Disease subtype discovery using multi-omics data integration

Elia Togni[1]

[1]Dipartimento di Informatica, Università degli Studi di Milano.

Abstract

Prostate adenocarcinoma, a prevalent malignancy among men worldwide, exhibits substantial heterogeneity in its clinical presentation and molecular characteristics. The work presented in this report 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 as a baseline.

After a preprocessing phase, the work integrates the multi-omics data using firstly an average baseline integration and secondly using a graph-based integration algorithm denominated **Similarity Network Fusion**.

For the clustering aspect, we choose to cluster the patients in the similarity networks using both the **Partitioning Around Medoids** and **Spectral Clustering** methods. Once the clustering process is completed, we are able to evaluate the computed outcomes in accordance with the identified disease subtypes for prostate cancer using *iCluster*.

This study demonstrated a heightened capability for clustering within the multi-omics data when integrated with SNF. Conversely, employing the simple integration approach may diminish the potential of a singular omics source. Moreover, the findings showed a preference for the PAM algorithm over Spectral Clustering. Consequently, the integration of SNF coupled with PAM clustering yielded a cluster more similar to the *iCluster* ones in our study.

I Introduction

The problem of disease subtype discovery from multi-omics data

Recents advancements in technology have facilitated the generation of diverse genome-wide high-throughput biological data types, collectively referred to as **omics** [1]. Omic is a suffix used to refer to different fields of study that involve comprehensive analysis of a specific biological component or aspect. The term **multi-omic** typically denotes a multidimensional approach to studying biological systems on a large scale, encompassing various molecular components, such as genes (**genomics**), proteins (**proteomics**), metabolites (**metabolomics**), and more [2]. The omic sciences aim to understand the complex interactions and functions of these components to gain insights into biological processes. By utilizing the omic approach, researchers seek a comprehensive grasp of biological systems at a molecular level, exploring the intricate networks and relationships that contribute to an organism's structure, function and behavior, trying to reveal more holistic, systems-level insights [1].

The wealth of these omic profiles gathered from large cohorts in recent years presents a unique opportunity to also gain a deeper understanding of human diseases. These profiles can serve as valuable resources for characterizing diseases more comprehensively, thus facilitating the development of personalized treatment strategies tailored to individual patients [1]. This is an integral part of what is known as **personalized medicine** [2].

In the field of oncology, the analysis of extensive datasets has led to the identification of novel cancer subtypes, revolutionizing treatment decision-making [1]. However, typically, the attained results are based on the analysis of a single omic rather than being derived from a comprehensive analysis of multiple data sources. Since the molecular complexity of a tumor manifests itself at the omics levels, genomic profiling at these multiple strata allows a better integrated characterization of tumor etiology [3].
Identifying tumor subtypes by simultaneously analyzing **multi-omic data** is, therefore, a relatively new problem. In fact, since initiatives like **The Cancer Genome Atlas** [4] have made multi-omic cohort data largely available, there has been a pressing need for improved and advanced methodologies that enable the integrated analysis of these datasets.

The Cancer Genome Atlas Program [4] (henceforth referred to as **TCGA**) is a pioneering cancer genomics program that originated in the United States, bringing together researchers from diverse disciplines. Over the years, TCGA has produced more than 2.5 petabytes of publicly available genomic, epigenomic, transcriptomic, and proteomic data [4]. It encompasses data from more than $11,000$ cases across 33 tumor types, generating an extensive and comprehensive dataset that describes the molecular changes occurring in cancer. This project is the coordinated effort of the **National Cancer Institute** (**NCI**) and of the **National Human Genome Research Institute** (**NHGRI**) to explore the entire spectrum of genomic alternations in human cancer to obtain an integrated view of the molecolar features of cancers [3, 5]. Therefore, the value of the TCGA dataset cannot be overstated. Its vastness has, in fact, allowed researchers to document specific genomic and molecular changes in cancer, establish a more meaningful taxonomy of cancer types and subtypes and even explore questions that were not initially envisioned during the project's planning [5].

The simplest way utilized in the past to combine biological data was to concatenate normalized measurements from various biological domains for each sample. Regrettably, the act of combining different types of data makes it more challenging to discern meaningful information from the background noise within each type of data. This is likely due to the inherent differences in the data formats and characteristics of the various data types being merged, which may create more complexity and noise when combined together. Concatenation further dilutes the already low signal-to-noise ratio in each data type [6]. To avoid this, a common strategy was to also analyze each data type independently before combining data. In fact, the most used approach to subtype discovery across multiple types in the past years was to separately cluster each type and, then, to manually integrate the result. Researchers often resorted to heuristic approaches where manual integration was performed after separate analysis of individual data types, and it was unlikely that two investigators would perform manual integration in the same manner. Manual integration also may require a considerable amount of prior knowledge about the underlying disease [3]. Therefore, such independent analyses often led to inconsistent conclusions that were hard to integrate [6].

Multi-omics clustering methods

The field was now confronted with the need to develop innovative computational methodologies that could seamlessly merge data from genomics, transcriptomics, proteomics, and beyond.

A pivotal journey in the evolution of bioinformatics has been the transition from the analysis of single omic data to the integration of multiple omic dimensions.

Within the data integration flow, two broad classifications (Figure 1) have been established by existing literature: the first classification refers to the **horizontal integration approaches** and the second one refers to the **vertical integration approaches**. Horizontal integration approaches fuse **multisets** (e.g. datasets where each view is acquired by the same source under different conditions) by independently applying the same procedure on each view and then merging the individual results. On the other hand, vertical integration approaches fuse **multimodal datasets** (e.g. datasets composed by semantically different views) through more complex techniques, further categorized as **hierarchical–vertical integration** methods and **parallel–vertical integration** techniques. The former fuse data views following a **hierarchy** driven by biological a priori knowledge whereas the latter do not exploit knowledge-based dependencies between views. **Parallel-vertical integration** methods are the most diffused integration methods; they are further classified based on the phase when the data **integration-step** is performed with respect to the model construction [7]. The simplest parallel-vertical integration approach, called **early integration** (also named **concatenation-based method**), is applied on the input data in an early stage and it concatenates all omic matrices into one and applies single-omic clustering on it. The evident advantage of early methods relies on their ability to uncover the individual information characterizing each of the different sources as well as the hidden relationships between them. Another considerable advantage is brought by the fact that early methods solve the integration problem in the first stage, so that any unimodal analysis (statistical analysis of data that involves a single variable) process may be subsequentially applied. Nevertheless, these methods suffer from the increasing of the dimensionality of the data. They also ignore the different distributions of values in different omics.

Another approach, called **late integration** method (also named **model-based method**), clusters each omic separately, and then integrates in a late phase the clustering results. Its strength relies on its capacity to use readily available tools designed specifically for each omics type, and compared to the other strategies, it does not suffer the challenges of trying to assemble different kinds of data. This approach has, however, the flaw of ignoring interactions that are weak but consistent across omics, discarding in this way an important piece of information [1, 8]. These approaches along with the early integration ones are classified as **model-agnostic**, because they are independent from the specific algorithm applied in the preceding unimodal analysis, which can be therefore tailored to the processed type [7].

Finally, an ulterior integrative clustering approach, which accounts for all omics, is the one called **middle integration**. It allows joint inference from multi-omic data and generates a single integrated cluster assignment through simultaneously capturing patterns of genomic alterations that are consistent across multiple data types, specific to individual data types or weak yet consistent across datasets that would emerge only as a result of combining levels of evidence.

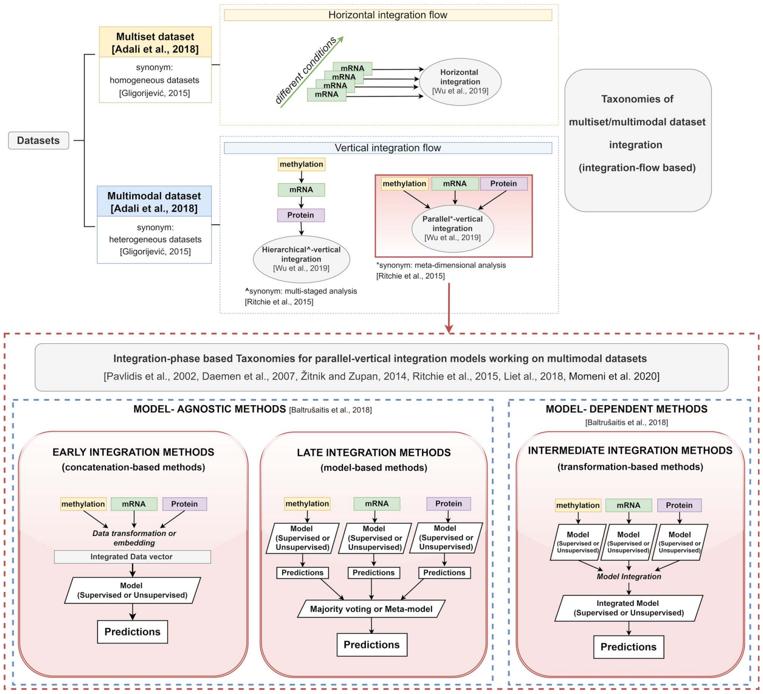

Horizontal integration flow
Multiset dataset [Adali et al., 2018]
synonym: homogeneous datasets [Gligorijević, 2015]
different conditions
mRNA
mRNA
mRNA
mRNA
Horizontal integration [Wu et al., 2019]
Taxonomies of multiset/multimodal dataset integration (integration-flow based)
Datasets
Vertical integration flow
Multimodal dataset [Adali et al., 2018]
synonym: heterogeneous datasets [Gligorijević, 2015]
methylation
mRNA
Protein
Hierarchical^-vertical integration [Wu et al., 2019]
^synonym: multi-staged analysis [Ritchie et al., 2015]
methylation
mRNA
Protein
Parallel*-vertical integration [Wu et al., 2019]
*synonym: meta-dimensional analysis [Ritchie et al., 2015]
Integration-phase based Taxonomies for parallel-vertical integration models working on multimodal datasets
[Pavlidis et al., 2002, Daemen et al., 2007, Žitnik and Zupan, 2014, Ritchie et al., 2015, Liet al., 2018, Momeni et al. 2020]
MODEL- AGNOSTIC METHODS [Baltrušaitis et al., 2018]
MODEL- DEPENDENT METHODS [Baltrušaitis et al., 2018]
EARLY INTEGRATION METHODS (concatenation-based methods)
methylation
mRNA
Protein
Data transformation or embedding
Integrated Data vector
Model (Supervised or Unsupervised)
Predictions
LATE INTEGRATION METHODS (model-based methods)
methylation
mRNA
Protein
Model (Supervised or Unsupervised)
Model (Supervised or Unsupervised)
Model (Supervised or Unsupervised)
Predictions
Predictions
Predictions
Majority voting or Meta-model
Predictions
INTERMEDIATE INTEGRATION METHODS (transformation-based methods)
methylation
mRNA
Protein
Model (Supervised or Unsupervised)
Model (Supervised or Unsupervised)
Model (Supervised or Unsupervised)
Model Integration
Integrated Model (Supervised or Unsupervised)
Predictions

Figure 1: Schema of the main taxonomies proposed in literature for categorizing multimodal integration methods [7, 9, 10, 11, 12, 13, 14, 15].

However, this data-integration method needs to overcome at least three **computational challenges** [6]:

- the small number of samples compared to the large number of measurements;
- the differences in scale, collection bias and noise in each data set;
- the complementary nature of the information provided by different types of data.

In addition to the computational ones, it is possible to underline two other major challenges to the development of a truly integrative approach. First, to capture both concordant and unique alterations across data types, separate modeling of the covariance between data types and the variance–covariance structure within data types is needed. Most of the existing deterministic clustering methods cannot be easily adapted in this way [3].
Second, **dimension reduction** plays a crucial role in making integrative clustering approaches feasible and efficient. Utilizing pairwise correlation matrices for such methods becomes computationally impractical, especially with modern high-resolution arrays. To address this, dimension reduction techniques like **Principal Component Analysis** (**PCA**) [16] and **Non-Negative Matrix Factorization** (**NMF**) have been proposed to work alongside clustering algorithms effectively. While

these techniques are successful for individual data types, they are limited in their ability to simultaneously reduce dimensions for multiple correlated datasets [3].
Therefore, because of the high number of features and because of the complexity of dimension reduction algorithms, feature selection is required [1].

Conversely, similarity based methods address these limitations by utilizing **inter-patient similarities**. Such approaches exhibit enhanced computational efficiency and reduced dependence on selecting features.
All middle integration methods for multi-omics clustering developed within the bioinformatics community assume full datasets, e.g. data from all omics were measured for each patient. Nonetheless, in practical experimental setups, it is a frequent scenario that, for certain patients, only a subset of the omics are actually measured. These datasets are called **partial datasets** [1]. This occurrence is already observable in existing multi-omic datasets and is projected to intensify as study cohorts expand. Being able to analyze partial data holds immense significance due to the substantial costs associated with experiments and the uneven expenses associated with obtaining data for diverse omics. Naive solutions like using only those patients with all omics measured or **imputation** (the assignment of a value to something by inference from the value of the products or processes to which it contributes) have obvious disadvantage [3].

Prostate adenocarcinoma

The case of study of this report is the **prostate cancer**, a cancer type that affects the prostate gland and which is the second most common cancer type among men and, in general, ranking fourth in frequency worldwide [4]. A combination of genetic and demographic factors like age, family history, genetic susceptibilityt and race contribute to its high incidence [4].

The clinical behavior of localized prostate cancer can be extremely heterogeneous, with some individuals having aggressive cancer that can spread and cause death, while others have indolent cancer that can be treated or observed safely [4]. Several genetic and epigenetic alterations are highly prevalent and appear to be essential factors in the tumorigenesis and progression of cancer.

To better predict the likelihood of progression and tailor treatment accordingly, **risk stratification systems** that take into account various clinical and pathological parameters have been developed [4]. **Risk stratification** is the process of categorizing individuals or entities into different **risk levels** based on certain characteristics or factors in order to predict the likelihood of an event or outcome occurring and, therefore, risk stratification systems are tools employed to assign individuals or entities to specific risk categories. These systems aim to identify individuals at higher risk for aggressive disease and guide treatment decisions, taking into account factors such as **prostate-specific antigen** (**PSA**) levels, **Gleason score** (a measure of cancer aggressiveness based on biopsy samples), clinical stage, and other factors [4].

Despite these systems' usefulness, it is fundamental to keep in mind that they are not perfect, and there is still a need for improved risk stratification. This is where molecular features come into play. Molecular and genetic profiles are, in fact, increasingly being used to subtype various cancer types and guide targeted treatment interventions [4].

Recent research has identified several genomic alterations as key features of primary prostate cancer, including **mutations** (changes in the DNA sequence, where one or more nucleotides are altered), **DNA copy-number changes** (changes in the number of copies of a specific DNA sequence or gene in a cell's genome), **rearrangements** (changes in the structure or arrangement of larger segments of DNA, such as genes or whole chromosomes), and **gene fusions** (the joining or fusion of two separate genes, resulting in the formation of a hybrid gene). The most common genomic alteration in prostate cancer is the fusion of **androgen-regulated promoters**, regions of DNA that control the expression of genes in response to androgen hormones, such as testosterone, with members of the **ETS family** of transcription factors such as ERG [4]. The ETS family is a group of genes that encode proteins involved in regulating gene expression. These transcription factors control the activity of various genes, influencing important cellular processes like growth, differentiation, and development.

However, individuals with fusion-bearing tumors do not appear to have a different prognosis following treatment than those without [4, 17].

Prostate cancers also have varying degrees of DNA copy-number alteration, with indolent and low-Gleason tumors having fewer alterations, while more aggressive tumors have a higher burden of copy-number alteration throughout the genome [4, 18].

Further research on the molecular basis of prostate cancer and risk stratification could help identify those at higher risk of developing aggressive disease, leading to better treatment options and outcomes for patients. Therefore, there is a need to continue studying the molecular characteristics of prostate cancer to develop better risk stratification and treatment strategies.

Synopsis

With the aforementioned in mind, the work presented in this report 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 [19] as a baseline. Their study involved the analysis of 333 primary prostate cancers through seven genomics platforms, leading to the recognition of three distinct categories of prostate cancers [4]. The first group displayed predominantly unaltered genomes, the second group, constituting half of all tumors, showed an intermediate level of **somatic copy number alterations** (**SCNAs**), and the third group exhibited a heightened frequency of genomic gains and losses affecting chromosome arms.

The main contribution of the research presented in our report is therefore 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.

Hence, the present study permit to formulate different Research Questions.

The first Research Question delves into the performance of multi-omics data utilization in comparison to the single omics similarity matrix. It seeks to discern whether the application of multi-omics data yields enhanced outcomes.

The second Research Question revolves around the evaluation of integration techniques within multi-omics data. Specifically, it investigates whether the Similarity Network Fusion method outperforms the simple averaging technique when integrating diverse data types.

The third Research Question refers to the choice of clustering algorithms in the context of Similarity Network Fusion. By comparing Spectral Clustering and the PAM algorithm, it aims to identify the superior clustering approach that complements the chosen integration technique.

Lastly, the fourth Research Question aims to assess the efficacy of the proposed techniques in approximating the *iCluster* disease subtypes. This inquiry seeks to determine if any of the suggested methods yield accurate estimations of these subtypes.

The work integrates the multi-omics data using firstly an average baseline integration and secondly using a graph-based integration algorithm denominated **Similarity Network Fusion** [6]. For the clustering aspect, we choose to cluster the patients in the similarity networks using both the **Partitioning Around Medoids** [20] and **Spectral Clustering** [21] methods. Once the clustering process is completed, we are able to evaluate the computed outcomes in accordance with the identified disease subtypes for prostate cancer using *iCluster*.

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. Finally, Section IV presents the main results obtained from the experimental procedure and presents the following conclusions.

II Methodology

1 Overview

This section describes an overview of the proposed procedure (Figure 2) for prostate adenocarcinoma subtype discovery on multi-omics data through similarity networks and clustering methods.

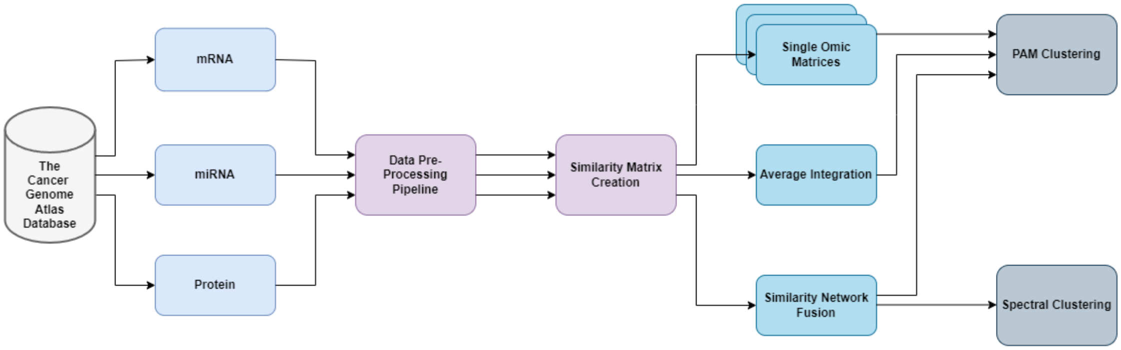

The Cancer Genome Atlas Database
mRNA
miRNA
Protein
Data Pre-Processing Pipeline
Similarity Matrix Creation
Single Omic Matrices
Average Integration
Similarity Network Fusion
PAM Clustering
Spectral Clustering

Figure 2: Overview of the proposed methodology

The process initiates by extracting the pertinent data from the TCGA dataset [4]. The disease

code to fetch for prostate adenocarcinoma is PRAD. Specifically, this endeavor acquires data related to **mRNA**, **miRNA** and **proteins**, corresponding to *RNASeq2Gene*, *miRNASeqGene*, and *RPPAArray* assays within the TCGA toolkit.
Upon obtaining the desired omics data, each type undergoes treatment in a data preprocessing pipeline 5. This pipeline encompasses various tasks, such as data cleansing and feature selection. Post-cleansing, the subsequent steps involve constructing the **similarity matrices** for each omic, utilizing 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. Consequently, two distinct approaches are employed for integration: a simple average method and the Similarity Network Fusion method [6].
Finally, both the integrated matrices and the individual omics' similarity matrices undergo independent clustering using the Partitioning Around Medoids clustering algorithm [20]. This facilitates a comparative evaluation of each integration technique (as depicted in Figure 2). Additionally, the SNF matrix undergoes clustering using the Spectral Clustering algorithm [21], enabling a comparative assessment of clustering methodologies as well.

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 formed the primary interface for dataset communication and manipulation. Furthermore, *SNFtool* was employed for the SNF and Spectral Clustering, *cluster* for the PAM algorithm, *mclustcomp* for clustering metrics, *factoextra* for dimensionality reduction and cluster representation, *NetPreProc* for computing the distance matrices and *caret* for data preprocessing.

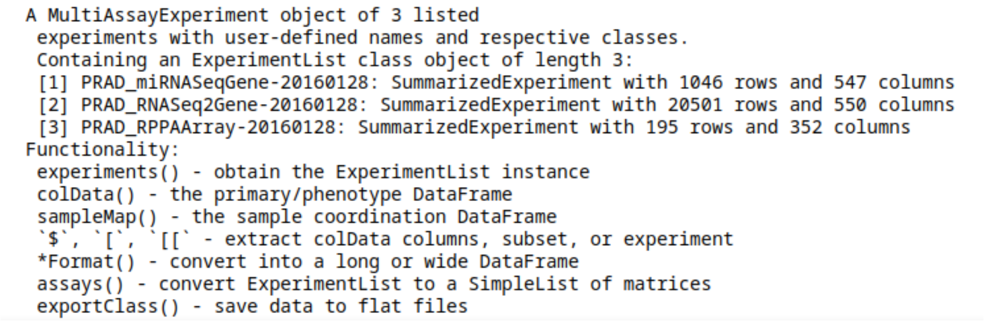
2 Multi-omic Dataset

As above-mentioned and saw in Figure 2, we will download **multi-omics** data from patients having prostate cancer. All omics data are high-dimensional and characterized by **small-n large-p** (e.g. few samples and a large number of features), which easily leads to the **curse of dimensionality** [2] in machine learning applications, which can be defined as the deterioration of algorithm performance caused by the exponential growth of data volume as the number of input features or dimensions increases. In fact, as the dimensionality of the data grows, the available data becomes increasingly sparse in the high-dimensional space, resulting in difficulties in accurately representing and analyzing the data.

In particular, we exploit the package *curatedTCGAData* to download the following data views from TCGA:

- **mRNA data**;
- **miRNA data**;
- **protein data**.

As we can see, we obtain a **MultiAssayExperiment** object, which, in its essence, is a data structure designed to store and coordinately analyze multi-omics experiments. The three main components of this data structure are [2]:



```
A MultiAssayExperiment object of 3 listed
 experiments with user-defined names and respective classes.
 Containing an ExperimentList class object of length 3:
 [1] PRAD_miRNASeqGene-20160128: SummarizedExperiment with 1046 rows and 547 columns
 [2] PRAD_RNASeq2Gene-20160128: SummarizedExperiment with 20501 rows and 550 columns
 [3] PRAD_RPPAArray-20160128: SummarizedExperiment with 195 rows and 352 columns
Functionality:
 experiments() - obtain the ExperimentList instance
 colData() - the primary/phenotype DataFrame
 sampleMap() - the sample coordination DataFrame
 `$`, `[`, `[[` - extract colData columns, subset, or experiment
 *Format() - convert into a long or wide DataFrame
 assays() - convert ExperimentList to a SimpleList of matrices
 exportClass() - save data to flat files
```

Figure 3: A MultiAssayExperiment object

- **colData**: it contains a dataframe having for each sample the corresponding phenotipic characteristics (in our case mainly clinical data);
- **ExperimentList**: a list with the considered experiments (e.g. data modalities acquired with a specific technology). Element of the list are usually matrices or dataframes;
- **sampleMap**: it is a map that connects all the considered elements.

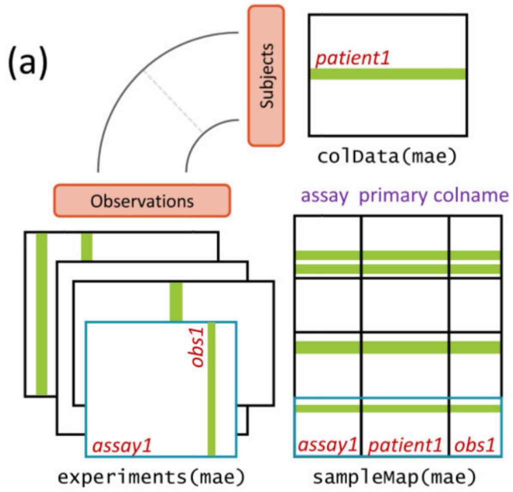


Figure 4: MultiAssayExperiment's structure [2, 22]

Especially the generated MultiAssayExperiment had three experiments:

- **PRAD_RNASeq2Gene-20160128**, comprised of 20501 rows and 550 columns;
- **PRAD_miRNASeqGene-20160128**, comprised of 1046 rows and 547 columns;
- **PRAD_RPPAArray-20160128**, comprised of 195 rows and 352 columns.

Each experiment is also a data structure of its own, denominated **SummarizedExperiment**.

3 Data Preprocessing

This section describes the developed pipeline (Figure 5) for prostate adenocarcinoma subtype discovery on multi-omics data through similarity networks and clustering methods, inspired by [2].

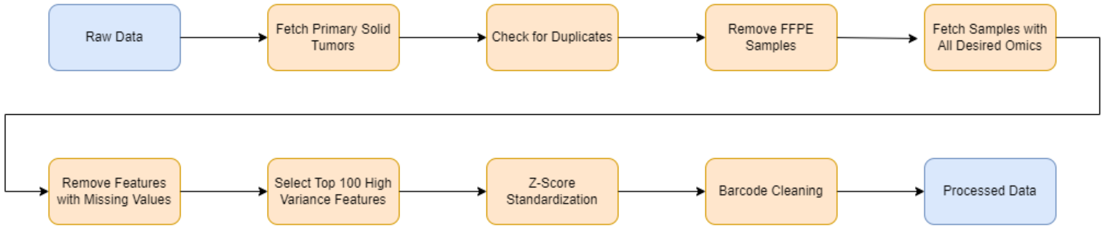

Raw Data
Fetch Primary Solid Tumors
Check for Duplicates
Remove FFPE Samples
Fetch Samples with All Desired Omics
Remove Features with Missing Values
Select Top 100 High Variance Features
Z-Score Standardization
Barcode Cleaning
Processed Data

Figure 5: Schema of the data preprocessing pipeline

Since data comes from TCGA, it is important to understand the structure of the **barcode** associated to each sample. A TCGA barcode is composed of a collection of identifiers. Each sample/patient is identified by one of these codes with a specific structure: in pratice, the first 12 characters of the barcode correspond to a specific individual, while the other parts give us indications about the type of sample (e.g. primary, metastatic, solid, blood derived, etc), the type of genomic material extracted (e.g. DNA, RNA) and other information related to technical replicates (e.g. repeated measurements from the same sample). Each specifically identifies a TCGA data element.
We use the barcode to retain only **Primary Solid Tumors** to have a more homogeneous group of samples and to check for the presence of technical replicates in the dataset. We consider only primary solid tumors because primary tumors originate in a specific organ or tissue and are generally more consistent in terms of location, size, and characteristics compared to metastatic tumors (secondary tumors that spread from the primary site). Focusing on these kind of tumors helps maintain statistical validity by comparing similar types of tumors, reducing variability and confounding factors that may arise from studying different metastatic sites. We also utilize the function *anyReplicated()*, which checks the so called biological or primary unit for replicates in the sampleMap of the MultiAssayExperiment object, that corresponds to the first 12 characters of the barcodes for TCGA data. In fact, if two samples have the same 12 characters in their barcodes, then they come from the same patient and can be identified as technical replicated (since we already filtered for the same sample type). Since repeated measurements taken from the same sample are not attractive to the application, we remove them. The outcome *FALSE* indicates that there were no replicates.

Then, other additional pre-processing steps are performed:

- the removal of **FFPE (formalin-fixed, paraffin-embedded)** samples. Broadly speaking, after a biopsy is performed we need to store and preserve the sample. Two major tissue preparation methods are generally used:

  1. FFPE;
  2. freezing the sample.

DNA and RNA molecules are preserved better if the tissue is frozen, thus we will remove samples preserved using FFPE technique;

- the restriction of samples to the ones having all the considered omics, disregarding all samples of patients which do not present mRNA, miRNA and protein data, and the extraction of the set of omics (one matrix for each omic) in a list;
- the transposition of each matrix to have samples in the rows and features in the columns;
- the removal of the features having missing values (e.g. NA). In this case, it is easier to remove features instead of performing imputation, since only few features in the proteomics data have missing values (Figure 5, second row);
- the selection of features having more variance across samples. Here we make a strong assumption: features that have more variance across samples bring more information and are the more relevant ones. This feature selection strategy is fast and commonly used in literature, however it has some drawbacks:

  1. it is univariate, thus does not considers interactions among features. In univariate analysis, any covariation with other variables is explicitly neglected and this may lead to important features being ignored [2];
  2. it is not able to remove redundant variables [2].

  Moreover, we need to identify a threshold for feature selection (in our case, the threshold we selected is defined by the top 100 features) but it is always an arbitrary choice;
- the standardization of features using **Z-score**. Z-score normalization is a statistical technique 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. It is used to transform a dataset so that it has a mean of zero and a standard deviation of one. This process allows us to compare and analyze data that originally had different scales or units. In fact, 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;
- the cleaning of barcodes retaining only *Project-TSS-Participant*, that is, the unique identifier for each patient. We substitute the names of the rows with the substring composed by their first 12 characters. With this modification, we can better identify our data rows and provide a simple method of matching the omics data with the disease subtypes;

4 Prostate adenocarcinoma subtypes

The classification of a sample to a specific disease subtype helps to predict patients' prognosis and it has an impact also on the definition of the therapy. Many different tests to define the disease subtype of a prostate cancer patient are available, which consider different subset of genes for the definition of the subtypes. TCGA Research Network provides the subtypes defined using the *iCluster* model

[19]. We will try to see if the clusters we compute are similar to the *iCluster* disease subtypes for prostate cancer.

We literally are in a new preprocessing phase, this one regarding the subtype dataset instead of the multi-omic set. We will begin fetching the disease subtypes (of prostate adenocarcinoma) from the *TCGAbiolinks* package.

Prostate cancer can be divided into multiple different subtypes classifications. The research by TCGA Network [19] unveiled a molecular taxonomy wherein 74% of the tumors were categorized into seven subtypes based on distinct gene fusions (ERG, ETV1, ETV4, and FLI1) or mutations (SPOP, FOXA1, and IDH1). This classification, regarded as the most significant by the *TCGAbiolinks* package, is employed in their data's *Subtype_Selected* column.
However, the TCGA Network in the same research [19], using the *iCluster* technique, were able to identify also three significant groups of prostate cancers, as already mentioned in the Synopsis: 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.

These subtypes serve as the foundational reference for the study due to their more accessible subtype classification, while still remains relevant to predicted patient prognoses. Moreover, the *iCluster* subtype holds significance within our particular context, as it originates from a multi-omics integration analysis. As such, this research seeks to ascertain whether the computed clusters using the proposed methodologies exhibit similarities to the prostate cancer disease subtypes, namely the *iCluster* subtypes, as designated by TCGA.

Upon data analysis, it becomes evident that not all subtypes are represented within the subset of samples encompassing all the considered omics data sources. Thus, it became imperative to exclusively incorporate samples from the multi-omics dataset with an associated subtype. Through this process, we were able to sift out samples with the disease subtypes in the designated omics, ultimately shaping the final configuration of the data points.

Subtype 1	**Subtype 2**	**Subtype 3**
60	83	105

Table 1: number of samples for each subtype found by *iCluster*

Consequently, the ultimate subtypes database comprised 60 patients of Subtype 1, 83 patients of Subtype 2, and 105 patients of Subtype 3, summing up to a total of 248 chosen patients for the multi-omics data clustering endeavor. To conclude, our data configuration comprises a vector of 3 matrices, each representing different omics. Within each matrix, there are 248 rows (samples/patients) and 100 columns (selected features with high variance).

5 Data Integration methods

We now apply different techniques to integrate the multi-omic data at our disposal. The first step, common to all the techniques utilized is the construction of a similarity matrix among samples for each data source. A similarity matrix is a square matrix that quantifies the similarity between pairs of objects or entities. In the context of clustering or data analysis, it is commonly used to represent the pairwise similarities between data points. The similarity measure exploited to determine the weight of the edge is the **scaled exponential Euclidean distance**. The reasoning for choosing this similarity measure 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 [7].

With the similarity matrices created, we fuse our prostate cancer multi-omic dataset using 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 Similarity Network Fusion [6], implemented in the package *SNFtool*.

5.1 Average

Averaging involves combining data from multiple omic datasets by taking the average values of corresponding measurements. However, it may overlook non-linear relationships, specific interactions, or complex regulatory mechanisms that exist between different omic layers. Overall, multi-omic data integration via averaging is a valuable strategy for gaining a first understanding of biological systems by leveraging multiple layers of molecular information and for providing a benchmark for confronting other results obtained.

5.2 Similarity Network Fusion

Similarity Network Fusion is a data integration method particolarly useful when dealing with heterogeneous data, where each data source provides complementary information about the underlying system.

SNF addresses the challenge of integration by establishing networks of samples (e.g. patients) for each accessible data type. These serve as the foundational elements for integration, subsequently merged into a singular network that encapsulates the entire range of underlying data. Consequently, to construct a holistic understanding of a medical condition using a group of patients, SNF calculates and merges patient similarity networks derived from each distinct data type. This process capitalizes on the complementarity in the data to provide a more comprehensive depiction [6].

This method consists of two main steps (Figure 6):

- the construction of a sample-by-sample similarity matrix for each chosen data type by using a similarity measure. The matrix is equivalent to a similarity network where nodes are samples and the weighted edges represent pairwise sample similarity;
- integration of these networks into a single similarity network using a nonlinear combination method that iteratively updates every network, making it more similar to the others with every iteration. After a few iterations, SNF converges to a single network.

This integrative approach offers the benefit of eliminating weak similarities (represented by low-weight edges), which aids in noise reduction. Concurrently, strong similarities (depicted by high-weight edges) found within one or more networks get incorporated into the rest. Moreover, low-

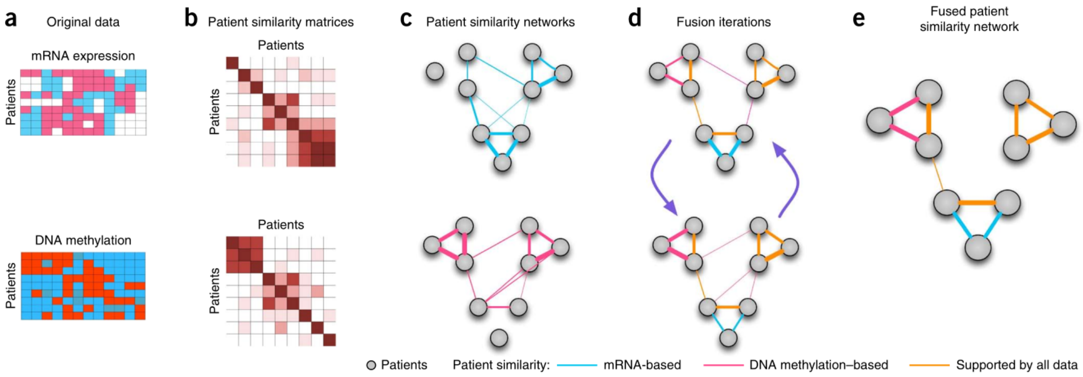


Figure 6: Illustrative example of SNF steps

weight edges endorsed by all networks are preserved based on the degree of interconnectedness observed within their respective neighborhoods across the networks [6].

In our case, $t = 20$ is the number of iterations and $K = 20$ is the number of neighbours to consider to compute the local similarity matrix. SNF has proven successful compared to other relevant fusion methods and outperforms single omics data [6].

6 Clustering Approaches

With the integrated multi-omics data, it is now possible to perform disease subtype discovery using the PAM and the Spectral Clustering algorithms.

As we strive to comprehend the influence of the integration techniques and the clustering algorithms, we executed various combinations of clusterings and similarity matrices, listed as follows:

- clustering of the mRNA similarity matrix using PAM;
- clustering of the miRNA data using PAM;
- clustering of the proteins data using PAM;
- clustering of the average data integration using PAM;
- clustering of SNF data integration using PAM;
- clustering of SNF data integration using Spectral Clustering.

To obtain the similarity matrices for the first three experiments we apply the scaled exponential Euclidean distance to each data source. Then, we cluster indipendently each similarity matrix. Each application of the PAM clustering algorithm requires a distance matrix as input. Therefore, to make each single source (mRNA, miRNA, proteins and average) comparable, it is necessary to normalize the distance matrices.

6.1 Partitioning Around Medoids

The **Partitioning Around Medoids** [20] algorithm is based on the search for a number k (given as input by the user) of representative objects but, instead of the **centrotypes** [20] typical of clustering methods such as K-Nearest Neighbors, PAM employs actual data points within the cluster called **medoids**, which are less sensitive to outliers and can provide better cluster representations when dealing with non-linear or asymmetric data [20]. After finding a set of k medoids, k clusters are constructed by assigning each observation to the nearest one. The goal is to find these k representative objects which minimize the sum of the dissimilarities of the observations to their closest representative object. In other words, the final goal is to obtain a set of clusters where the average distances of objects belonging to the cluster and the cluster representative is minimized (equivalently the sum of the distances can be minimized) [20].
The algorithm has two phases [7]:

- **build phase**: initially, the algorithm randomly selects k data points from the dataset as the initial medoids. The first object selected is the one that has minimal distance with all the other objects, thus the most central data point. The other points are individually evaluated to be selected as representatives and chosen if they have a high number of unselected objects that are closer to them than to already selected representatives. These steps are performed until a number of selected medoids k is reached. Each data point is, then, assigned to the medoid chosen as the closest representative based on a distance metric. The cost of the current clustering solution is calculated as the sum of distances between each data point and its assigned medoid;

- **swap phase**: for each medoid and for each non-medoid data point assigned to it, the algorithm evaluates the potential improvement in the cost if the medoid and non-medoid are swapped. If a swap results in a lower cost, the medoid and non-medoid are swapped, and the cost of the clustering solution is updated accordingly. The swap phase iteratively explores all possible medoid swaps until a predefined stopping criterion is met, such as a maximum number of iterations or a negligible improvement in the cost.

6.2 Spectral Clustering

Spectral Clustering is the other approach we employed. It is a technique with roots in graph theory used for clustering data points based on the spectral properties of a similarity matrix derived from the data and is particularly useful for identifying non-linear and complex structures within the data [21]. Spectral properties are characteristics or information extracted from the eigenvalues and eigenvectors of a matrix. 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 [21]. The Spectral Clustering algorithm allows for detecting clusters that may have complex shapes or are not linearly separable in the original feature space [21]. In this project, Spectral Clustering is performed with the support of the *SNFtool* library, using $k = 3$ as number of clusters.

III Statistical Analysis

When comparing clusterings of multi-omic data, it is important to consider the specific characteristics and requirements of the data. To evaluate the efficacy of our techniques in identifying disease subtypes from the provided multi-omic samples we used the following indices:

- **Rand Index**: (**RI**): the RI is a statistical measure used in data clustering to evaluate the similarity between two clusterings of data. The range of this index spans from 0 to 1, where 0 means that there is no agreement between the two data clusterings on any pair of data points and 1 means perfect agreement, that is, the data clusterings are identical. This measure counts how many pairs of objects are in the same clusters in both C_1 and C_2 (depicted by n_{11}), and how many pairs are in different clusters in both C_1 and C_2 (depicted by n_{00}), considering all the possible pairings [2, 23]. The Rand Index R is calculated as:

$$R(C_1, C_2) = \frac{2(n_{00} + n_{11})}{n(n-1)} = \frac{n_{00} + n_{11}}{n_{00} + n_{10} + n_{01} + n_{11}}$$

- **Adjusted Rand Index** (**ARI**): the ARI is a modified version of the Rand Index widely used measure for comparing clusterings in various domains, including multi-omic data analysis. It accounts for chance agreements and provides a normalized similarity score. While the Rand Index is limited to values between 0 and 1, the Adjusted Rand index can generate negative values (ranging from -1 to 1) if the index falls below the expected value [23];

- **Variation of Information** (**VI**): this measure is commonly used in multi-omic data analysis as it considers the shared information and entropy (informally, the entropy of a clustering C is a measure for the uncertainty about the cluster of a randomly picked element) between two clusterings. It is a measure of the distance between two clusterings and it quantifies the information loss when one clustering is used to represent another, providing insights into the similarity or dissimilarity between the clusterings [23]. The VI index's range goes from 0 to the maximum value determined by the logarithm of the number of elements being clustered;

- **Normalized Mutual Information** (**NMI**): NMI is another commonly employed measure in multi-omic data analysis. The mutual information index provides a means of quantifying the extent to which we can decrease uncertainty about an element's cluster when we possess knowledge about its cluster in another clustering:

$$MI(C_1, C_2) = \sum_{i=1}^{k} \sum_{j=1}^{l} P(i,j) \log_2 \frac{P(i,j)}{P(i)P(j)}$$

  where $P(i,j) = \frac{|C_{1i} \cap C_{2j}|}{n}$ is the probability that an element belongs to cluster $C_i \in C_1$ and cluster $C_j \in Cluster_2$. Since mutual information has no upper bound, a normalized version is easier to interpret:

$$NMI(C_1, C_2) = \frac{MI(C_1, C_2)}{\sqrt{H(C_1)H(C_2}}$$

where $H(C_1)$ and $H(C_2)$ are the entropies associated with clustering C_1 and C_2. NMI values can span from 0 to 1, with the highest value of NMI achieved when C_1 is the same as C_2. he NMI provides a normalized measure that accounts for the inherent differences in the sizes and entropies of the compared sets [23];

- **Fowlkes-Mallows Index**: the Fowlkes-Mallows Index measures the geometric mean of pairwise precision and recall. It can be applicable in comparing clusterings of multi-omic data when considering precision and recall aspects. A higher value for the Fowlkes–Mallows index indicates a greater similarity between the clusters and the benchmark classifications [23];
- **Jaccard Coefficient**: the Jaccard Coefficient is a simple measure that compares the similarity between two clusterings based on the presence or absence of samples in the same or different clusters. It is very similar to the Rand Index, however it disregards the pairs of elements that are in different clusters for both clusterings. It can be used as a quick similarity measure for multi-omic data clusterings [23]. In detail, it compares the intersection and union of the data points assigned to each cluster. It provides a value between 0 and 1, with 1 indicating complete similarity and 0 indicating no similarity. Although it's easy to interpret, it is sensitive to small sample sizes.

$$J(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} = \frac{n_{11}}{n_{10} + n_{01} + n_{11}}$$

Those metrics are already implemented in the *mclustcomp* R package, which we used in our research.

IV Results and conclusions

The process of experimentation involved crafting a set of experiments aimed at establishing connections between the methodologies devised for data integration and clustering, and the Research Questions (RQs) put forth in the Synopsis.

A series of experiments has been designed, encompassing six individual trials. The initial trio of experiments employed PAM clustering along with similarity matrices derived from distinct data sources (miRNA, mRNA, and proteins). These matrices were generated using the usual scaled exponential Euclidean distance. These experiments are denoted as **single omics experiments**, given that each clustering was executed on an independent data source, without implementing any data integration methods.

The subsequent three experiments leveraged multi-omics data and are thus termed **multi-omics experiments**. These trials encompassed PAM clustering with the average data integration technique, PAM clustering in conjunction with SNF integration, and Spectral Clustering combined with SNF integration. This collection of experiments is deemed comprehensive enough to adequately address the proposed research questions.

Experiments	Cluster #1	Cluster #2	Cluster #3
miRNA	91	94	63
mRNA	76	71	101
Proteins	45	120	83
Average integration	79	97	72
SNF integration	78	92	78
Spectral Clustering	83	84	81
iCluster	105	60	83

Table 2: Count of samples in clusters for each experiment

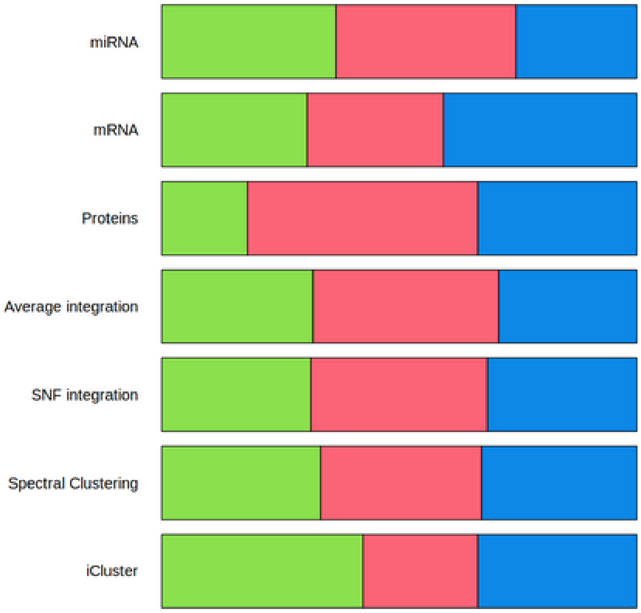


Figure 7: Distribution of the clusters in the computer clusterings

Table II provides an initial summary of the outcomes. The purpose of this table is to display the sample counts within each cluster, illustrate how samples are distributed across clusters in each experiment and validate the clustering process.

Table III summarizes the resulting metrics for each experiment. From this table, it result evident that there is a degree of overlap in the clustering outcomes, albeit to a limited extent. In approaching our questions through table analysis, we are firstly interested in discovering if the multi-omics data outperform the usage of single omics-data I. A discernible trend emerges as all metrics related to SNF Integration and Spectral Clustering outperform the corresponding metrics derived from single omics data. This suggests that a reasoned choice for the integration of the multi-omics data provides better results. However, this pattern doesn't extend to the average integration method, which surpasses the single-omics metrics for proteins and miRNA, but not for mRNA. Noteworthy is the fact that, excluding NMI and VI, the metrics associated with mRNA outperform those of the average integration technique. This might imply that mRNA holds greater clustering power

Experiments	**ARI**	**FMI**	**Jaccard**	**NMI**	**RI**	**VI**
miRNA	0.0270	0.3614	0.2205	0.0426	0.5611	2.9797
mRNA	0.0621	0.3839	0.2375	0.0581	0.5772	2.9352
Proteins	0.0003	0.3618	0.2206	0.0144	0.5379	2.9871
Average integration	0.0419	0.3692	0.2263	0.0704	0.5690	2.9024
SNF integration	**0.1795**	**0.4584**	**0.2973**	**0.1567**	**0.6317**	**2.6388**
Spectral Clustering	0.1191	0.4176	0.2638	0.1172	0.6051	2.7663

Table 3: Summary of the experiments' result metrics

concerning miRNA and proteins, or alternatively, that simplistic data integration methods lack efficacy if the information obtained from a single omic outweighs others in significance. Furthermore, if the biological samples are highly heterogeneous, it's possible that mRNA data captures certain subpopulations more accurately while integration methods might smooth over these differences. All these suppositions could potentially be confirmed by employing a weighted average, assigning a higher weight to mRNA omics.

As the second point of our analysis, we focus on whether the similarity networks fusion integration method can outperform a simple average data integration I. From table III, it is easily noticeable that the SNF integration outperforms the average integration in every possible metrics in both clustering approaches.

Since we are also interested in comparing the PAM algorithm against the Spectral Clustering algorithm, both with SNF data integration I, we can notice that it is equally clear from the table that the PAM clustering outperforms the Spectral clustering since all its metrics are higher (or lower, in the case of the Variation of Information), providing a better similarity with the *iCluster* diseases subtypes.

Our focus shifts now towards a comprehensive analysis of the obtained metrics to provide a response to the last question I. As already mentioned, the data in the table indicates a degree of overlap in the clustering outcomes, although quite limited. Even though the RI stands out as the most positive metric at 0.6317 suggesting that the clusters are more alike to the *iCluster* ones than they are not, the other metrics make clear that the agreement is closer to a result purely by chance than to a perfect agreement. This is confirmed by the 0.1567 NMI and 0.2973 Jaccard similarity, which both indicate a small similarity value between the sets, and by the 2.6388 returned by the VI index, which differs slightly from the ones returned by the same index for the other similarity matrices. This indicates that all the clusters obtained are quite dissimilar than the *iCluster* one. These results were foreseeable because of the differences between the PAM algorithm and the *iCluster* method. The former one is based on traditional clustering techniques while *iCluster* incorporates flexible modeling of the associations between different data types and the variance–covariance structure within data types in a single framework, while simultaneously reducing the dimensionality of the datasets [3].

Furthermore, it is possible that our straightforward data processing might not have adequately extracted pertinent information suitable for the clustering algorithms. As a result, pinpointing subtypes based on the derived features could be more intricate. It is, therefore, possible to retry the experiments mantaining a higher number of features.

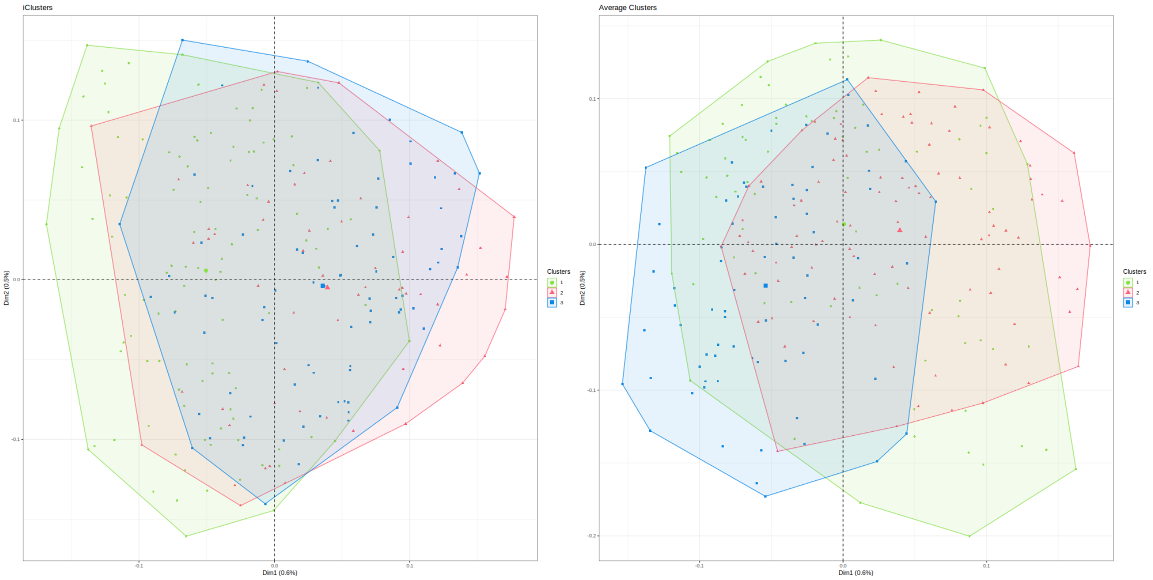


Figure 8: *iCluster* Visualization and Average Integration Cluster Visualization

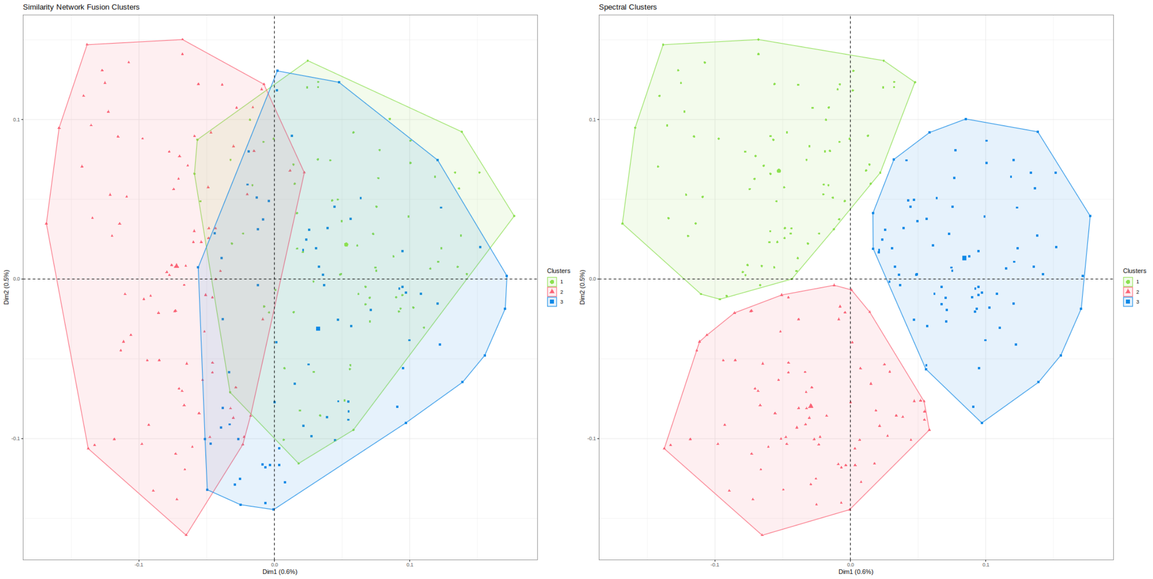


Figure 9: Similarity Network Fusion Integration Partitioning Around Medoids Cluster Visualization and Similarity Network Fusion Integration Spectral Cluster Visualization

To better understand the results of our investigation method, we propose the visualization of the clusters regarding the integration techniques. To reduce the dimensionality of the dataset utilized, we apply the **Principal Component Analysis** (**PCA**) [16, 24]. PCA is a statistical technique used for dimensionality reduction and data transformation in order to simplify complex data while retaining its important underlying patterns [16]. Analizing the *iCluster* plot 8 it is easily understandable that considering only the features with higher variance is not sufficient to permit the clusters' separation. The same pattern appears on the visualization of the other clusters 8 9, with the exception of the Spectral Cluster one 9. In this figure, the clusters are significantly separate: this can be explained by the fact that it operates by transforming the data into a lower-dimensional space where the clusters become more distinguishable [21].

Hence, it can be concluded that the outcomes demonstrated a heightened capability for clustering within the multi-omics data when integrated with SNF. Conversely, employing the simple integration approach may diminish the potential of a singular omics source. Moreover, the findings showed a preference for the PAM algorithm over Spectral Clustering. Consequently, the integration of SNF coupled with PAM clustering yielded a cluster more similar to the *iCluster* ones in our study.

References

[1] Nimrod Rappoport and Ron Shamir. "NEMO: cancer subtyping by integration of partial multi-omic data". In: *Bioinformatics* 35 (2019), pp. 3348–3356. DOI: `10.1093/bioinformatics/btz058`. URL: `https://doi.org/10.1093/bioinformatics/btz058`.

[2] Jessica Gliozzo. "Practice 2 - Disease subtype discovery on multi-omics data". In: *Department of Computer Science, Università degli Studi di Milano* (2023). URL: `https://github.com/GliozzoJ/Bioinformatics_practice2023/blob/main/practice2_clustering.nb.html`.

[3] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis". In: *Bioinformatics* 25 (2009), pp. 2906–2912. DOI: `10.1093/bioinformatics/btp543`. URL: `https://doi.org/10.1093/bioinformatics/btp543`.

[4] Center for Cancer Genomics. *The Cancer Genome Atlas Program.* 2023. URL: `https://www.cancer.gov/ccg/research/genome-sequencing/tcga`.

[5] Carolyne Hutter and Jean Claude Zenklusen. "The cancer genome atlas: Creating lasting value beyond its data". In: *Cell* 173 (2018), pp. 283–285. DOI: `10.1016/j.cell.2018.03.042`. URL: `https://doi.org/10.1016/j.cell.2018.03.042`.

[6] Bo Wang et al. "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature Methods* 11 (2014), pp. 333–337. DOI: `10.1038/nmeth.2810`. URL: `https://doi.org/10.1038/nmeth.2810`.

[7] Jessica Gliozzo et al. "Heterogeneous data integration methods for patient similarity networks". In: *Briefings in Bioinformatics* 23 (2022), pp. 1–26. DOI: `10.1093/bib/bbac207`. URL: `https://doi.org/10.1093/bib/bbac207`.

[8] Milan Picard et al. "Integration strategies of multi-omics data for machine learning analysis". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 3735–3746. DOI: `https://doi.org/10.1016/j.csbj.2021.06.030`. URL: `https://doi.org/10.1016/j.csbj.2021.06.030`.

[9] Paul Pavlidis et al. "Learning gene functional classifications from multiple data types". In: *Journal of Computational Biology* 9 (2002), pp. 401–411. DOI: 10.1089/10665270252935539. URL: https://doi.org/10.1089/10665270252935539.

[10] Anneleen Daemen, Olivier Gevaert, and Bart De Moor. "Integration of clinical and microarray data with kernel methods". In: *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2007), pp. 5411–5415. DOI: 10.1109/IEMBS.2007.4353566. URL: https://doi.org/10.1109/IEMBS.2007.4353566.

[11] Marinka Žitnik and Blaž Zupan. "Data Fusion by Matrix Factorization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (2015), pp. 41–53. DOI: 10.1109/TPAMI.2014.2343973. URL: https://doi.org/10.1109/TPAMI.2014.2343973.

[12] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. "A review on machine learning principles for multi-view biological data integration". In: *Briefings in Bioinformatics* 19 (2016), pp. 325–340. DOI: 10.1093/bib/bbw113. URL: https://doi.org/10.1093/bib/bbw113.

[13] Zahra Momeni et al. "A survey on single and multi omics data mining methods in cancer data classification". In: *Journal of Biomedical Informatics* 107 (2020), p. 103466. DOI: 10.1016/j.jbi.2020.103466. URL: https://doi.org/10.1016/j.jbi.2020.103466.

[14] Vladimir Gligorijević and Nataša Pržulj. "Methods for biological data integration: perspectives and challenges". In: *Journal of the Royal Society Interface* 12 (2015), pp. 1–19. DOI: 10.1098/rsif.2015.0571. URL: https://doi.org/10.1098/rsif.2015.0571.

[15] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2019), pp. 423–443. DOI: 10.1109/TPAMI.2018.2798607. URL: https://doi.org/10.1109/TPAMI.2018.2798607.

[16] Rasmus "Bro and Age K." Smilde. ""Principal component analysis"". In: *"Analytical Methods"* "6" ("2014"), "2812–2831". DOI: "10.1039/C3AY41907J". URL: http://dx.doi.org/10.1039/C3AY41907J.

[17] Anuradha Gopalan et al. "TMPRSS2-ERG Gene Fusion Is Not Associated with Outcome in Patients Treated by Prostatectomy". In: *Cancer Research* 69 (2009), pp. 1400–1406. DOI: 10.1158/0008-5472.CAN-08-2467. URL: https://doi.org/10.1158/0008-5472.CAN-08-2467.

[18] Barry S. Taylor et al. "Integrative genomic profiling of human prostate cancer". In: *Cancer Cell* 18 (2010), pp. 11–22. DOI: 10.1016/j.ccr.2010.05.026. URL: https://doi.org/10.1016/j.ccr.2010.05.026.

[19] Jaeil Ahn et al. Abeshouse. "The Molecular Taxonomy of Primary Prostate Cancer". In: *Cell* 163 (2015), pp. 1011–1025. DOI: 10.1016/j.cell.2015.10.025. URL: https://doi.org/10.1016/j.cell.2015.10.025.

[20] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990. DOI: 10.1002/9780470316801. URL: https://doi.org/10.1002/9780470316801.

[21] Ulrike Von Luxburg. “A Tutorial on Spectral Clustering”. In: *Statistics and Computing* 17 (2007), pp. 395–416. DOI: `10.48550/arXiv.0711.0189`. URL: `https://doi.org/10.48550/arXiv.0711.0189`.

[22] Marcel Ramos et al. “Software for the Integration of Multiomics Experiments in Bioconductor”. In: *Cancer Research* 77 (2017), e39–e42. DOI: `10.1158/0008-5472.CAN-17-0344`. URL: `https://doi.org/10.1158/0008-5472.CAN-17-0344`.

[23] Silke Wagner and Dorothea Wagner. “Comparing Clusterings - An Overview”. In: *Interner Bericht, Universität Karlsruhe (TH) Fakultät für Informatik* 173 (2007), pp. 1–19. DOI: `10.5445/IR/1000011477`.

[24] Statistical tools for high-throughput data analysis. *Principal Component Analysis in R: prcomp vs princomp.* 2017. URL: `http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/`.