PAMOGK: A Pathway Graph Kernel based Multi-Omics Clustering Approach for Discovering Cancer Patient Subgroups

Yasin Ilkagan Tepeli[1 [0000−0002−3375−6678]] ⋆, Ali Burak Ünal[2,3 [0000−0002−7279−620X]] ⋆, Furkan Mustafa Akdemir[3 [0000−0003−0948−5756]], and Oznur Tastan[1 [0000−0001−7058−5372]] ⋆⋆

[1] Faculty of Engineering and Natural Sciences, Sabanci University, 34956, Istanbul, Turkey
[2] Dept of Computer Science, University of Tübingen, 72076, Tübingen, Germany
[3] Dept of Computer Engineering, Bilkent University, 06800, Ankara, Turkey

Abstract. Accurate classification of patients into homogeneous molecular subgroups is critical for the development of effective therapeutics and for deciphering what drives these different subtypes to cancer. However, the extensive molecular heterogeneity observed among cancer patients presents a challenge. The availability of multi-omic data catalogs for large cohorts of cancer patients provides multiple views into the molecular biology of the tumors with unprecedented resolution. In this work, we develop PAMOGK, which integrates multi-omics patient data and incorporates the existing knowledge on biological pathways. PAMOGK is well suited to deal with the sparsity of alterations in assessing patient similarities. We develop a novel graph kernel which we denote as smoothed shortest path graph kernel, which evaluates patient similarities based on a single molecular alteration type in the context of pathway. To corroborate multiple views of patients evaluated by hundreds of pathways and molecular alteration combinations, PAMOGK uses multi-view kernel clustering. We apply PAMOGK to find subgroups of kidney renal clear cell carcinoma (KIRC) patients, which results in four clusters with significantly different survival times (p-value = 7.4e-10). The patient subgroups also differ with respect to other clinical parameters such as tumor stage and grade, and primary tumor and metastasis tumor spreads. When we compare PAMOGK to 8 other state-of-the-art existing multi-omics clustering methods, PAMOGK consistently outperforms these in terms of its ability to partition patients into groups with different survival distributions. PAMOGK enables extracting the relative importance of pathways and molecular data types. PAMOGK is available at github.com/tastanlab/pamogk

Keywords: Patient Stratification · Graph Kernels · Multi-view Clustering · Pathways

1 Introduction

Cancer is a molecular diverse disease; within the same cancer type, patients bear different molecular alterations, which manifest themselves as different clinical trajectories [9, 53]. Finding subgroups of patients that show coherent molecular profiles is essential for developing better diagnostic tools and subtype-specific treatment strategies. Discovering coherent subgroups of patients with similar molecular profiles is also key to discovering the molecular mechanisms that drive these different subtypes to cancer.

The availability of multi-omics characterization of patients opens up possibilities for better stratification of cancer patients[48, 46, 9]. Towards this goal, several multi-omics clustering methods have been proposed (reviewed in [35]) to integrate these multi-dimensional data collected on patients. The simple form of integration is early integration. In this case, features derived from a single omic data are concatenated, and standard clustering is applied to this combined feature representation. However, this approach equally weighs each data type and suffers from a curse of dimensionality as the higher dimensional features dominate the clustering. There are sophisticated early integration approaches that aim to overcome these problems. iClusterBayes and its earlier variants [42, 30, 29] and LRACluster assume a latent lower dimensional distribution of data and uses regularization. A different strategy is to deploy late integration approaches, in which the samples are clustered with each omic data type separately, and the ensemble's cluster assignments are combined into a single solution. The consensus clustering by Monti et al. [31] is frequently used for cancer subtyping [16, 48]. PINS[34] and COCA[17] fall into this category as well. These approaches

⋆ These authors contributed equally to the work as first authors.
⋆⋆ Corresponding author: otastan@sabanciuniv.edu

have the drawback that they do not capture the correlations between the different data types. This leads to poor clustering when each view individually contains a weak signal.

Alternatively, several intermediate integration algorithms have been developed [36]. SNF [51] constructs a patient similarity network using each data type, and these similarities are then fused in a single similarity network through an algorithm based on message passing. Meng *et al.* [28] applies dimension reduction to the axes of maximal covariance between data types, JIVE [27] utilizes the variations in data. MCCA [54, 5] extends the canonical correlation analysis (CCA) [14] to a multi-view setting. There are also several algorithms, which are developed as generic multi-view algorithms ([57]). For example, [21] and [7] extend the spectral clustering [50], which relies on partitioning a similarity network of samples. There are also kernel-based multi-view clustering algorithms. Kernel methods are powerful methods, where the samples' similarities are implicitly calculated in a higher dimensional space [40]. Several generic multi-view kernel clustering methods (reviewed in [56]) have been developed where some have been applied to cancer subtyping. rMKL-LPP [45] extends the [23] multi-view kernel framework to the multi-omics clustering. A kernel matrix is computed from each omic data type, and a linear combination of kernels is sought for the clustering of the patients in kernel k-means. Localized multiple kernel k-means (LMKMM) [13] also assumes a linear combination of the views but learns a sample specific kernel matrix weight in a k-means framework.

Although corroborating multi-omics data is important to construct a better view of patient similarities, it might not be sufficient to boost the signal as often times only a small fraction of molecular alterations is common among the patients. Analyzing molecular data in the context of molecular networks is a widely used approach to overcome this sparsity problem (reviewed in [8]). In this work, we present PAMOGK, a multi-view kernel clustering approach, which integrates multi-omics patient data with pathways using graph kernels. PAMOGK represents each patient as a set of vertex labeled undirected graphs, where each graph represents the gene interactions in a biological pathway, and the vertex labels are assigned based on patient specific molecular alterations. To quantify patient similarity over a pathway and to attain an omic view, we introduce a novel graph kernel, smoothed shortest path graph kernel (SmSPK), which extends the shortest path graph kernel [4]. While existing graph kernels are designed to capture the topological similarities of the graphs, this kernel captures the similarities of the vertex label within the graph context. This allows us to capture patients' similarities that stem from the dysregulation of similar processes in the pathways. By utilizing multi-view kernel clustering approaches, PAMOGK stratifies patients into subgroups. The method also offers additional insights by showing how informative each pathway and data type is to the clustering process based on the assigned kernel weights.

We apply our methodology to kidney renal cell carcinoma(KIRC) data made available through the Cancer Genome Atlas Project (TCGA) [3]. We utilize patient somatic mutations, gene expression levels, and protein expression datasets. We find four patient subgroups that are significantly different in their survival times. Compared to the state-of-the-art multi-omics clustering methods, PAMOGK consistently outperforms in terms of its ability to partition into groups with different prognosis. PAMOGK also allows extracting the relative importance of pathways in the clustering process. PAMOGK is available at github.com/tastanlab/pamogk.

2 Methods

Given a set of cancer patients, $\mathcal{S}$, for which molecular profiles of the tumors are available, PAMOGK aims to stratify them into k subgroups through integrating pathways. Formally, we would like to find a partitioning $\mathcal{C}$ such that: $\mathcal{S}$ is grouped into k number of disjoint subsets C_i's where, $\mathcal{S} = \cup_{i=1}^{k} C_i$ and where $C_i \cap C_i = \emptyset$. In this section, we detail the steps of PAMOGK and data processing used in our experiment. Let M be the number of pathways, D be the molecular alteration types (mutations, altered expression, etc.) available for the patients and N be the number of patients.

2.1 PAMOGK Overview

PAMOGK involves three main steps (Figure 1). In the first step, each pathway is represented with an undirected graph. Next, for a given molecular alteration type, i.e., somatic mutations, a patient's molecular alterations are mapped on the pathway. These alterations constitute the patient-specific node labels of the patient's graph. This way, each patient is represented with a set of $M \times D$ graphs. To asses a pair of patients' similarity under a view, in the second step, the novel graph kernel, SmSPK, is computed o quantify a patient pair's similarity over a pathway and a molecular alteration type. Each $N \times N$ kernel matrix constitute a *view* to the patient similarities. In the final step, to stratify cancer patients into meaningful subgroups, these multiple kernels are input to a multi-view kernel clustering algorithm. In the following sections, we elaborate on each step of PAMOGK with more technical details.

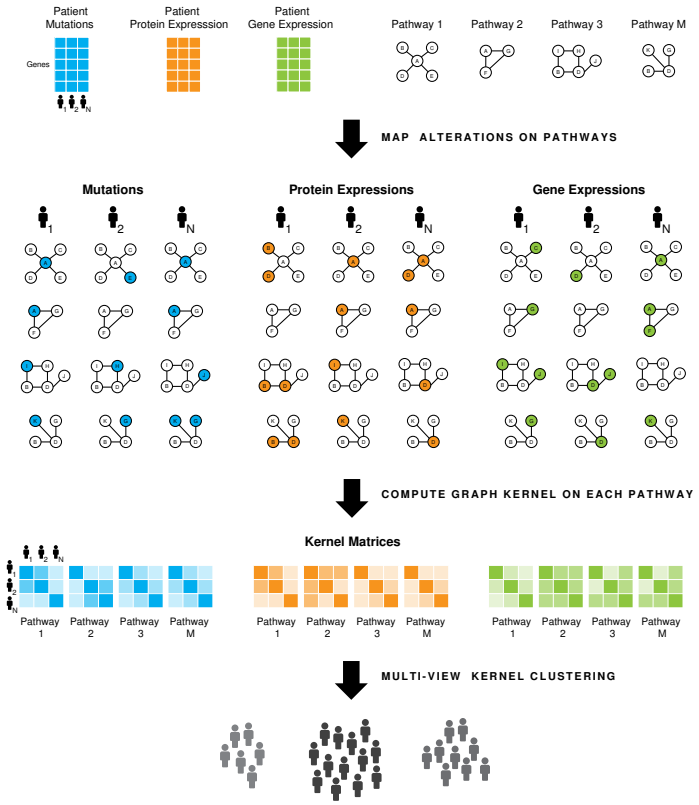


Fig. 1: PAMOGK framework. Each patient is represented with a set of undirected graphs, whose interactions are based on pathways and node labels are molecular alterations of the genes for that patient. Each pathway-omic pair constitute a view, and for each of these views, a patient-by-patient graph kernel matrix is computed. In the final step, these views are input to a multi-view kernel clustering method to obtain patient clusters.(Note that pathway graphs are shown smaller than usual due to size constraints.)

2.2 Step 1: Patient graph representation

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

interactions in the pathway while the vertex labels are different and are based on each patient's individual molecular alterations.

For a patient j, $\ell_i^{(j)}$ entries are assigned based on the patient's molecular alteration profile. For example, in the case of somatic mutations, if the corresponding gene k is mutated in patient j, label of value 1 is assigned to this gene (node), and 0 otherwise. At the end of this step, we have $N \times M \times D$ labeled pathway graphs.

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

In this step, we would like to asses the similarities of the patients on a given pathway for a given molecular data type. While typical kernels take vectors as input, a graph kernel function takes two graphs as input and returns a real-valued number that quantifies the similarity of two input graphs: $\mathcal{K} : \mathcal{G} \times \mathcal{G} \mapsto \mathbb{R}$ [49]. Powerful graph kernels are presented in earlier work [43, 4, 33]. However, these graph kernels are designed to compare graphs with different graph structures and to identify similarities and differences that arise from these different structures. In our case, though, we would like to compare graphs with identical topology but different node label distribution. For this, we devise a new graph kernel.

Inspired from the shortest path graph kernel [4], SmSPK makes use of all shortest paths of the graphs to characterize them. We also smooth the node labels of a patient in the pathway so that if two patients have alterations in genes in close proximity, they contribute to the similarity even though the set of altered genes are not identical. To propagate node labels along the pathway, we use the random walk with restart [8]. For a single graph indexed by g, the label propagation is performed by employing the following formula for all patients:

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

,where $\mathbf{S_g}^{(0)}$ is a patient-by-gene matrix which represents the labels of the vertices in the graph g at time $t = 0$ and each row (patient) is determined by $\ell_g^{(j)}$. In this case, $\mathbf{S}_{g,ji}^{(0)} = 1$ where j is index of the patient and i is index of the vertex. $\mathbf{S_g}^{(t)}$ is the node label matrix at time t. $\mathbf{A}_g$ is the degree normalized adjacency matrix of the pathway graph g. $\alpha \in [0, 1]$ is the parameter that defines the degree of smoothing. We iterate over propagation until convergence is attained. We assign the labels of the vertices of the graph based on the final $\mathbf{S}$.

Once we attain the label smoothed graphs of $G_g^{(i)}$ and $G_g^{(j)}$, we compute the similarities of these two graphs to each other as follows:

$$\mathcal{K}(G_g^{(i)}, G_g^{(j)}) = \sum_{p=1}^{P} \mathbf{s}_p^{(i)} . \mathbf{s}_p^{(j)} \tag{2}$$

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

For a given molecular alteration type and a pathway, we compute the kernel over all pairs of patients. $\mathbf{K}$ matrix is a symmetric $N \times N$ matrix, for which the i, j-th entry is the kernel function evaluated for patient i and patient j pair. By computing kernel matrices for each pathway and for each molecular alteration type, we obtain $M \times D$ kernel matrices.

2.4 Step 3: Multi-View Kernel Clustering

Each of the kernel matrices computed in the previous section represents a view of the patients' similarities. To integrate these views, we resort to existing multi-view kernel clustering approaches. We experiment with different approaches(See section 3.3); multiple kernel k-means with matrix-induced regularization (MKKM-MR) [25] performs the best. Thus the final model of PAMOGK uses MKKM-MR; yet, this step can be replaced by any multi-view clustering approach as long as the method accepts kernel matrices as input.

In this section, for completeness, we provide a brief overview of the selected multi-view kernel clustering methods with which we experimented.

Multiple Kernel K-Means with Matrix-Induced Regularization: MKKM-MR algorithm objective is to minimize sum-of-squared loss over the cluster assignments. To reduce redundancy among kernel matrices and enhance the diversity of the selected kernel matrices, MKKM-MR [25] uses the matrix-induced regularization. The algorithm solves the following optimization problem:

$$\begin{aligned} \min_{\mathbf{H}\in\mathbb{R}^{nxk},\gamma\in\mathbb{R}_+^m} \quad & \mathrm{Tr}(\mathbf{K}_\gamma(\mathbf{I}_n - \mathbf{H}\mathbf{H}^{\mathrm{T}})) + \frac{\lambda}{2}\gamma^{\mathrm{T}}\mathbf{M}\gamma \\ \text{s.t.} \quad & \mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I}_k \\ & \gamma^{\mathrm{T}}1_m = 1 \end{aligned} \tag{3}$$

Here, k is number of clusters, n denotes the number of samples, m is number of kernel matrices. $\mathbf{H}$ is the relaxed clustering assignment matrix, $\gamma = [\gamma_1, \gamma_2, \cdots, \gamma_m]$ are the weights of input kernel matrices. $\mathbf{K}_\gamma$ is the best kernel matrix, $\mathbf{M}$ is the matrix which measures the relation between kernel matrices. $\mathbf{I}_x$ is the x-by-x dimensional identity matrix, 1_m is m dimensional vector of ones. λ is the parameter that adjusts the trade-off between clustering cost and the regularization term.

Average Kernel K-Means (AKKM): Kernel k-means (KKM) [39] is simple but a strong baseline. Since it accepts single kernel matrix, we input the average of the kernel matrices. We will refer to this method as average kernel k-means (AKKM).

Localized multiple kernel k-means (LMKMM): LMKMM is another powerful method that optimizes not only the weight of the kernel matrices but also the weight of the samples[13]. We reimplemented LMKKM in Python, which is originally provided in Matlab and R.

2.5 Dataset and Data Preprocessing

Pathway data: As the pathway source, we use National Cancer Institute - Pathway Interaction Database (NCI-PID) at NDEXBio [38][4]. NCI-PID is a curated database with focus on processes that are relevant to cancer research (download date: Apr 24, 2019). We filter out a pathway if it does not contain any overlapping gene with the omic data genes, which leaves out 165 pathways.

Patient molecular and clinical data: he molecular and clinical data for KIRC is obtained from TCGA PanCancer project [52]. We retrieve the data directly from Synapse[5]. We only consider the primary solid tumour samples and make use of three different molecular data types that can directly be mapped to pathways: somatic mutations, transcriptomics and proteomics data. The transcriptomic data include the RNAseq gene expression levels, while protein expression is quantified through Reverse Phase Protein Array (RPPA). The exact data files are listed in Supplementary Table 1. We eliminate genes that are not expressed in more than half of the samples. In processing the proteomics data, we only consider the unphoshorylated protein expressions. We obtain the clinical data of cancer patients from TCGA at the Genomics Data Commons (GDC) data portal[6]. For patients who have passed away, days to death is used for calculating survival time, while for patients with censored information, the days to the last follow-up information is used. The final patient set is formed with patient tumour samples where all three types of molecular data are present and for which the survival information of the patient is available. This results in 361 patients; 236 of them are right-censored, and 125 of them had passed away.

Assigning node labels based on molecular alterations: The gene and protein expression values are converted to z-scores and for each with z-score greater than 1.96(which stands for 95% confidence), the gene(the

[4] https://ndexbio.org/#/networkset/8a2d7ee9-1513-11e9-bb6a-0ac135e8bacf
[5] https://www.synapse.org/#!Synapse:syn300013
[6] https://portal.gdc.cancer.gov

protein) is considered overexpressed while the genes(the proteins) with z-score lower than -1.96 is considered underexpressed. Thus from the three different omic data sources, five different types of alterations are defined: a somatic mutation in a gene, over and underexpression of a gene, over and underexpression of a protein. In each case, the patient node label is assigned as a binary label based on the presence or absence of the molecular alteration.

3 Results and Discussion

3.1 Experimental Set up

noindent We apply PAMOGK to discover different subgroups of KIRC patients. The dataset contains 361 patients whose molecular profiles come from three different data types: somatic mutation, gene expression, and protein expression. We define five different molecular alteration types based on these three types of omics data (see Section 2.5). In each case, the graph node labels are binary labels based on the presence of molecular alterations. We compute one kernel matrix for each pathway-molecular alteration type; this results in 825 kernels (165 pathways x 5 molecular alterations), each one of which constitutes a distinct view.

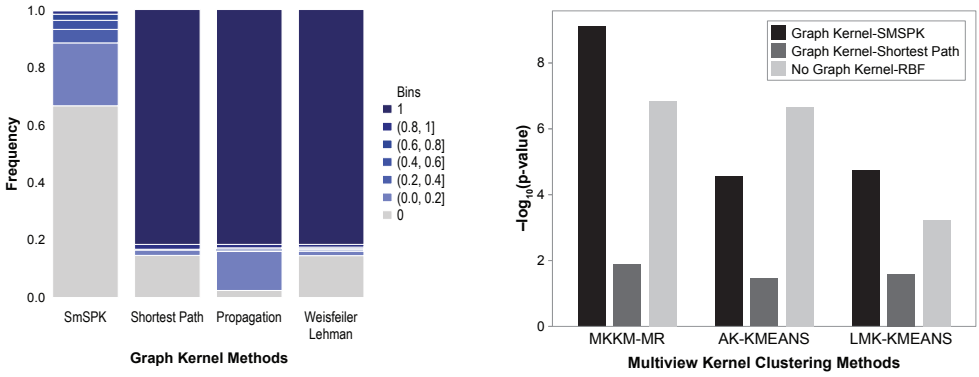
Throughout all experiments, we evaluate four different cluster numbers, $k = 2, 3, 4, 5$. When computing SmSPK, we try 12 different alpha α values (Supplementary Table 2). We conduct experiments by using different multi-view clustering methods. These include average kernel k-means(AKKM), LMKKM, and MKKM-MR. If a pathway kernel includes a few or no altered genes, we eliminate it before inputting it into multi-view kernel clustering methods to increase time efficiency. The criteria for this is to eliminate those whose nonzero entries constitute 1% of all entries. The parameter λ in MKKM-MR is chosen using grid-search (Supplementary Table 2).

We evaluate the clustering solutions through survival analysis in accordance with previous work [2, 22, 1, 12]. We compare the survival distributions of the clusters using Kaplan-Meier (KM) survival curves [20] and log-rank test's p-value [15]. In the log-rank test, we test whether there is a statistical difference between the survival times of the clusters. In comparing alternative methods, we use the p-value of this log-rank test as the performance criteria.

3.2 Assessing the Need of a new Graph Kernel

Constructing kernels, which reflect the similarity of patients, is a crucial step of PAMOGK. First, we would like to understand whether there is any merit in using SmSPK as opposed to deploying an already existing and powerful graph kernel. The motivation behind proposing a new kernel is that the existing graph kernels are designed to capture topological similarities. Since we compare the two patients on the same pathway, the structure of graphs shall always be the same. On the other hand, the node label distribution is different as it is patient specific. Thus, the existing graph kernels computed over the same pathway will consider patients as overly similar and would not serve our purpose. To check if this intuition holds, we analyze the distribution of the kernel values computed over all the pathways and the overexpressed molecular alteration type. Since the overexpressed genes are the densest kernels, we choose this data type. We compare SmSPK with the shortest path kernel [4], propagation kernel [33] and Weisfeiller Lehman subtree kernel [43]. We use the implementation provided by the Grakel library [44]. In order to make the comparison fair, we also apply smoothing and choose the results with the best smoothing parameter assignment (Supplementary Table 2).

To analyze the distribution of kernel values assigned to patients by each of the different kernels, we bin the kernel matrix entries into groups for each kernel and calculate the average of the bins across different kernels (Figure 2a). Next, for a kernel matrix computed over a pathway, we calculate the frequency of entries of the kernel in a bin. We repeat this for all the kernels and calculate the average frequency for each bin. Figure 2a shows, for each graph kernel, how the kernel values are distributed on average. All the kernels other than SmSPK, assign patient similarities of 1 very frequently (the darkest bin). Five randomly chosen



(a) **(b)**

Fig. 2: a) The average frequency of patient similarities for different kernels over all pathways with the overexpression molecular data. **b)** The log-rank test p-values obtained with different choices of multi-view clustering algorithms and kernels. Kernel construction methods consist of SmSPK(our method), shortest path graph kernel [4] and radial basis function (RBF) kernel. The clustering methods include average kernel k-means (AKKM), localized multiple kernel k-means (LMKKM) [13] , multiple kernel k-means with matrix-induced regularization (MKKM-MR) [25]. MKKM-MR and SmSPK combination corresponds to PAMOGK.

kernel matrices computed with each kernel are provided in Supplementary Figure 2 which clearly shows how these kernel values are excessively 1. These results confirm our intuition that due to the identical graph structures, the existing graph kernels are unable to distinguish patients with different molecular alterations on the same pathway graph. Therefore, the use of a new graph kernel, SmSPK is justified.

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

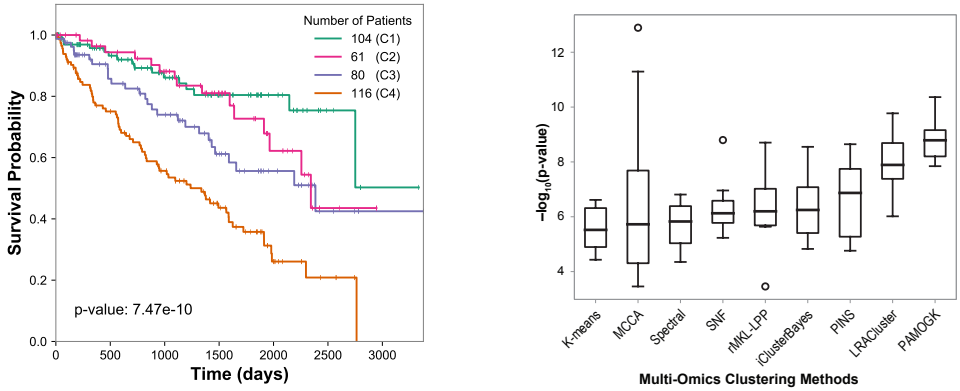
To determine the multi-view kernel clustering algorithm to be used in PAMOGK, we experiment with alternatives. The multi-view kernel clustering methods we analyze include the MKKM-MR [25], AKKM and LMKKM [13] (see Section 2.4). For each method, we report the best clustering solution, which is determined based on the lowest p-value attained in the log-rank test on the survival distributions of clusters. In each experiment, we allow the methods to choose from a set of predetermined values for each of the hyperparameters. These include k for clustering, the smoothing parameter α for SmSPK, gamma(γ) parameter for RBF kernel, λ for MKKM-MR.

We test these clustering methods coupling them with alternative kernels to ensure that the resulting performances are not due to SmSPK. As the alternative graph kernel function, we use the shortest path graph kernel, the closest alternative graph kernel. Since we establish that other graph kernels fail to capture patient similarities in the previous experiments (see Section 3.2), we do not include them in the experiments. We also include the radial basis function (RBF) kernel to judge if there is any need at all for a graph kernel. When RBF kernels are computed, they are directly evaluated on the omic data. Thus, they are computed over all the genes regardless of their participation in a pathway. The gamma values of RBF is determined by the median heuristic [41] (Supplementary Table 2).

Figure 2b summarizes the results in these experiments for the best clustering solution, where $k = 4$. (Other k values are provided in Supplementary Figure 1). When comparing the three multi-view kernel clustering methods, we observe that MKKM-MR produces the best results regardless of the kernel type employed. LMKKM and AKKM yield similar results with the difference that LMKKM performs slightly better when SmSPK is used, and AKKM performs better when the RBF kernel is employed. We also observe that regardless of the clustering method, the shortest path graph kernel yields the poorest results. This can be explained based on the previous remark that this graph kernel is formulated to distinguish graphs with different topologies. Thirdly, although the use of RBF kernel generally yields good results, the integration of

pathway information through SmSPK brings an improvement to the cluster separations in terms of survival. Overall, PAMOGK that uses the SmSPK with MKKM-MR multi-view clustering outperforms all the other combinations of clustering and kernel alternatives. Thus, we employ MKKM-MR in PAMOGK.

The best clustering solution by PAMOGK is obtained when $k = 4$, smoothing parameter, α is set to 0.3, and λ for MKKM-MR is set to 8. The KM plot of the resulting clustering is provided in Figure 3a. The survival distributions significantly differ (log-rank test, p-value=7.4e-10). We should note that the solution with $k = 3$ is also quite good, p-value=8.53e-10 (Supplementary Figure 3b).



(a) **(b)**

Fig. 3: a) Kaplan-Meier survival curves of the best clustering solution for KIRC. The p-value was obtained from a log-rank test between the groups **b)** Comparison of PAMOGK with the multi-omics clustering methods over 10 different trials. Each trial contains a random subsample of KIRC patients. (Note that PINS method results are over 9 experiments since in one of trial, it did not return a result.)

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

Performance comparison: We compare PAMOGK with eight other multi-omics methods. These include k-means [26], MCCA [54], LRACluster [55], rMKL-LPP [45], iClusterBayes [29], PINS [34], SNF [51], and finally Spectral Clustering [58]. These methods cover all methods that are included in a recent comparative benchmark study by Rappoport et al. [36] with the exception of multiNMF [24], which we are not able to run properly. In running these algorithms, we set the maximum number of clusters to five and choose the other parameter configurations for each algorithm exactly as in the benchmark study [36].

To asses the performance of different methods, we repeatedly subsample the original patient set, and for each subsample, run the algorithms to find the patient clusters. Each subsample contained 300 patients. Due to prohibiting runtime of iClusterBayes, we were able to conduct this experiment 10 times. The distribution of log-rank test p-values attained by each method is displayed in Figure 3b. The comparison over ten runs shows that PAMOGK is the best performer among the nine methods. Not only the median performance is high, but even the 90-th percentile of the trials is superior to almost all methods. It also displays low variance across different runs. For all methods, for all trials, the resulting clusters are balanced in terms of the number of patients participating in the clusters except two trials of MCCA. Log-rank test is known to result in unrealistically low p-values when one of cluster size is small [47]. In those two trials, MCCA's extremely low p-values are due to cluster sizes of 9 and 14.

Table 1: The runtime in seconds for clustering 361 KIRC patients with the three types of omic data.

Method	**PAMOGK**	**LRACluster**	**PINS**	**SNF**	**rMKL-LPP**	**iClusterBayes**	**Spectral**	**K-means**	**MCCA**
Time	$1,472$	289	56	7	109	$10,898$	3	47	6

Runtime comparisons: We conduct a runtime comparison of the algorithms for clustering all the KIRC patients using the three different data types. PAMOGK demands more time to run in comparison to the other methods, with the exception of iClusterBayes. This is because it calculates many more views of the data based on pathways. A second time limiting step is the weight optimization of the kernels in the MKKM-MR algorithm. Despite these additional requirements, the runtime is within reasonable limits, and a typical run takes less than 30 minutes without any parallelization. Replacing the multi-view clustering step with a less demanding algorithm and parallelization could reduce the runtime. Experiments are conducted on the following system configuration: CPU: Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz CPU. Memory: 256Gb. Operating system: Ubuntu 16.04.4 LTS.

3.5 Detailed Analysis of KIRC Subgroups Discovered by PAMOGK

Table 2: Statistical analysis of other clinical variables.

Clinical Parameter	Test	p-value
Age	One-way ANOVA	1.430e-01
Gender	χ^2	2.510e-01
Stage	χ^2	1.087e-09
Primary Tumor Pathologic Spread	χ^2	3.801e-08
Distant Metastasis Pathologic Spread	χ^2	3.163e-06
Neoplasm Histologic Grade	χ^2	1.532e-09

KIRC Subgroups' Associations with Other Clinical Parameters: We analyze the association of clinical parameters of the discovered subgroups other than survival. The parameters include age, gender, tumor stage, primary tumor pathological spread, distant metastasis pathological spread and neoplasm histological grade. The association of categorical variables are determined using χ^2 test while the continuous variables are tested with one-way ANOVA. We find no statistical difference in terms of age (p-value = 0.143) and gender (p-value = 0.251). All the other clinical parameters differ across groups at a statistically significant level (see Table 2). The distribution of these variables across groups are provided in Supplementary Section 3.

The best prognosis group is cluster 1, and the worst prognosis group is cluster 4 (Figure 3a). There are clear differences between these two groups in terms of these additional clinical parameters. More specifically, 53.8% of the patients in cluster 1 are in stage I, whereas 69.8% of the patients in Cluster 4 are either in stage III or Stage IV (Supplementary Table 5). Also, half of the patients in cluster 1 have primary tumor T1, whereas 60.3% of the patients in cluster 2 have primary tumor T3 (see Supplementary Table 6). While only 7.69% of the patients of cluster 1 have distant metastasis, this ratio is 32.8% for cluster 4 patients (Supplementary Table 7). Finally, the fraction of cluster 1 patients with histologic grade G1 is 59.6%, and those with G4 is 5.7%. For cluster 4, the percentage for G1 drops to 18% and G4 increases to 35.3% (Supplementary Table 8). For all prognostic tumor-related features, cluster 1 always has more patients with a lower degree stage and grade, whereas cluster 4 always has more patients with a higher degree stage and grade. Overall, this analysis provides additional evidence that PAMOGK partitions KIRC patients into clinically meaningful subgroups.

Influential pathways and data types: By inspecting the assigned kernel weights, we can quantify the relative importance to pathways and molecular data types. For KIRC ($k = 4$), the *IL-6 mediated signaling events* pathway and gene overexpression kernel emerge as the most important pathway-molecular alteration pair (see Supplementary Figure 4 for the top 10 pairs). By averaging the weights associated with each omic data type, we find that the gene expression is the top important data type while protein expression data have almost no effect on the clustering (Supplementary Figure 5). This could be arising from the fact that the protein expression data covers only a small number of proteins.

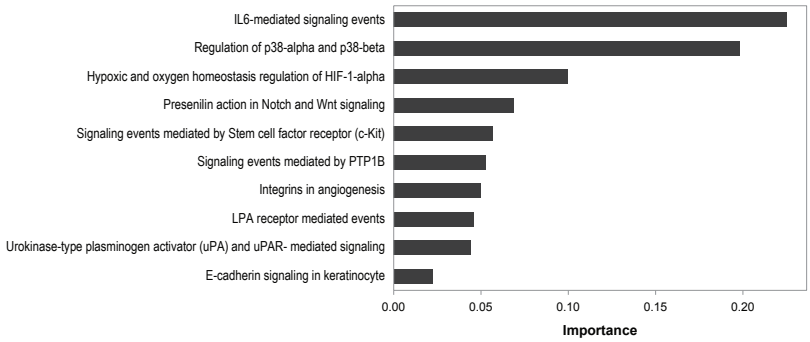


Fig. 4: Top 10 pathways, which have the highest relative importance in clustering.

The top relevant pathway in clustering the patients into subgroups emerges as the *IL6-Mediated signaling events* (Figure 4). Interleukin-6 (IL6) is a cytokine that regulates transcriptional acute and chronic inflammatory responses. [19] reports that IL6 and its related signaling pathways are key factors of development of KIRC and its spread. Others [32, 37, 11] also report the prognostic value of IL6 for KIRC. The second most influential pathway is the *regulation of p38-alpha and p38-beta* pathway. p38 mitogen-activated protein kinases (also known as MAPKs) regulate critical processes such as cell proliferation, cell differentiation, cell death, cell migration, and invasion [6]. [18] shows that the inactivation of this pathway suppresses KIRC growth. Mutations in VHL and PTEN are reported as drivers of KIRC [3]. VHL is part of the *Hypoxic and oxygen homeostasis regulation of HIF-1-alpha* pathway and HIF-1 is part of the *Signaling events mediated by Stem cell factor receptor (c-Kit)* pathway. Both of these pathways are among the top 5 influential pathways according to our analysis.

4 Conclusion and Future Work

We present PAMOGK for discovering subgroups of patients which not only operates by integrating different omics data sets derived from patients but also incorporates existing knowledge on biological pathways. To corroborate these data sources, we develop a novel graph kernel that evaluates patient similarities based on their molecular alterations in the context of known pathways. We employ a multi-view kernel clustering technique to leverage views constructed by different molecular alteration types and pathways. Our results indicate that the suggested methodology, when applied to KIRC, results in patient clusters that differ significantly in their survival distributions and other clinical parameters. The proposed methodology also provides quantitative evidence for the decisive role of known driver pathways on the clustering process. We further show that PAMOGK performs better compared to the state-of-the art multi-omics approaches.

One limitation of the current work is that we used the bulk expression results provided by the TCGA project. However, it is known that there could be high level of intra-tumour heterogeneity [10] and the bulk tumour might include a diverse collection of cells harbouring distinct molecular signatures. A future work would be to adapt PAMOGK framework to single cell measurements as they become available for large cohorts of patients.

The work can be extended in other several directions. In this current work we use binary node labels to indicate if the gene for that patient is perturbed. PAMOGK can be easily extended to accept continuous node labels to incorporate the extent of the molecular alterations. The proposed kernel matrix characterizes the similarities of patients based on the shortest path on the graphs. Other graph kernels can be devised to capture patient similarities on the graphs using other topological features of the graphs. Furthermore, in the present study, we ignore the direction information of the edges in the pathways. A kernel that explicitly accounts for edge directions can be more devised. In lieu or addition to the pathways, protein-protein interaction networks can be used to generate views, which will be explored in future work.

5 Acknowledgements

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under Grant #117E140. Oznur Tastan thanks to the Science Academy of Turkey under The Young Scientist Award Program (BAGEP). Yasin Tepeli and Ali Burak Unal acknowledge TUBITAK-BIDEB for the 2210-A scholarship program.

References

1. et al., C.J.R.: The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. Cell Reports **23**(1), 313–326.e5 (Apr 2018). https://doi.org/10.1016/j.celrep.2018.03.075, https://doi.org/10.1016/j.celrep.2018.03.075
2. et al., J.L.: An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell **173**(2), 400–416.e11 (Apr 2018). https://doi.org/10.1016/j.cell.2018.02.052, https://doi.org/10.1016/j.cell.2018.02.052
3. et al., T.C.G.A.R.N.: Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature **499**(7456), 43–49 (Jun 2013). https://doi.org/10.1038/nature12222, https://doi.org/10.1038/nature12222
4. Borgwardt, K.M., Kriegel, H.P.: Shortest-path kernels on graphs. In: Data Mining, Fifth IEEE International Conference on. pp. 8–pp. IEEE (2005)
5. Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K.: Multi-view clustering via canonical correlation analysis. In: Proceedings of the 26th annual international conference on machine learning. pp. 129–136. ACM (2009)
6. Chen, Z., Gibson, T.B., Robinson, F., Silvestro, L., Pearson, G., Xu, B.e., Wright, A., Vanderbilt, C., Cobb, M.H.: Map kinases. Chemical reviews **101**(8), 2449–2476 (2001)
7. Chikhi, N.F.: Multi-view clustering via spectral partitioning and local refinement. Information Processing & Management **52**(4), 618–627 (Jul 2016). https://doi.org/10.1016/j.ipm.2015.12.007, https://doi.org/10.1016/j.ipm.2015.12.007
8. Cowen, L., Ideker, T., Raphael, B.J., Sharan, R.: Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics **18**(9), 551 (2017)
9. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al.: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature **486**(7403), 346 (2012)
10. Dagogo-Jack, I., Shaw, A.T.: Tumour heterogeneity and resistance to cancer therapies. Nature reviews Clinical oncology **15**(2), 81 (2018)
11. Fu, Q., Chang, Y., An, H., Fu, H., Zhu, Y., Xu, L., Zhang, W., Xu, J.: Prognostic value of interleukin-6 and interleukin-6 receptor in organ-confined clear-cell renal cell carcinoma: a 5-year conditional cancer-specific survival analysis. British journal of cancer **113**(11), 1581 (2015)
12. Gabasova, E., Reid, J., Wernisch, L.: Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. PLoS computational biology **13**(10), e1005781 (2017)
13. Gönen, M., Margolin, A.A.: Localized data fusion for kernel k-means clustering with application to cancer biology. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1. pp. 1305–1313. NIPS'14, MIT Press, Cambridge, MA, USA (2014), http://dl.acm.org/citation.cfm?id=2968826.2968972
14. H., H.: Relations between two sets of variables. Biometrika pp. 321–327 (1936)
15. HARRINGTON, D.P., FLEMING, T.R.: A class of rank test procedures for censored survival data. Biometrika **69**(3), 553–566 (1982). https://doi.org/10.1093/biomet/69.3.553, https://doi.org/10.1093/biomet/69.3.553
16. Hayes, D.N., Monti, S., Parmigiani, G., Gilks, C.B., Naoki, K., Bhattacharjee, A., Socinski, M.A., Perou, C., Meyerson, M.: Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. Journal of Clinical Oncology **24**(31), 5079–5090 (2006)
17. Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D., Niu, B., McLellan, M.D., Uzunangelov, V., et al.: Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell **158**(4), 929–944 (2014)
18. Huang, D., Ding, Y., Luo, W.M., Bender, S., Qian, C.N., Kort, E., Zhang, Z.F., VandenBeldt, K., Duesbery, N.S., Resau, J.H., et al.: Inhibition of mapk kinase signaling pathways suppressed renal cell carcinoma growth and angiogenesis in vivo. Cancer research **68**(1), 81–88 (2008)
19. Kamińska, K., Czarnecka, A.M., Escudier, B., Lian, F., Szczylik, C.: Interleukin-6 as an emerging regulator of renal cell cancer. Urologic Oncology: Seminars and Original Investigations **33**(11), 476–485 (Nov 2015). https://doi.org/10.1016/j.urolonc.2015.07.010, https://doi.org/10.1016/j.urolonc.2015.07.010
20. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association **53**(282), 457–481 (Jun 1958). https://doi.org/10.1080/01621459.1958.10501452, https://doi.org/10.1080/01621459.1958.10501452
21. Kumar, A., Rai, P., Daumé, III, H.: Co-regularized multi-view spectral clustering. In: Proceedings of the 24th International Conference on Neural Information Processing Systems. pp. 1413–1421. NIPS'11, Curran Associates Inc., USA (2011), http://dl.acm.org/citation.cfm?id=2986459.2986617

22. Liang, R., Wang, M., Zheng, G., Zhu, H., Zhi, Y., Sun, Z.: A comprehensive analysis of prognosis prediction models based on pathway-level, gene-level and clinical information for glioblastoma. International Journal of Molecular Medicine (Jul 2018). https://doi.org/10.3892/ijmm.2018.3765, https://doi.org/10.3892/ijmm.2018.3765
23. Lin, Y.Y., Liu, T.L., Fuh, C.S.: Multiple kernel learning for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(6), 1147–1160 (2010)
24. Liu, J., Wang, C., Gao, J., Han, J.: Multi-view clustering via joint nonnegative matrix factorization. In: Proceedings of the 2013 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics (May 2013). https://doi.org/10.1137/1.9781611972832.28, https://doi.org/10.1137/1.9781611972832.28
25. Liu, X., Dou, Y., Yin, J., Wang, L., Zhu, E.: Multiple kernel k-means clustering with matrix-induced regularization. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. pp. 1888–1894. AAAI'16, AAAI Press (2016), http://dl.acm.org/citation.cfm?id=3016100.3016163
26. Lloyd, S.: Least squares quantization in PCM. IEEE Transactions on Information Theory **28**(2), 129–137 (Mar 1982). https://doi.org/10.1109/tit.1982.1056489, https://doi.org/10.1109/tit.1982.1056489
27. Lock, E.F., Hoadley, K.A., Marron, J.S., Nobel, A.B.: Joint and individual variation explained (jive) for integrated analysis of multiple data types. The annals of applied statistics **7**(1), 523 (2013)
28. Meng, C., Kuster, B., Culhane, A.C., Gholami, A.M.: A multivariate approach to the integration of multi-omics datasets. BMC bioinformatics **15**(1), 162 (2014)
29. Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K.S., Hilsenbeck, S.G.: A fully bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics **19**(1), 71–86 (May 2017). https://doi.org/10.1093/biostatistics/kxx017, https://doi.org/10.1093/biostatistics/kxx017
30. Mo, Q., Wang, S., Seshan, V.E., Olshen, A.B., Schultz, N., Sander, C., Powers, R.S., Ladanyi, M., Shen, R.: Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences **110**(11), 4245–4250 (2013)
31. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning **52**(1-2), 91–118 (2003)
32. Negrier, S., Perol, D., Menetrier-Caux, C., Escudier, B., Pallardy, M., Ravaud, A., Douillard, J.Y., Chevreau, C., Lasset, C., Blay, J.Y.: Interleukin-6, interleukin-10, and vascular endothelial growth factor in metastatic renal cell carcinoma: prognostic value of interleukin-6—from the groupe francais d'immunotherapie. Journal of clinical oncology **22**(12), 2371–2378 (2004)
33. Neumann, M., Garnett, R., Bauckhage, C., Kersting, K.: Propagation kernels: Efficient graph kernels from propagated information. Mach. Learn. **102**(2), 209–245 (Feb 2016). https://doi.org/10.1007/s10994-015-5517-9, http://dx.doi.org/10.1007/s10994-015-5517-9
34. Nguyen, T., Tagett, R., Diaz, D., Draghici, S.: A novel approach for data integration and disease subtyping. Genome research **27**(12), 2025–2039 (2017)
35. Rappoport, N., Shamir, R.: Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic acids research **46**(20), 10546–10562 (2018)
36. Rappoport, N., Shamir, R.: Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Research **46**(20), 10546–10562 (10 2018). https://doi.org/10.1093/nar/gky889, https://doi.org/10.1093/nar/gky889
37. Rini, B.I., Campbell, S.C., Escudier, B.: Renal cell carcinoma. The Lancet **373**(9669), 1119 – 1132 (2009). https://doi.org/https://doi.org/10.1016/S0140-6736(09)60229-4, http://www.sciencedirect.com/science/article/pii/S0140673609602294
38. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., Buetow, K.H.: PID: the pathway interaction database. Nucleic Acids Research **37**(suppl_1), D674–D679 (Oct 2008). https://doi.org/10.1093/nar/gkn653, https://doi.org/10.1093/nar/gkn653
39. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation **10**(5), 1299–1319 (Jul 1998). https://doi.org/10.1162/089976698300017467, https://doi.org/10.1162/089976698300017467
40. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA (2001)
41. Sejdinovic, D., Gretton, A., Bergsma, W.: A kernel test for three-variable interactions (2013)
42. Shen, R., Olshen, A.B., Ladanyi, M.: Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics **25**(22), 2906–2912 (09 2009). https://doi.org/10.1093/bioinformatics/btp543, https://doi.org/10.1093/bioinformatics/btp543
43. Shervashidze, N., Schweitzer, P., Leeuwen, E.J.v., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler-lehman graph kernels. Journal of Machine Learning Research **12**(Sep), 2539–2561 (2011)
44. Siglidis, G., Nikolentzos, G., Limnios, S., Giatsidis, C., Skianis, K., Vazirgiannis, M.: Grakel: A graph kernel library in python. arXiv preprint arXiv:1806.02193 (2018)
45. Speicher, N.K., Pfeifer, N.: Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics **31**(12), i268–i275 (2015)

46. Toss, A., Cristofanilli, M.: Molecular characterization and targeted therapeutic approaches in breast cancer. Breast Cancer Research **17**(1), 60 (2015)
47. Vandin, F., Papoutsaki, A., Raphael, B.J., Upfal, E.: Accurate computation of survival statistics in genome-wide studies. PLOS Computational Biology **11**(5), e1004071 (May 2015). https://doi.org/10.1371/journal.pcbi.1004071, https://doi.org/10.1371/journal.pcbi.1004071
48. Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B.A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H.S., Hodgson, J.G., James, C.D., Sarkaria, J.N., Brennan, C., Kahn, A., Spellman, P.T., Wilson, R.K., Speed, T.P., Gray, J.W., Meyerson, M., Getz, G., Perou, C.M., Hayes, D.N.: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell **17**(1), 98–110 (Jan 2010). https://doi.org/10.1016/j.ccr.2009.12.020, https://doi.org/10.1016/j.ccr.2009.12.020
49. Vishwanathan, S.V.N., Borgwardt, K.M., Kondor, I.R., Schraudolph, N.N.: Graph kernels. CoRR **abs/0807.0093** (2008), http://arxiv.org/abs/0807.0093
50. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing **17**(4), 395–416 (2007)
51. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.: Similarity network fusion for aggregating data types on a genomic scale. Nature methods **11**(3), 333 (2014)
52. Weinstein, J.N., , Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. Nature Genetics **45**(10), 1113–1120 (Sep 2013). https://doi.org/10.1038/ng.2764, https://doi.org/10.1038/ng.2764
53. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Network, C.G.A.R., et al.: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113 (2013)
54. Witten, D.M., Tibshirani, R.J.: Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical Applications in Genetics and Molecular Biology **8**(1), 1–27 (Jan 2009). https://doi.org/10.2202/1544-6115.1470, https://doi.org/10.2202/1544-6115.1470
55. Wu, D., Wang, D., Zhang, M.Q., Gu, J.: Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC Genomics **16**(1) (Dec 2015). https://doi.org/10.1186/s12864-015-2223-8, https://doi.org/10.1186/s12864-015-2223-8
56. Yu, S., Tranchevent, L., Liu, X., Glanzel, W., Suykens, J.A., De Moor, B., Moreau, Y.: Optimized data fusion for kernel k-means clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(5), 1031–1039 (2011)
57. Zhao, J., Xijiong, X., Xu, X., Sun, S.: Multi-view learning overview: Recent progress and new challenges. Information Fusion **38** (02 2017). https://doi.org/10.1016/j.inffus.2017.02.007
58. Zhou, D., Burges, C.J.: Spectral clustering and transductive learning with multiple views. In: Proceedings of the 24th international conference on Machine learning. pp. 1159–1166. ACM (2007)