSPK experiment with $k = 4$ and $\alpha = 0.01$. In that case, Lysine degradation pathway is not selected even for the gene expression data. PI(3)K-Akt signalling pathway always contributes to the stratification of patients when gene expression data and protein expression data are mapped on it. For somatic mutation data mapped PI(3)K-Akt signalling pathway is selected only for smSPK experiment with $k = 3$ and $\alpha = 0.7$. This analysis shows that the selection of pathways for stratification of patients is compatible with the potential driver pathways. The detailed information on selected pathways along with the best corresponding methods are provided in 4.24.

Method \ k	2	3	4	5
RBF on all genes with LMKKM	3.5e-02 (141-220)	6.3e-05 (102-97-162)	5.1e-05 (120-53-27-161)	1.2e-04 (22-49-50-140-100)
RBF on pathway genes with LMKKM	3.4e-03 (230-131)	1.6e-04 (143-91-127)	3.4e-04 (44-76-121-120)	3.9e-04 (54-106-60-106-35)
RBF on all genes with MVKKM	7.5e-09 (65-296)	7.9e-07 (14-67-280)	3.3e-06 (13-66-272-10)	3.0e-06 (272-30-17-14-28)
RBF on pathway genes with MVKKM	3.0e-06 (265-96)	2.2e-05 (76-12-273)	2.2e-08 (17-284-26-34)	4.1e-08 (15-273-19-23-31)
smSPK on all pathways	6.1e-08 (125-236)	1.5e-08 (82-88-191)	2.1e-06 (109-91-80-81)	5.7e-07 (64-67-75-91-64)
smSPK along with RBF on all genes	9.9e-07 (259-102)	2.0e-07 (169-94-98)	6.4e-07 (98-122-68-73)	1.6e-06 (52-89-76-92-52)

Table 4.23: The best results obtained for each method and each k value.

Table 4.24: The selection of potential driver pathways of RCC is displayed for smSPK and smSPK along with RBF kernel. Each k value of each setting is given. The rows of each k are somatic mutation (SE), gene expression (GE) and protein expression (PE) mapped versions of that pathway. "✓" indicates that the pathway on which the corresponding data type mapped is selected for that k value by the corresponding method.

Method	k	Data Type	Lysine degradation pathway	HIF signalling pathway	PI(3)K-Akt signalling pathway
smSPK	2	SM		✓	
		GE	✓	✓	✓
		PE		✓	✓
	3	SM		✓	✓
		GE	✓	✓	✓
		PE		✓	✓
	4	SM		✓	
		GE		✓	✓
		PE		✓	✓
	5	SM			
		GE	✓	✓	✓
		PE		✓	✓
smSPK with RBF	2	SM			
		GE	✓	✓	✓
		PE		✓	✓
	3	SM		✓	
		GE	✓	✓	✓
		PE		✓	✓
	4	SM		✓	
		GE	✓	✓	✓
		PE		✓	✓
	5	SM			
		GE	✓	✓	✓
		PE		✓	✓

Chapter 5

Conclusion and Future Work

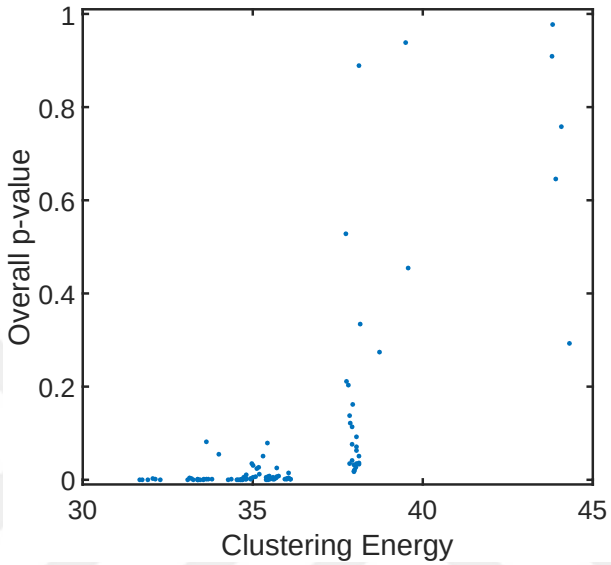
Identification of patient subgroups are important for many cancer types where the molecular subtypes are not well defined. One such cancer is kidney renal cell carcinoma. Accurate classification of patients into clinically relevant subtypes paves the way for the development of effective diagnostic and therapeutic strategies. Molecular data that catalogues DNA and RNA alterations for large cohorts of cancer patients provide unprecedented multiple views into the biology of the tumor. Yet there is still an extensive amount of heterogeneity observed in cancer patients. This presents a particular challenge for studies which set out to identify clinically meaningful subtypes of the same cancer type.

Integrative approaches concurrently account for the known interactions among molecules in the cell and therefore play an instrumental role in overcoming the complications brought out by this heterogeneity. In this thesis, we develop a framework 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 develop a novel graph kernel that evaluates patient similarities based on their genomic, transcriptomic and proteomic alterations in the context of known pathways and employ a multi-view kernel clustering techniques. Our results indicate that the suggested methodology, when applied to RCC, results in patient clusters that differ significantly in

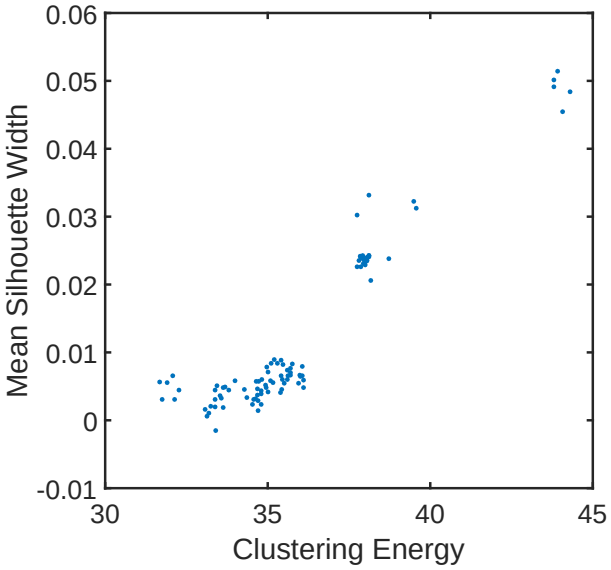
their survival distributions. The proposed methodology also provides quantitative evidence for the decisive role of known driver pathways on the clustering process.

In tackling this project the largest hurdle comes from the nature of clustering task. Since we lack the labels of patients indicating their exact subtypes, it is not straightforward to assess the performance of the results. We rely on survival analysis approach to accomplish that. However, our experiments also indicate that p-values obtained by testing the survival differences of the clusters are not always positively correlated with the clustering energy, which is the objective function of the clustering algorithm. Therefore, there is an inherent discrepancy between the objective function and the evaluation strategy. To remedy this situation we also tried other clustering evaluation metrics, such as the silhouette width, which is widely used to assess the performance of clustering [24]. However, we again observe that this evaluation metric is not always positively correlated with the objective function value. We believe this stems from the fact that the silhouette width works well when the clusters are globular and well separated but might fail if they are not so. Figure 5.1 demonstrates these relations over 100 MVKKM repetitions.

The work can be extended in several directions. When there is a relatively high number of kernel matrices, MVKKM do not perform well, and the assigned weights of the selected matrices are close to each other. To improve the proposed framework, it is possible to employ a multi-view kernel clustering method, which is capable of handling large number of kernel matrices kernel matrices. Furthermore, log-rank test has reliability problems when the sizes of some clusters are small compared to the rest of the clusters [47]. Therefore, alternative tests that do not suffer from the same problem can be explored. The proposed kernel matrix characterizes the similarities of patients based on the shortest path on the graphs. Other graph kernels can be devised. In addition, one can employ a different labeling strategy to emphasize rarer mutations. In the present study we ignore the directionality of the edges in the pathways. A kernel that explicitly accounts for edge directions can be more expressive. Finally, the framework can be tested in other cancer types.



(a) The change of p-value according to the clustering energy of MVKKM for 100 repetitions.



(b) The change of mean silhouette width according to the clustering energy of MVKKM for 100 repetitions.

Figure 5.1: In (a), there are cases contradicting with the correlation of p-value and clustering energy. In (b), we see that mean silhouette width and clustering energy positively correlated whereas we expected the negative correlation.

Bibliography

[1] C. M. Perou, T. Sorlie, M. B. Eisen, M. Van De Rijn, *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, p. 747, 2000.

[2] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences*, vol. 98, no. 19, pp. 10869–10874, 2001.

[3] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, *et al.*, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of Clinical Oncology*, vol. 27, no. 8, pp. 1160–1167, 2009.

[4] A. Toss and M. Cristofanilli, "Molecular characterization and targeted therapeutic approaches in breast cancer," *Breast Cancer Research*, vol. 17, no. 1, p. 60, 2015.

[5] B. I. Rini, S. C. Campbell, and B. Escudier, "Renal cell carcinoma," *The Lancet*, vol. 373, no. 9669, pp. 1119–1132, 2009.

[6] C. G. A. R. Network *et al.*, "Comprehensive molecular characterization of clear cell renal cell carcinoma," *Nature*, vol. 499, no. 7456, pp. 43–49, 2013.

[7] P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson, C. Sander, *et al.*, "Pathway and

network analysis of cancer genomes," *Nature Methods*, vol. 12, no. 7, p. 615, 2015.

[8] N. Shervashidze, P. Schweitzer, E. J. v. Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-lehman graph kernels," *Journal of Machine Learning Research*, vol. 12, no. Sep, pp. 2539–2561, 2011.

[9] K. M. Borgwardt and H.-P. Kriegel, "Shortest-path kernels on graphs," in *Data Mining, Fifth IEEE International Conference on*, pp. 8–pp, IEEE, 2005.

[10] G. Tzortzis and A. Likas, "Kernel-based weighted multi-view clustering," in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pp. 675–684, IEEE, 2012.

[11] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. Springer, 2013.

[12] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1," *Cancer Cell*, vol. 17, no. 1, pp. 98–110, 2010.

[13] A. A. Alizadeh, M. B. Elsen, R. E. Davis, C. Ma, *et al.*, "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, p. 503, 2000.

[14] R. W. Tothill, A. V. Tinker, J. George, R. Brown, S. B. Fox, S. Lade, D. S. Johnson, M. K. Trivett, D. Etemadmoghadam, B. Locandro, *et al.*, "Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome," *Clinical Cancer Research*, vol. 14, no. 16, pp. 5198–5208, 2008.

[15] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, *et al.*, "Repeated observation of breast tumor subtypes in independent gene expression data sets.," *Proceedings of*

the National Academy of Sciences of the United States of America, vol. 100, no. 14, pp. 8418–8423, 2003.

[16] J. Lapointe, C. Li, J. P. Higgins, M. Van De Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, *et al.*, "Gene expression profiling identifies clinically relevant subtypes of prostate cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 3, pp. 811–816, 2004.

[17] A. K. Virmani, J. A. Tsou, K. D. Siegmund, L. Y. Shen, T. I. Long, P. W. Laird, A. F. Gazdar, and I. A. Laird-Offringa, "Hierarchical clustering of lung cancer cell lines using dna methylation markers," *Cancer Epidemiology and Prevention Biomarkers*, vol. 11, no. 3, pp. 291–297, 2002.

[18] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, no. 1, pp. 91–118, 2003.

[19] C. G. A. Network *et al.*, "Comprehensive molecular portraits of human breast tumors," *Nature*, vol. 490, no. 7418, p. 61, 2012.

[20] A. R. Brannon, A. Reddy, M. Seiler, A. Arreola, D. T. Moore, R. S. Pruthi, E. M. Wallen, M. E. Nielsen, H. Liu, K. L. Nathanson, *et al.*, "Molecular stratification of clear cell renal cell carcinoma by consensus clustering reveals distinct subtypes and survival patterns," *Genes & Cancer*, vol. 1, no. 2, pp. 152–163, 2010.

[21] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

[22] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations," *Nature Methods*, vol. 10, no. 11, pp. 1108–1115, 2013.

[23] C. G. A. R. Network *et al.*, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–615, 2011.

[24] N. K. Speicher and N. Pfeifer, "Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery," *Bioinformatics*, vol. 31, no. 12, pp. i268–i275, 2015.

[25] S. Yu, L. Tranchevent, X. Liu, W. Glanzel, J. A. Suykens, B. De Moor, and Y. Moreau, "Optimized data fusion for kernel k-means clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1031–1039, 2012.

[26] J. Chen, Z. Zhao, J. Ye, and H. Liu, "Nonlinear adaptive distance metric learning for clustering," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 123–132, ACM, 2007.

[27] M. B. Blaschko and C. H. Lampert, "Correlational spectral clustering," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[28] G. F. Tzortzis and A. C. Likas, "Multiple view clustering using a weighted combination of exemplar-based mixture models," *IEEE Transactions on Neural Networks*, vol. 21, no. 12, pp. 1925–1938, 2010.

[29] M. Gönen and A. A. Margolin, "Localized data fusion for kernel k-means clustering with application to cancer biology," in *Advances in Neural Information Processing Systems*, pp. 1305–1313, 2014.

[30] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS Computational Biology*, vol. 6, no. 1, p. e1000641, 2010.

[31] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, *et al.*, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, 2013.

[32] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014.

[33] Y. Liu, Q. Gu, J. P. Hou, J. Han, and J. Ma, "A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression," *BMC Bioinformatics*, vol. 15, no. 1, p. 37, 2014.

[34] S. Kim, M. Kon, and C. DeLisi, "Pathway-based classification of cancer subtypes," *Biology Direct*, vol. 7, no. 1, p. 21, 2012.

[35] T. Gärtner, P. Flach, and S. Wrobel, "On graph kernels: Hardness results and efficient alternatives," *Learning Theory and Kernel Machines*, pp. 129–143, 2003.

[36] W. Imrich and S. Klavzar, *Product graphs: structure and recognition*. Wiley, 2000.

[37] P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, and J.-P. Vert, "Extensions of marginalized graph kernels," in *Proceedings of the Twenty-first International Conference on Machine Learning*, p. 70, ACM, 2004.

[38] K. M. Borgwardt, *Graph kernels*. PhD thesis, lmu, 2007.

[39] R. W. Floyd, "Algorithm 97: shortest path," *Communications of the ACM*, vol. 5, no. 6, p. 345, 1962.

[40] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Kegg as a reference resource for gene and protein annotation," *Nucleic Acids Research*, p. gkv1070, 2015.

[41] A. Arakelyan and L. Nersisyan, "Keggparser: parsing and editing kegg pathway maps in matlab," *Bioinformatics*, vol. 29, no. 4, pp. 518–519, 2013.

[42] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[43] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[44] D. P. Harrington and T. R. Fleming, "A class of rank test procedures for censored survival data," *Biometrika*, vol. 69, no. 3, pp. 553–566, 1982.

[45] T. M. Therneau, *A Package for Survival Analysis in S*, 2015. version 2.38.

[46] Terry M. Therneau and Patricia M. Grambsch, *Modeling Survival Data: Extending the Cox Model*. New York: Springer, 2000.

[47] F. Vandin, A. Papoutsaki, B. J. Raphael, and E. Upfal, "Accurate computation of survival statistics in genome-wide studies," *PLoS Computational Biology*, vol. 11, no. 5, p. e1004071, 2015.

Appendix A

Selected Pathways

Table A.1: Pathway selection table of multi-view kernel clustering with pathway based shortest path graph kernels experiment with $k = 3$ and $\alpha = 0.7$. Here, we demonstrate pathways which have non-zero weight in at least one data type.

Pathway \ Dataset	**SM**	**GE**	**PE**
hsa00010 Glycolysis / Gluconeogenesis		✓	
hsa00030 Pentose phosphate pathway		✓	
hsa00051 Fructose and mannose metabolism		✓	
hsa00052 Galactose metabolism		✓	
hsa00140 Steroid hormone biosynthesis		✓	
hsa00230 Purine metabolism		✓	
hsa00240 Pyrimidine metabolism		✓	
hsa00270 Cysteine and methionine metabolism		✓	
hsa00310 Lysine degradation		✓	
hsa00330 Arginine and proline metabolism		✓	
hsa00350 Tyrosine metabolism		✓	
hsa00360 Phenylalanine metabolism		✓	
hsa00410 betaAlanine metabolism		✓	
hsa00480 Glutathione metabolism		✓	
hsa00500 Starch and sucrose metabolism		✓	

hsa00510 NGlycan biosynthesis		✓	
hsa00512 Mucin type OGlycan biosynthesis		✓	
hsa00520 Amino sugar and nucleotide sugar metabolism		✓	
hsa00531 Glycosaminoglycan degradation		✓	
hsa00561 Glycerolipid metabolism		✓	
hsa00562 Inositol phosphate metabolism		✓	
hsa00563 Glycosylphosphatidylinositol(GPI) anchor biosynthesis		✓	
hsa00564 Glycerophospholipid metabolism		✓	
hsa00565 Ether lipid metabolism		✓	
hsa00590 Arachidonic acid metabolism		✓	
hsa00591 Linoleic acid metabolism		✓	
hsa00600 Sphingolipid metabolism		✓	
hsa00601 Glycosphingolipid biosynthesis lacto and neolacto series		✓	
hsa00620 Pyruvate metabolism		✓	
hsa00760 Nicotinate and nicotinamide metabolism		✓	
hsa00860 Porphyrin and chlorophyll metabolism		✓	
hsa00983 Drug metabolism other enzymes		✓	
hsa03013 RNA transport		✓	
hsa03015 mRNA surveillance pathway		✓	
hsa03460 Fanconi anemia pathway		✓	
hsa04010 MAPK signaling pathway		✓	✓
hsa04012 ErbB signaling pathway	✓	✓	✓
hsa04014 Ras signaling pathway	✓	✓	✓
hsa04015 Rap1 signaling pathway	✓	✓	✓
hsa04020 Calcium signaling pathway		✓	✓
hsa04022 cGMPPKG signaling pathway	✓	✓	✓
hsa04024 cAMP signaling pathway	✓	✓	✓
hsa04062 Chemokine signaling pathway	✓	✓	✓
hsa04064 NFkappa B signaling pathway		✓	

hsa04066 HIF1 signaling pathway	✓	✓	✓
hsa04068 FoxO signaling pathway	✓	✓	✓
hsa04070 Phosphatidylinositol signaling system		✓	
hsa04071 Sphingolipid signaling pathway	✓	✓	✓
hsa04072 Phospholipase D signaling pathway	✓	✓	✓
hsa04110 Cell cycle		✓	
hsa04114 Oocyte meiosis		✓	✓
hsa04115 p53 signaling pathway		✓	✓
hsa04130 SNARE interactions in vesicular transport		✓	
hsa04141 Protein processing in endoplasmic reticulum		✓	
hsa04144 Endocytosis		✓	
hsa04150 mTOR signaling pathway	✓	✓	✓
hsa04151 PI3KAkt signaling pathway	✓	✓	✓
hsa04152 AMPK signaling pathway	✓	✓	✓
hsa04210 Apoptosis	✓	✓	✓
hsa04211 Longevity regulating pathway mammal	✓	✓	✓
hsa04261 Adrenergic signaling in cardiomyocytes	✓	✓	✓
hsa04270 Vascular smooth muscle contraction		✓	✓
hsa04310 Wnt signaling pathway		✓	
hsa04330 Notch signaling pathway		✓	
hsa04340 Hedgehog signaling pathway		✓	
hsa04350 TGFbeta signaling pathway		✓	
hsa04360 Axon guidance		✓	
hsa04370 VEGF signaling pathway	✓	✓	✓
hsa04380 Osteoclast differentiation	✓	✓	✓
hsa04390 Hippo signaling pathway		✓	
hsa04510 Focal adhesion	✓	✓	✓
hsa04520 Adherens junction		✓	✓
hsa04530 Tight junction		✓	✓
hsa04540 Gap junction		✓	✓

hsa04550 Signaling pathways regulating pluripotency of stem cells	✓	✓	✓
hsa04611 Platelet activation	✓	✓	✓
hsa04620 Tolllike receptor signaling pathway	✓	✓	✓
hsa04621 NODlike receptor signaling pathway		✓	
hsa04622 RIGIlike receptor signaling pathway		✓	
hsa04623 Cytosolic DNAsensing pathway		✓	
hsa04630 JakSTAT signaling pathway	✓	✓	✓
hsa04650 Natural killer cell mediated cytotoxicity	✓	✓	✓
hsa04660 T cell receptor signaling pathway	✓	✓	✓
hsa04662 B cell receptor signaling pathway	✓	✓	✓
hsa04664 Fc epsilon RI signaling pathway	✓	✓	✓
hsa04666 Fc gamma Rmediated phagocytosis	✓	✓	✓
hsa04668 TNF signaling pathway	✓	✓	✓
hsa04670 Leukocyte transendothelial migration		✓	✓
hsa04710 Circadian rhythm		✓	
hsa04713 Circadian entrainment		✓	
hsa04720 Longterm potentiation		✓	✓
hsa04722 Neurotrophin signaling pathway	✓	✓	✓
hsa04724 Glutamatergic synapse		✓	✓
hsa04725 Cholinergic synapse	✓	✓	✓
hsa04726 Serotonergic synapse		✓	✓
hsa04728 Dopaminergic synapse		✓	
hsa04730 Longterm depression		✓	✓
hsa04740 Olfactory transduction		✓	
hsa04750 Inflammatory mediator regulation of TRP channels		✓	✓
hsa04810 Regulation of actin cytoskeleton		✓	✓
hsa04910 Insulin signaling pathway	✓	✓	✓
hsa04911 Insulin secretion		✓	
hsa04912 GnRH signaling pathway		✓	✓

hsa04913 Ovarian steroidogenesis		✓	
hsa04914 Progesteronemediated oocyte maturation	✓	✓	✓
hsa04915 Estrogen signaling pathway	✓	✓	✓
hsa04916 Melanogenesis		✓	✓
hsa04917 Prolactin signaling pathway	✓	✓	✓
hsa04918 Thyroid hormone synthesis		✓	
hsa04919 Thyroid hormone signaling pathway	✓	✓	✓
hsa04920 Adipocytokine signaling pathway		✓	
hsa04921 Oxytocin signaling pathway		✓	✓
hsa04922 Glucagon signaling pathway		✓	✓
hsa04923 Regulation of lipolysis in adipocytes	✓	✓	✓
hsa04924 Renin secretion		✓	
hsa04925 Aldosterone synthesis and secretion		✓	
hsa04930 Type II diabetes mellitus	✓	✓	✓
hsa04931 Insulin resistance	✓	✓	✓
hsa04932 Nonalcoholic fatty liver disease (NAFLD)	✓	✓	✓
hsa04933 AGERAGE signaling pathway in diabetic complications	✓	✓	✓
hsa04960 Aldosteroneregulated sodium reabsorption	✓	✓	✓
hsa04961 Endocrine and other factorregulated calcium reabsorption		✓	
hsa04970 Salivary secretion		✓	
hsa04971 Gastric acid secretion		✓	
hsa05010 Alzheimer's disease		✓	
hsa05012 Parkinson's disease		✓	
hsa05014 Amyotrophic lateral sclerosis (ALS)		✓	
hsa05016 Huntington's disease		✓	
hsa05020 Prion diseases			✓
hsa05030 Cocaine addiction		✓	
hsa05031 Amphetamine addiction		✓	
hsa05032 Morphine addiction		✓	

hsa05034 Alcoholism		✓	✓
hsa05100 Bacterial invasion of epithelial cells	✓	✓	✓
hsa05120 Epithelial cell signaling in Helicobacter pylori infection		✓	
hsa05131 Shigellosis		✓	
hsa05132 Salmonella infection		✓	
hsa05133 Pertussis		✓	
hsa05134 Legionellosis		✓	
hsa05142 Chagas disease (American trypanosomiasis)	✓	✓	✓
hsa05145 Toxoplasmosis	✓	✓	✓
hsa05152 Tuberculosis		✓	✓
hsa05160 Hepatitis C	✓	✓	✓
hsa05161 Hepatitis B		✓	✓
hsa05162 Measles	✓	✓	
hsa05164 Influenza A	✓	✓	✓
hsa05166 HTLVI infection	✓	✓	✓
hsa05168 Herpes simplex infection		✓	
hsa05169 EpsteinBarr virus infection	✓	✓	
hsa05200 Pathways in cancer	✓	✓	✓
hsa05205 Proteoglycans in cancer	✓	✓	✓
hsa05210 Colorectal cancer	✓	✓	✓
hsa05211 Renal cell carcinoma	✓	✓	✓
hsa05212 Pancreatic cancer	✓	✓	✓
hsa05213 Endometrial cancer	✓	✓	✓
hsa05214 Glioma	✓	✓	✓
hsa05215 Prostate cancer	✓	✓	✓
hsa05218 Melanoma	✓	✓	✓
hsa05219 Bladder cancer		✓	✓
hsa05220 Chronic myeloid leukemia	✓	✓	✓
hsa05221 Acute myeloid leukemia	✓	✓	✓
hsa05222 Small cell lung cancer	✓	✓	

hsa05223 Nonsmall cell lung cancer	✓	✓	✓
hsa05231 Choline metabolism in cancer	✓	✓	✓
hsa05321 Inflammatory bowel disease (IBD)		✓	
hsa05410 Hypertrophic cardiomyopathy (HCM)		✓	
hsa05416 Viral myocarditis		✓	

Table A.2: Pathway selection table of multi-view kernel clustering with pathway based shortest path graph kernels along with RBF kernels using all genes experiment with $k = 3$ and $\alpha = 0.2$. Here, we demonstrate pathways which have non-zero weight in at least one data type.

Pathway \ Dataset	**SM**	**GE**	**PE**
hsa00051 Fructose and mannose metabolism		✓	
hsa00052 Galactose metabolism		✓	
hsa00230 Purine metabolism		✓	
hsa00240 Pyrimidine metabolism		✓	
hsa00310 Lysine degradation		✓	
hsa00330 Arginine and proline metabolism		✓	
hsa00360 Phenylalanine metabolism		✓	
hsa00500 Starch and sucrose metabolism		✓	
hsa00510 NGlycan biosynthesis		✓	
hsa00520 Amino sugar and nucleotide sugar metabolism		✓	
hsa00561 Glycerolipid metabolism		✓	
hsa00562 Inositol phosphate metabolism		✓	
hsa00564 Glycerophospholipid metabolism		✓	
hsa00565 Ether lipid metabolism		✓	
hsa00590 Arachidonic acid metabolism		✓	
hsa00591 Linoleic acid metabolism		✓	
hsa00600 Sphingolipid metabolism		✓	
hsa00760 Nicotinate and nicotinamide metabolism		✓	
hsa00860 Porphyrin and chlorophyll metabolism		✓	
hsa00983 Drug metabolism other enzymes		✓	

hsa03013 RNA transport		✓	
hsa03015 mRNA surveillance pathway		✓	
hsa03460 Fanconi anemia pathway		✓	
hsa04010 MAPK signaling pathway		✓	✓
hsa04012 ErbB signaling pathway	✓	✓	✓
hsa04014 Ras signaling pathway	✓	✓	✓
hsa04015 Rap1 signaling pathway		✓	✓
hsa04020 Calcium signaling pathway		✓	✓
hsa04022 cGMPPKG signaling pathway	✓	✓	✓
hsa04024 cAMP signaling pathway		✓	✓
hsa04062 Chemokine signaling pathway	✓	✓	✓
hsa04064 NFkappa B signaling pathway		✓	
hsa04066 HIF1 signaling pathway	✓	✓	✓
hsa04068 FoxO signaling pathway	✓	✓	✓
hsa04070 Phosphatidylinositol signaling system		✓	
hsa04071 Sphingolipid signaling pathway	✓	✓	✓
hsa04072 Phospholipase D signaling pathway	✓	✓	✓
hsa04110 Cell cycle		✓	
hsa04114 Oocyte meiosis		✓	✓
hsa04115 p53 signaling pathway			✓
hsa04130 SNARE interactions in vesicular transport		✓	
hsa04141 Protein processing in endoplasmic reticulum		✓	
hsa04144 Endocytosis		✓	✓
hsa04150 mTOR signaling pathway		✓	✓
hsa04151 PI3KAkt signaling pathway		✓	✓
hsa04152 AMPK signaling pathway	✓	✓	✓
hsa04210 Apoptosis	✓	✓	✓
hsa04211 Longevity regulating pathway mammal	✓	✓	✓
hsa04261 Adrenergic signaling in cardiomyocytes		✓	✓
hsa04270 Vascular smooth muscle contraction		✓	✓
hsa04310 Wnt signaling pathway		✓	

hsa04330 Notch signaling pathway		✓	
hsa04350 TGFbeta signaling pathway		✓	
hsa04360 Axon guidance		✓	
hsa04370 VEGF signaling pathway	✓	✓	✓
hsa04380 Osteoclast differentiation	✓	✓	✓
hsa04390 Hippo signaling pathway		✓	
hsa04510 Focal adhesion		✓	✓
hsa04514 Cell adhesion molecules (CAMs)		✓	
hsa04520 Adherens junction		✓	✓
hsa04530 Tight junction		✓	✓
hsa04540 Gap junction		✓	✓
hsa04550 Signaling pathways regulating pluripotency of stem cells	✓	✓	✓
hsa04611 Platelet activation	✓	✓	✓
hsa04620 Tolllike receptor signaling pathway	✓	✓	✓
hsa04621 NODlike receptor signaling pathway		✓	
hsa04622 RIGIlike receptor signaling pathway		✓	
hsa04623 Cytosolic DNAsensing pathway		✓	
hsa04630 JakSTAT signaling pathway		✓	✓
hsa04650 Natural killer cell mediated cytotoxicity	✓	✓	✓
hsa04660 T cell receptor signaling pathway	✓	✓	✓
hsa04662 B cell receptor signaling pathway	✓	✓	✓
hsa04664 Fc epsilon RI signaling pathway	✓	✓	✓
hsa04666 Fc gamma Rmediated phagocytosis	✓	✓	✓
hsa04668 TNF signaling pathway	✓	✓	✓
hsa04670 Leukocyte transendothelial migration		✓	✓
hsa04710 Circadian rhythm		✓	
hsa04713 Circadian entrainment		✓	✓
hsa04720 Longterm potentiation		✓	✓
hsa04722 Neurotrophin signaling pathway	✓	✓	✓
hsa04724 Glutamatergic synapse		✓	✓

hsa04725 Cholinergic synapse	✓	✓	✓
hsa04726 Serotonergic synapse		✓	✓
hsa04728 Dopaminergic synapse		✓	✓
hsa04730 Longterm depression		✓	✓
hsa04740 Olfactory transduction		✓	
hsa04750 Inflammatory mediator regulation of TRP channels	✓	✓	✓
hsa04810 Regulation of actin cytoskeleton		✓	✓
hsa04910 Insulin signaling pathway	✓	✓	✓
hsa04911 Insulin secretion		✓	
hsa04912 GnRH signaling pathway		✓	✓
hsa04913 Ovarian steroidogenesis		✓	
hsa04914 Progesteronemediated oocyte maturation	✓	✓	✓
hsa04915 Estrogen signaling pathway	✓	✓	✓
hsa04916 Melanogenesis		✓	✓
hsa04917 Prolactin signaling pathway	✓	✓	✓
hsa04918 Thyroid hormone synthesis		✓	
hsa04919 Thyroid hormone signaling pathway	✓	✓	✓
hsa04920 Adipocytokine signaling pathway		✓	
hsa04921 Oxytocin signaling pathway		✓	✓
hsa04922 Glucagon signaling pathway		✓	✓
hsa04923 Regulation of lipolysis in adipocytes	✓	✓	✓
hsa04924 Renin secretion		✓	
hsa04925 Aldosterone synthesis and secretion		✓	
hsa04930 Type II diabetes mellitus	✓	✓	✓
hsa04931 Insulin resistance	✓	✓	✓
hsa04932 Nonalcoholic fatty liver disease (NAFLD)	✓	✓	✓
hsa04933 AGERAGE signaling pathway in diabetic complications	✓	✓	✓
hsa04960 Aldosteroneregulated sodium reabsorption	✓	✓	✓

hsa04961 Endocrine and other factorregulated calcium reabsorption		✓	
hsa04970 Salivary secretion		✓	
hsa04971 Gastric acid secretion		✓	
hsa05010 Alzheimer's disease		✓	
hsa05012 Parkinson's disease		✓	
hsa05014 Amyotrophic lateral sclerosis (ALS)		✓	
hsa05016 Huntington's disease		✓	
hsa05020 Prion diseases		✓	✓
hsa05030 Cocaine addiction		✓	
hsa05031 Amphetamine addiction		✓	
hsa05032 Morphine addiction		✓	
hsa05034 Alcoholism		✓	✓
hsa05100 Bacterial invasion of epithelial cells	✓	✓	✓
hsa05120 Epithelial cell signaling in Helicobacter pylori infection		✓	
hsa05131 Shigellosis		✓	
hsa05132 Salmonella infection		✓	
hsa05133 Pertussis		✓	
hsa05140 Leishmaniasis		✓	
hsa05142 Chagas disease (American trypanosomiasis)		✓	✓
hsa05145 Toxoplasmosis	✓	✓	✓
hsa05152 Tuberculosis		✓	✓
hsa05160 Hepatitis C	✓	✓	✓
hsa05161 Hepatitis B		✓	✓
hsa05162 Measles	✓	✓	✓
hsa05164 Influenza A	✓	✓	✓
hsa05166 HTLVI infection		✓	✓
hsa05168 Herpes simplex infection		✓	
hsa05169 EpsteinBarr virus infection	✓	✓	✓
hsa05200 Pathways in cancer	✓	✓	✓

hsa05205 Proteoglycans in cancer	✓	✓	✓
hsa05210 Colorectal cancer	✓	✓	✓
hsa05211 Renal cell carcinoma	✓	✓	✓
hsa05212 Pancreatic cancer	✓	✓	✓
hsa05213 Endometrial cancer	✓	✓	✓
hsa05214 Glioma	✓	✓	✓
hsa05215 Prostate cancer	✓	✓	✓
hsa05218 Melanoma	✓	✓	✓
hsa05219 Bladder cancer		✓	✓
hsa05220 Chronic myeloid leukemia	✓	✓	✓
hsa05221 Acute myeloid leukemia	✓	✓	✓
hsa05222 Small cell lung cancer	✓	✓	✓
hsa05223 Nonsmall cell lung cancer	✓	✓	✓
hsa05231 Choline metabolism in cancer	✓	✓	✓
hsa05410 Hypertrophic cardiomyopathy (HCM)		✓	
hsa05414 Dilated cardiomyopathy		✓	
hsa05416 Viral myocarditis		✓	

Appendix B

Weight Distributions

Figure B.1: The weight distribution of multi-view kernel clustering with pathway based shortest path graph kernels experiment with $k = 3$ and $\alpha = 0.7$. The top figure is the weights of pathways on which somatic mutations are mapped. In the middle, weights of gene expression mapped pathways are shown. In the bottom, we see the weights of protein expression mapped pathways.

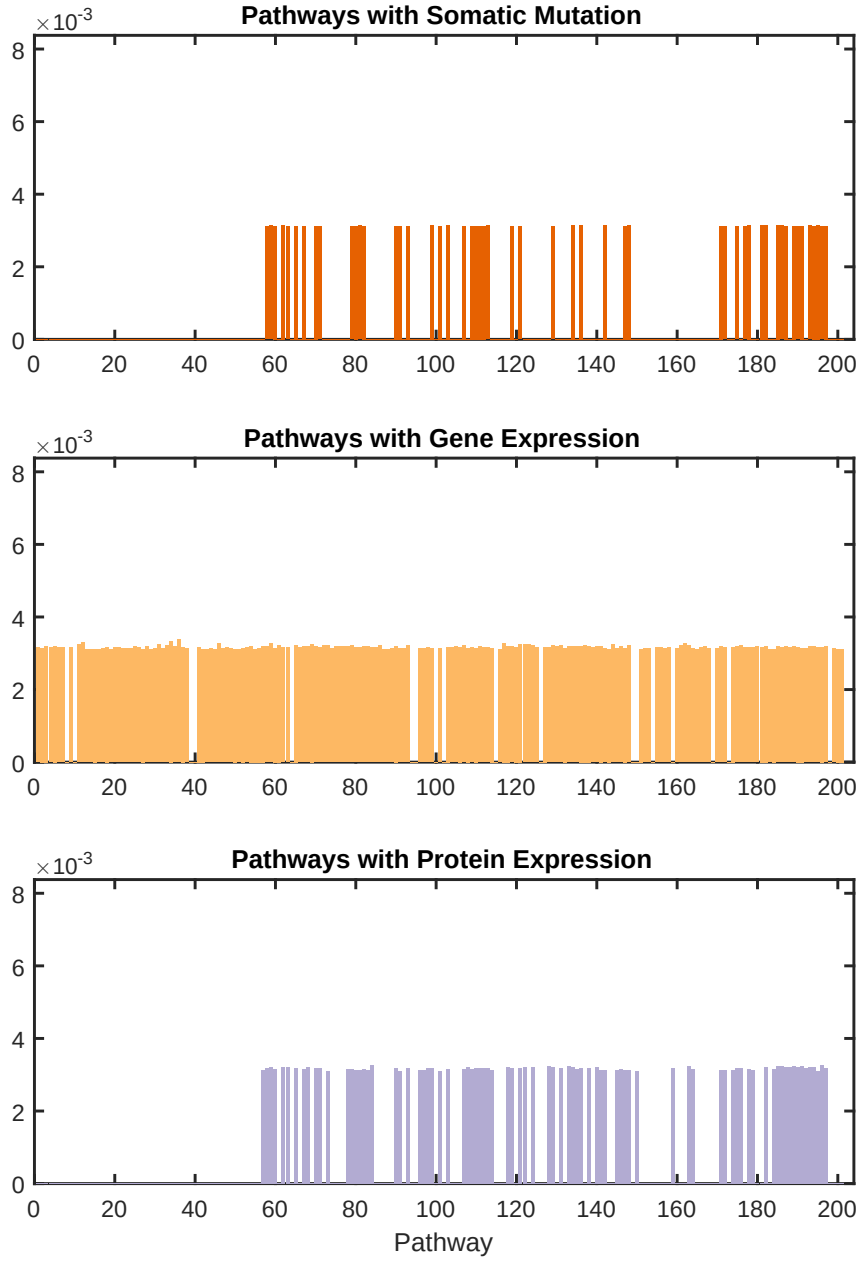


Figure B.2: The weight distribution of multi-view kernel clustering with pathway based shortest path graph kernels along with RBF kernels using all genes experiment with $k = 3$ and $\alpha = 0.2$. The top figure is the weights of pathways on which somatic mutations are mapped. In the middle, weights of gene expression mapped pathways are shown. In the bottom, we see the weights of protein expression mapped pathways. In addition to these, the weight of RBF kernel on all genes for that data type is shown in purple at the end of each figure.