PRADC: Prostate adenocarcinoma subtype discovery on multi-omics data through clustering

1nd Renato Avellar Nobre
Computer Science Department
Università degli Studi di Milano
Course “Bioinformatics” - Exam Project, A.Y. 22/23

***Index Terms*—bioinformatics, clustering, multi-omics data, cancer, similarity networks**

I. Introduction

The Cancer Genome Atlas Program (TCGA) [1] is a landmark cancer genomics program that originated in the United States, bringing together researchers from diverse disciplines. The TCGA over the years generated over 2.5 petabytes of publicly available genomic, epigenomic, transcriptomic, and proteomic data [1]. The data has already improved the ability to diagnose, treat, and prevent cancer [1].

Among the multiple cancer genomics presented in the TCGA, prostate cancer is a prominent study case, being the second most common cancer in men and the fourth most common tumour type worldwide [2]. In 2020, an estimated 1,414,259 people were diagnosed with prostate cancer [3], which might have multiple genetic and demographic factors contributing to its high incidences, such as age, family history, genetic susceptibility, and race [4].

Prostate cancer is also highly heterogeneous, ranging from asymptomatic to a rapidly fatal systemic malignancy [5]. Mortality varies dramatically with age and detection stage, increasing exponentially when detected later in older patients [6]. Several genetic and epigenetic alterations are highly prevalent and appear to be essential factors in the tumorigenesis and progression of cancer [5]. Therefore, developing more precise tools to identify the signs of prostate cancer and the relation of the disease subtypes with genetic factors is essential to reduce the dramatic consequences.

The TCGA Network, in a groundbreaking research [7], kick-started the studies of using comprehensively integrated diverse omics (referred to as multi-omics) data types to assess the robustness of previously defined prostate cancer subtypes [7]. Their work gained further insight into the molecular-genetic heterogeneity of primary prostate cancer by comprehensively characterizing 333 primary prostate cancers using seven genomics platforms. Their technique used an integrative clustering model, denominated iCluster [8] on multi-omics data and revealed novel molecular features that provide a better understanding of the disease [7].

With the aforementioned in mind, this project proposes and explores additional methods to identify the prostate cancer molecular disease subtypes using multi-omics data, considering the three subtypes identified by the TCGA Network with the iCluster technique in [7]. Since multiple genetics and demographic factors can contribute to a high incidence of prostate cancer [4], identifying disease subtypes aims to find homogeneous groups of patients with similar clinical or molecular characteristics [9], grouping similar individuals may help better to predict their future health status (prognosis) and also direct the more appropriate therapy [9]. This is a crucial aspect of a field known as “personalized medicine”, which involves customizing medical treatment based on factors like genetics, environment, and lifestyle of individuals [9].

Among multiple methods, one traditional way of finding similarity between groups is through clustering [8]. Specifically, this work uses a multi-omics dataset composed of prostate cancer samples and attempts to find patient disease subtypes through multiple methods of integrating different biological data sources and applying state-of-the-art clustering approaches. The work integrates the multi-omics data using an average baseline integration and a graph-based integration algorithm denominated Similarity Network Fusion (SNF) [10]. For the clustering aspect, we cluster the patients in the similarity networks by using both the Partitioning Around Medoids (PAM) [11] and Spectral Clustering [12] methods. After the cluster is performed, we can evaluate the computed results aligned with the disease subtypes for prostate cancer found by the TCGA Network with the iCluster technique [7].

Results indicated a more significant clustering power for the multi-omics data integrated with SNF, while the simple integration method can decrease the power of a single-omics source. Additionally, the results favoured the PAM algorithm compared to spectral clustering. Therefore our more robust cluster was generated with the SNF integration with PAM clustering, in which the metrics demonstrated overlap with the iClusting disease subtypes, although limited.

A. Related Works

In the last few years, we have seen many works using TCGA data to study multiple types of cancer. Specifically, several works have been published using the TCGA to study prostate cancer [5]–[7], [13]–[15]. This section presents an overview of the results and challenges of the work in this area.

Robinson *et al.* [13] conducted a study that shows how broad genomic principles developed for the TCGA can be applied to a group of metastatic tumours with clinical relevance. The study set up a collaborative clinical sequencing system involving multiple institutions to perform whole-exome and transcriptome sequencing of biopsies taken from bone or soft tissue tumours in 150 individuals affected by metastatic, castration-resistant prostate cancer (mCRPC). Long *et al.* [14] conducted a study investigating the connections between DNA CpG methylation density and distribution, gene expression, and tumour outcomes using the extensive genomic data available in the TCGA. Specifically, the study focused on the prostate cancer cohort within the TCGA and examined how the expression of genes involved in DNA methylation control, known targets of DNA methylation, and tumour status are related. The findings revealed that genes responsible for producing S-adenosyl-L-methionine (SAM) are associated with altered expression of DNA methylation targets in a subset of aggressive tumours.

Boldrini *et al.* [6] studied using transcriptome data from a group of 243 patients from the TCGA database. They specifically focused on identifying essential regulatory genes involved in the tumour microenvironment's proliferation activity, stress regulation, and inflammation regulation. The data was utilized for training multi-dimensional scaling (MDS) models and creating a combined score that could improve the prediction of patients' survival and disease-free intervals. The results suggested that selecting patients with a high level of proliferation and DNA repair activity could enable early identification of aggressive prostate cancer with potential metastatic development. Singh *et al.* [5] developed a review to summarize the critical mechanisms involved in the epigenetic interactions and highlight their potential as functional biomarkers for clinical use. Their analysis of prostate cancer data from the TCGA revealed that the expression of specific miRNAs is associated inversely with DNA methylation, emphasizing their importance in prostate cancer.

Finally, Huang *et al.* [15] conducted a differential expression analysis using the TCGA database to investigate the relationship between Long non-coding RNAs (LncRNAs) and prostate cancer. They analyzed the data of 500 prostate cancer patients and identified 6 LncRNAs as independent prognostic factors. Their study demonstrates that LncRNAs have a predictive effect on the occurrence and prognosis of prostate cancer, making them potential new biomarkers for prostate cancer survival and potential targets for treatment.

The work presented in this report uses as a baseline and analysis comparing the results obtained by the TCGA Network research [7]. Their work characterized 333 primary prostate cancers using seven genomics platforms and identified three significant groups of prostate cancers. One with mostly unaltered genomes, a second group comprised 50% of all tumours, exhibited an intermediate level of somatic copy number alterations (SCNAs), and a third group with a high frequency of genomic gains and losses at the level of chromosome arms. With this in mind, the main contribution of the research presented in our report is to analyze how the integration of different biological data sources to create similarity networks combined with clustering methods can approximate the clustering into the three subtypes obtained by the TCGA Network research [7].

The remainder of this work is organized as follows. Section II describes the data integration and clustering approaches exploited while describing the used dataset, considering disease subtypes, and explaining the data-preprocessing techniques applied. Section III explains the methodology used to validate the proposed solution, considering the employed metrics to validate the clustering with the disease subtypes and presents the results obtained from the experimental procedure. Finally, Section IV presents the main conclusions and opportunities for future research.

II. Methodology

This section describes the developed pipeline for prostate adenocarcinoma subtype discovery on multi-omics data through similarity networks and clustering methods. For that, this work takes advantage of three different features creation (single omics similarity matrices, average integrated similarity matrix and similarity networks fusion) and two different clustering methods: PAM [11] and Spectral Clustering [12].

On the one hand, PAM has been widely used, and it is known to be a robust version of k-means as it is considered less sensitive to outliers [11]. On the other hand, Spectral Clustering has provided valuable results when the structure of the clusters is non-convex or when a measure of the centre and spread of the cluster is not a suitable description of the complete cluster [12]. Therefore, this project was designed to evaluate the following Research Questions:

- **RQ1** - Does multi-omics data use outperform the single omics similarity matrix?
- **RQ2** - Considering the integration methods in multi-omics data, can the similarity networks fusion outperform simple average data integration?
- **RQ3** - Does the usage of the Spectral Clustering provide better results when compared to the PAM algorithm while clustering with the similarity networks fusion?
- **RQ4** - Does any of the proposed techniques provide a good estimation of the iCluster disease subtypes?

A. Overview

Figure 1 shows an overview of the proposed prostate adenocarcinoma subtype discovery method. The process starts by fetching the appropriate date from the TCGA dataset [1]. The disease code to fetch for prostate adenocarcinoma is *PRAD*. Specifically, this project fetched data of mRNA, miRNA and proteins, which corresponds to *RNASeq2Gene*, *miRNASeqGene*, and *RPPAArray* assays in the TCGA package. After fetching the desired omics, each is treated in a data preprocessing pipeline, responsible for multiple data cleaning and feature selection tasks. After the data is cleaned, the next step is to create the similarity matrices of each omic using the scaled exponential Euclidean distance. With the similarity matrices, we can perform the data integration of the omics, creating a single integrated matrix representing multi-omics information. We perform the integration with two different methodologies, a baseline method of a simple average and the SNF. Finally, the integrated matrices and the single omics similarity matrices are clustered independently using the PAM clustering algorithm, which enables the comparison e evaluation of each integration technique (see Figure 1 again). As an extra step, the SNE matrix is also clustered with the spectral clustering algorithm, allowing us to compare also the clustering method. The in-depth information on the dataset, data preprocessing phase, data integration methods and clustering approaches are further detailed in this section.

This research was implemented using the R programming language with the help of the BiocManager package manager for the *curatedTCGAData*, *TCGAutils*, *TCGAbiolinks* libraries. Those BiocManager libraries were the main interface of communication and manipulation with the dataset. Additionally, *SNFtool* was used for the SNF and spectral clustering, *cluster* for the PAM algorithm, and *mclustcomp* for clustering metrics.

B. Dataset

The idea behind the PRADC methodology is to work with a multi-omics dataset from prostate cancer patients. As saw in Figure 1, we fetch a prostate cancer multi-omics dataset from The Cancer Genome Atlas (TCGA) program [1]. In particular, we exploit the package "curatedTCGAData" to download the following assays: mRNA, miRNA and proteins. Each assay represents a different aspect of the biological state within the cells. The rationale behind utilizing multiple data sources is that the interactions of diverse molecules influence a biological system. Thus, considering multiple biological data sources simultaneously, we can better understand the underlying processes at work [9].

Fetching the data yielded a data structure object, denominated *MultiAssayExperiment*, designed to store and coordinately analyze multi-omics experiments. Especially the generated MultiAssayExperiment had three experiments:

- **PRAD_RNASeq2Gene-20160128**: with 20501 rows and 550 columns
- **PRAD_miRNASeqGene-20160128**: with 1046 rows and 547 columns
- **PRAD_RPPAArray-20160128**: with 195 rows and 352 columns

Each experiment is also a data structure of its own, denominated SummarizedExperiment. It is essential to highlight that the aforementioned describes the raw data, and after the preprocessing phase and the iCluster disease subtype matching, the specific amount of user data changes, as described in Subsection II-D.

1) Barcodes: It is essential to understand the barcode structure associated with each sample to work with data coming from TCGA. Each sample corresponds to a patient and is identified by a barcode with a specific structure shown in Figure 2.

The initial 12 characters in the barcode serve as identifiers for a particular individual, indicating the Project, TSS (Tissue Source Site), and Participant. These characters uniquely identify a specific person within the dataset. The remaining characters in the barcode provide information about the sample type, such as a primary tumour, metastatic sample, solid tissue, blood-derived sample, and others. They also indicate the type of extracted genomic material, such as DNA or RNA. Additionally, these characters may contain details about technical replicates, which refer to repeated measurements taken from the same sample. Overall, the barcode encodes essential information about the individual, sample type, genomic material, and any technical replicates associated with the data. Understanding the data barcode is crucial to the data preprocessing phase.

C. Data Preprocessing

Figure 3 summarizes the data preprocessing step and is inspired by [9]. The preprocessing starts with the individual omics from the selected assays and is applied independently for each of the three. The first two steps are related to the barcode and are used to filter the data. We start the preprocessing by fetching only primary solid tumours (original tumours). In the TCGA Barcode, primary solid tumours are identified by the code "01" in the sample part of the code (see Figure 2 again). The idea behind using only primary solid tumours is to have a more homogeneous group of samples [9]. Additionally, the barcode enables us to perform a second preprocessing step: checking for technical duplicates. Since repeated measurements taken from the same sample are not attractive to the application, we use the barcode to check for any repeated string in the 12 first characters and remove any possible duplicates.

Further, an additional set of preprocessing activities is performed to standardize, remove unused data and filter the features and samples. We are starting by removing FFPE (Formalin-Fixed Paraffin-Embedded) samples. After performing a biopsy, storing and preserving the sample properly is essential. There are two main methods for preparing the tissue: (1) FFPE (Formalin-Fixed Paraffin-Embedded) and (2) freezing the sample. Freezing the tissue helps to preserve DNA and RNA molecules more effectively [9]. Thus we exclude samples that have been preserved using the FFPE technique. Following, we filter only the samples with all considered omics, disregarding all samples of patients which do not present mRNA, miRNA and protein data. The fetched samples are extracted in a list of three matrices, where each matrix is information on the samples for that specific omics data. The matrices are also transposed to have samples in rows and features in columns.

Now that we have our sample and features as matrices, further preprocessing is more straightforward (at this point, we are at Figure 3, second column). The following step is to remove features having missing values. Since only a few features in the proteomics data have missing values, it is easier and not significantly impacting to only remove instead of performing some process of imputation [9]. Further, we select the features that have higher variance across the samples. This is based on a strong assumption that features with more variance across samples bring more information, thus being the more relevant ones [9]. This feature selection strategy is widely used in literature due to its speed, but it has some limitations [9]. Besides being univariate (not considering interactions among features), it does not effectively remove redundant variables [9]. Additionally, we need to select a threshold for the number of features to select. For the PRADC project, we arbitrarily select 100 features.

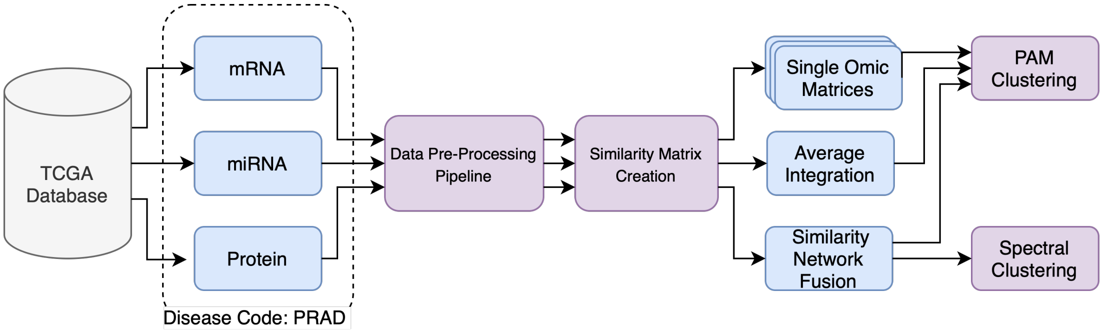


Fig. 1: Overview of PRADC proposed methodology

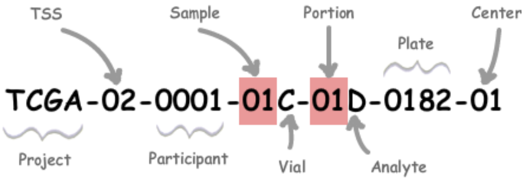


Fig. 2: Sample data barcode (Source: TCGA Barcode [16])

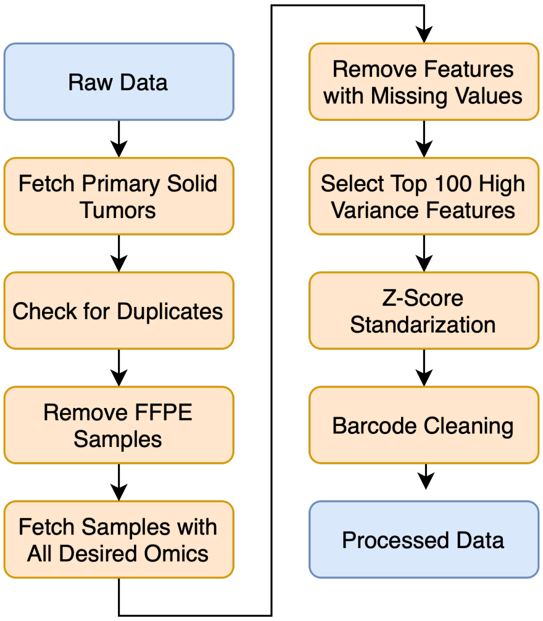

Raw Data
Fetch Primary Solid Tumors
Check for Duplicates
Remove FFPE Samples
Fetch Samples with All Desired Omics
Remove Features with Missing Values
Select Top 100 High Variance Features
Z-Score Standardization
Barcode Cleaning
Processed Data

Fig. 3: Overview of the preprocessing pipeline

The previous to last step of the preprocessing phase is feature standardization. PRADC standardize features using Z-score. The Z-score is a statistical measure representing the number of standard deviations an individual data point is from the mean of a given dataset. This standardization helps understand how far away a data point is from the mean relative to the spread of the data. A positive Z-score indicates that the data point is above the mean, while a negative Z-score indicates that it is below the mean. Finally, the last step of the data preprocessing is cleaning the sample barcodes to retain only the 12-character part specific to each individual. With this modification, we can better identify our data rows and provide a simple method of matching the omics data with the disease subtypes.

D. Prostate Cancer Disease Subtypes

The classification of a cancer sample to a specific disease subtype helps predict patients' prognosis, and it also impacts the definition of the therapy. Defining and identifying subtypes allows grouping similar individuals, which may help the field of "personalized medicine" better predict their prognosis and appropriate therapy [9]. Therefore this research uses the *TCGAbiolinks* package to fetch the subtypes data.

Prostate cancer can be divided into multiple different subtypes classifications. TCGA Network's research [7], revealed a molecular taxonomy in which 74% of the tumours fell into one of seven subtypes defined by specific gene fusions (ERG, ETV1, ETV4, and FLI1) or mutations (SPOP, FOXA1, and IDH1). The *TCGAbiolinks* package considers this classification the most prominent and is used in their data "Subtype_Selected" column. However, the TCGA Network in the same research [7], using the iCluster technique, were able to identify also three significant groups of prostate cancers. One with mostly unaltered genomes, a second group comprised 50% of all tumours, exhibited an intermediate level of somatic copy number alterations (SCNAs), and a third group with a high frequency of genomic gains and losses at the level of chromosome arms. The iCluster subtypes are labelled as the "Subtype_Integrative" column in the data. They are used as the baseline for the PRADC research since its subtype classification is more approachable while still being relevant to patients' predicted prognosis. Additionally, the iCluster subtype is also interesting for our specific application since it also comes from a multi-omics integrative analysis. Therefore this research tries to verify if the computed clusters with the proposed techniques are similar to the iCluster disease subtypes provided by TCGA for prostate cancer.

Analyzing the data is noticed that not all subtypes are present in the subset of samples that contain all the considered omics data sources. Therefore, it was necessary only to include samples from the multi-omics dataset with an associated subtype. With this step, we could filter samples with the disease subtypes in the selected omics, achieving the final shape of the data points. Therefore the final subtypes database consisted of 60 patients of Type-1, 83 patients of Type-2 and 105 patients of Type-3, totalizing 248 select patients for the multi-omics data clustering. Finally, our data shape is made of a vector of 3 matrices (one for each omic), where each matrix contains 248 rows (samples/patients) and 100 columns (high variance selected features).

E. Data Integration

For working with multi-omics data, a suitable method for fusing the diverse omics into a single data source is needed for most machine learning applications. Integrating different omics data is challenging in scientific research, and numerous methods have been proposed to address this issue [9]. Several reviews have been written on these approaches. Specifically, Gliozzo *et al.* [17] provide a comprehensive review of existing methods to integrate multiple biomedical data views and construct patient similarity networks. In this research, the basis of every integration method is the construction of a similarity matrix among samples for each data source, exploiting the scaled exponential Euclidean distance as a similarity measure.

The reasoning for choosing the scaled exponential Euclidean distance is based on its local normalization of the distance between a central node and any of its neighbours so that distances are independent of the neighbourhood scales [17]. The neighbourhood size set to PRADC study for the scaled exponential Euclidean distance was arbitrarily set to 20.

With the similarity matrices created, we fuse our prostate cancer multi-omic dataset with two different strategies. The first is a simple average of the matrices, which can be considered a trivial multi-omics data integration strategy. The second is the state-of-the-art approach SNF [10], implemented in the package *SNFtool*.

The SNF approach integrates multiple data types from the same set of samples. It creates separate networks for each data type and combines them into a single similarity network. The process involves using similarity matrices to represent the pairwise similarities between samples in each data type's network. These matrices can be visualized as networks, where samples are nodes and weighted edges represent pairwise similarities [10]. The network-fusion step iteratively updates each network to make them more similar. SNF eliminates weak similarities (low-weight edges) to reduce noise and fortify strong similarities (high-weight edges) from one or more networks [10]. Therefore, SNF's nonlinear approach allows it to integrate standard and complementary information across networks [10]. SNF has proven successful compared to other relevant fusion methods and outperforms single omics data [10], [17].

F. Clustering Approaches

Finally, with the integrated multi-omics data, we can perform disease subtype discovery using the PAM and the Spectral Clustering algorithms. Since we are aiming to understand the impact of the integrated data method and the clustering algorithm, we performed the following clustering/similarity matrices combinations:

- mRNA single sources PAM clustering
- miRNA single source PAM clustering
- Proteins single source PAM clustering
- Average data integration PAM clustering
- SNF PAM clustering
- SNF Spectral Clustering

The similarity matrices for the first three experiments were obtained from single data sources (i.e., miRNA, mRNA, proteins) using the usual scaled exponential Euclidean distance. Thus, we cluster independently three different similarity matrices. For all PAM clustering methods, we need to provide the input as a distance matrix, and for the single source and average, these distance matrices need to be normalized.

1) Partitioning Around Medoids (PAM): We first attempt to identify disease subtypes using the PAM clustering algorithm [11]. The algorithm finds a fixed number of clusters k (in our case 3 since we are further comparing with the iCluster baseline) represented by their central points called *medoids*. The main objective is to form clusters where the average distances between objects within a cluster and their respective *medoid* are minimized. The PAM algorithm consists of two main phases:

1) **Initialization:** In this phase, the algorithm randomly selects k objects from the dataset as the initial medoids, which act as representatives for the initial clusters [11].
2) **Update:** Iteratively updates the assignment of objects to clusters and selects new medoids to improve the clustering. The algorithm goes through each object and calculates the dissimilarity between the object and each medoid. The object is then assigned to the cluster represented by the medoid with the minimum dissimilarity. After the initial assignment, the algorithm improves the clustering by considering a potential swap between a medoid and a non-medoid object within the same cluster. If the swap reduces the total dissimilarity, it is accepted, and the medoid is updated. This process is repeated until no further improvements can be made or a stop criterion is reached [11].

2) Spectral Clustering: As a bonus experiment, we cluster the SNF using the Spectral Clustering algorithm. This approach leverages the concept of spectral decomposition to transform the data into a lower-dimensional space where the clusters can be more easily identified [12]. The spectral clustering algorithm allows for detecting clusters that may have complex shapes or are not linearly separable in the original feature space [12]. Therefore, it often outperforms traditional clustering algorithms such as the k-means and PAM algorithm [12]. In our work, spectral clustering is performed with the support of the *SNFtool* library, using $k = 3$ for the number of clusters.

III. Results

This section presents the methodology adopted and the results obtained to draw insights from the proposed solution. For this, the experimentation was made by designing a few

TABLE I: Count of samples in clusters for every experiment

Experiments	Cluster1	Cluster2	Cluster3
mRNA	76	71	101
miRNA	91	94	63
Proteins	45	120	83
Avg Integration	79	97	72
SNF Integration	78	92	78
Spectral Clustering	83	84	81
iCluster	**105**	**60**	**83**

TABLE II: Summary of the experiments results metrics

Experiments	ARI	Jaccard	NMI	RI
mRNA	0.0620	0.2375	0.0580	0.5771
miRNA	0.0270	0.2205	0.0425	0.5610
Proteins	0.0003	0.2205	0.0144	0.5378
AVG Integration	0.0418	0.2263	0.0704	0.5689
SNF Integration	**0.1794**	**0.2972**	**0.1567**	**0.6316**
Spectral Clustering	0.1190	0.2638	0.1172	0.6051

experiments to establish links between the methodologies developed for data integration and clustering and the proposed Research Questions (RQs) in Section II.

To answer these questions, a set of experiments has been designed. Especially six experiments were performed. The first three experiments used the PAM clustering with the similarity matrices obtained from single data sources (miRNA, mRNA, and proteins) using the usual scaled exponential Euclidean distance. Those are referred to as single omics experiments since each clustering is performed on an independent data source, and there is no data integration technique. The other three experiments use multi-omics data, referred to as multi-omics experiments. Those experiments include the PAM clustering with the average data integration technique, the PAM clustering with the SNF integration, and the spectral clustering with the SNF integration. This set of experiments should be sufficient to cover the RQs proposed. Table I shows an initial overview of the results. The idea of this Table is to present the count of samples in each cluster and visualize the distributions of samples per cluster in each experiment, also serving as a "sanity-check" for the clustering process.

However, we must first define cluster-comparing metrics to evaluate the proposed solutions and compare the clusters obtained by each considered approach for the iCluster disease subtypes. Many measures to compare clusters are available in the literature [18]. Those metrics are already implemented in the *mclustcomp* R package, which we used in our research. From the 24 different scores available in the package, we hand-picked four commonly used measures to compare clusters:

- **Rand Index (RI):** Measure that quantifies the agreement between two sets of clusters by "counting pairs" [9]. It calculates the similarity by comparing pairs of data points and determining if they are assigned to the same or different clusters in both sets. The Rand Index ranges from 0 to 1, where 0 indicates no agreement, and 1 indicates perfect agreement [18].
- **Adjusted Rand Index (ARI):** A variation of the Rand Index that considers agreement by chance [18]. The ARI considers that some agreement between two clusters can occur by chance and provides a value ranging from -1 to 1. A value of 1 indicates perfect agreement, 0 suggests agreement by chance, and negative values indicate worse agreement than expected by chance.
- **Normalized Mutual Information (NMI):** Measure derived from Mutual Information, which measures the amount of information obtained about one set when the other set is known [18]. The NMI adjusts the Mutual Information by dividing it by the average entropy of the two sets, thus normalizing the value between 0 and 1. A higher NMI value indicates a more significant similarity, while a lower value suggests less similarity. The NMI provides a normalized measure that accounts for the inherent differences in the sizes and entropies of the compared sets.
- **Jaccard Index:** Measure calculated by dividing the size of the intersection of the sets by the size of the union of the sets [18]. The resulting value ranges from 0 to 1, where 0 represents no similarity, and 1 represents complete similarity. Although it's easy to interpret, it is sensitive to small sample sizes.

Table II summarizes the resulting metrics for each experiment. The Table shows that all the clustering results have overlapped, although they are minimal. Starting from the Table analysis to answer the *RQ1*, we are interested in discovering if the multi-omics data outperform the usage of single omics-data. First, notice that all metrics of the SNF Integration and Spectral clustering outperform the metrics of the single omics data, which shows a strong indicator that the multi-omics data provides better results. However, this analysis is not valid for the average integration technique, which outperforms the protein and miRNA single-omics but not the mRNA. Except for the NMI, the mRNA metrics outperform the average integration metrics, which could indicate that mRNA has higher clustering power concerning the miRNA and proteins. It is also a possible indicator that simple data integration techniques don't add value if a single omic's information is more relevant than others. This hypothesis could be validated with a weighted average with a higher weight for the mRNA omics.

Considering the *RQ2*, we focus on whether the similarity networks fusion integration method can outperform a simple average data integration. From Table II, we notice that the answer is pretty straightforward since the SNF integration outperformed the average integration in all metrics in both clustering approaches. Not only, SNF integration consistently outperforms single-omics methods as well. For the *RQ3*, we are interested in comparing the PAM algorithm against the Spectral Clustering algorithm, both with SNF data integration. It is also clear from the Table that the PAM clustering outperforms the Spectral clustering since all its metrics are higher, providing a better similarity with the iCluster diseases subtypes.

With all the previous RQs answered, we can discuss RQ4 on whether any of the proposed techniques reasonably estimate the iCluster disease subtypes. Since we agreed that the SNF Integration with PAM clustering provided the best results in the previous discussion, we will focus on analyzing its metrics.

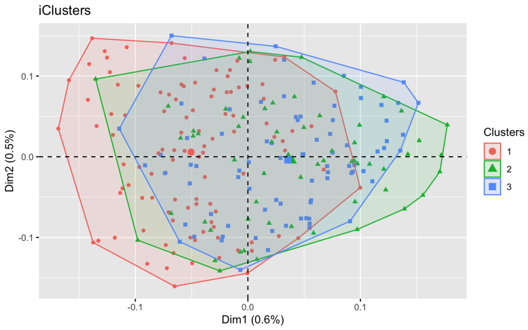
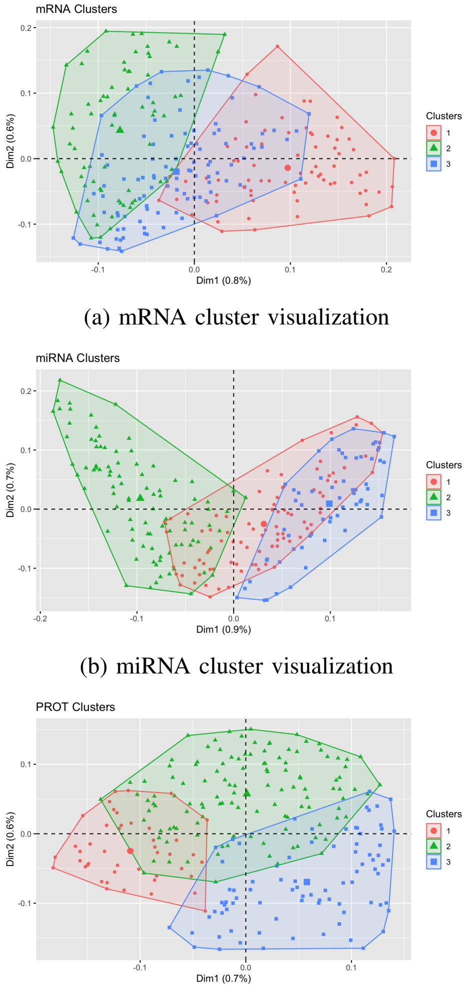


Fig. 4: iCluster visualization with the PCA of SNF data



(a) mRNA cluster visualization

(b) miRNA cluster visualization

(c) Protein cluster visualization

Fig. 5: Single-omics data clustering visualization

As stated, the Table shows that the clustering results have overlapped, although very limited. The most optimistic metric is the RI, with a value of 0.6316, which indicates that the clusters are slightly more alike than they are unlike. However, when we consider the agreement by chance introduced by the ARI of 0.1794, it's clear that the agreement is closer to a result by chance than to a perfect agreement. This can also be confirmed by the 0.1567 NMI and 0.2972 Jaccard similarity, which both indicate a small similarity value between the sets.

The limited similarity between the iCluster disease subtypes and our clustering experiments is unsurprising. Even though both approaches derive from multi-omics data, the PAM clustering with SNF Integrated data is a more straightforward methodology than the one employed for the disease subtypes. The iCluster technique incorporates flexible modelling of the associations between different data types and the variance-covariance structure within data types in a single framework [8]. At the same time, PAM clustering is a traditional and simpler technique. Additionally, it is possible that our simple data processing has not been sufficient to extract relevant information for the clustering algorithms, thus making it more challenging to identify the subtypes based on the extracted features.

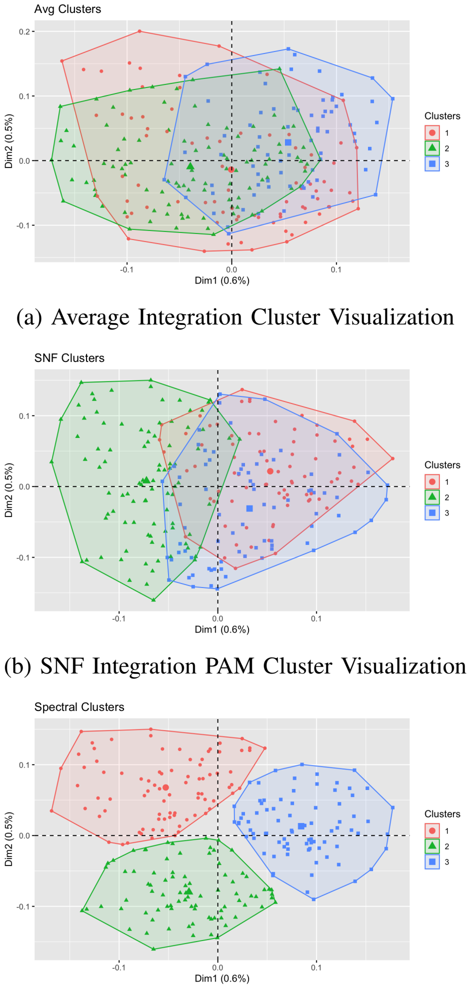
In addition to the metrics result, we provide a plot visualization of the obtained clusters. To be able to visualize, we performed a dimensionality reduction using Principal Component Analysis (PCA). PCA transforms high-dimensional data into a lower-dimensional representation while retaining the most critical information [19]. First, we generated a clustering for the iCluster results using the data from the SNF integration since it was the most prominent method. Figure 4 shows the obtained clusters, and the overlap between the clusters in two dimensions is clear. The overlap could indicate that considering only the two features with the higher variance is insufficient to separate the clusters.

Further, we perform the same process to visualize the results. Figure 5a, 5b and 5c represent the cluster results obtained for single-omics and Figure 6a, 6b and 6c represent the cluster results obtained for multi-omics data. It is worth mentioning that the positioning of the data is not comparable to the iCluster in the single omics and the average integration since the iCluster PCA was performed with SNF-integrated data. However, for the single-omics data, we can see that the clusters are more well separated in two dimensions when compared to the iCluster result, which could indicate that the PAM clustering depends on fewer features to find the clusters.

For the multi-omics data clusters, we see a more substantial overlap in the clusters, except for the Spectral clustering. This result can reflect the spectral clustering using spectral decomposition to transform the data into a lower-dimensional space where the clusters can be more easily identified [12]. Therefore, the clusters are easily separated when we reduce the dimensionality with the PCA.

IV. Conclusion and Future Work

This report proposed multiple approaches for integrating omics data and clustering information for prostate adenocarcinoma subtype discovery. Specifically, we fetched mRNA, miRNA and proteins omics and integrated their similarity matrixed with a simple average and the SNF. We also clustered the single- and multi-omics data with the PAM and Spectral clustering algorithms while comparing their results to the disease subtypes discovered in the iCluster in the TCGA Network research [7]. Results indicated a more significant clustering power for the multi-omics data integrated with SNF, while the simple integration method can decrease the power



(a) Average Integration Cluster Visualization

(b) SNF Integration PAM Cluster Visualization

(c) SNF Integration Spectral Cluster Visualization

Fig. 6: Multi-omics data clustering visualization

of a single-omics source. Additionally, the results favoured the PAM algorithm compared to spectral clustering. Therefore our more robust cluster was generated with the SNF integration with PAM clustering, in which the metrics demonstrated overlap with the iClusting disease subtypes with an RI of 0.6316. However, the overlap was limited since we obtained an ARI of 0.1794, showing that the overlap is closer to a result by chance than a perfect agreement. Future works could explore the impact of weighted averages for data integration as a better baseline method. Additionally, further improvements in data processing can be performed, such as using auto-encoders for dimensionality reduction and feature selection.

References

[1] T. Network. (2023) The cancer genome atlas program (tcga). [Online]. Available: https://www.cancer.gov/ccg/research/genome-sequencing/tcga

[2] W. H. O. (WHO) *et al.*, "Cancer incidence and mortality worldwide: Iarc," 2015.

[3] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: a cancer journal for clinicians*, vol. 73, no. 1, pp. 17–48, 2023.

[4] A. A. Al Olama, Z. Kote-Jarai, S. I. Berndt, D. V. Conti, F. Schumacher, Y. Han, S. Benlloch, D. J. Hazelett, Z. Wang, E. Saunders *et al.*, "A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer," *Nature genetics*, vol. 46, no. 10, pp. 1103–1109, 2014.

[5] P. K. Singh and M. J. Campbell, "The interactions of microrna and epigenetic modifications in prostate cancer," *Cancers*, vol. 5, no. 3, pp. 998–1019, 2013.

[6] L. Boldrini, P. Faviana, L. Galli, F. Paolieri, P. A. Erba, and M. Bardi, "Multi-dimensional scaling analysis of key regulatory genes in prostate cancer using the tcga database," *Genes*, vol. 12, no. 9, p. 1350, 2021.

[7] A. Abeshouse, J. Ahn, R. Akbani, A. Ally, S. Amin, C. D. Andry, M. Annala, A. Aprikian, J. Armenia, A. Arora *et al.*, "The molecular taxonomy of primary prostate cancer," *Cell*, vol. 163, no. 4, pp. 1011–1025, 2015.

[8] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.

[9] J. Gliozzo. (2023) Practical workshop of bioinformatics course (department of computer science, university of milan). [Online]. Available: https://github.com/GliozzoJ/Bioinformatics_practice2023

[10] 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.

[11] L. Kaufman, "Partitioning around medoids (program pam)," *Finding groups in data*, vol. 344, pp. 68–125, 1990.

[12] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[13] D. Robinson, E. M. Van Allen, Y.-M. Wu, N. Schultz, R. J. Lonigro, J.-M. Mosquera, B. Montgomery, M.-E. Taplin, C. C. Pritchard, G. Attard *et al.*, "Integrative clinical genomics of advanced prostate cancer," *Cell*, vol. 161, no. 5, pp. 1215–1228, 2015.

[14] M. D. Long, D. J. Smiraglia, and M. J. Campbell, "The genomic impact of dna cpg methylation on gene expression; relationships in prostate cancer," *Biomolecules*, vol. 7, no. 1, p. 15, 2017.

[15] H. Huang, Y. Tang, X. Ye, W. Chen, H. Xie, and S. Chen, "The influence of lncrnas on the prognosis of prostate cancer based on tcga database," *Translational Andrology and Urology*, vol. 10, no. 3, p. 1302, 2021.

[16] T. Network. (2023) Tcga barcode. [Online]. Available: https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/

[17] J. Gliozzo, M. Mesiti, M. Notaro, A. Petrini, A. Patak, A. Puertas-Gallardo, A. Paccanaro, G. Valentini, and E. Casiraghi, "Heterogeneous data integration methods for patient similarity networks," *Briefings in Bioinformatics*, vol. 23, no. 4, p. bbac207, 2022.

[18] S. Wagner and D. Wagner, *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

[19] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical methods*, vol. 6, no. 9, pp. 2812–2831, 2014.