Disease subtype discovery using multi-omics data integration

Samuele Bompani, Luigi Foscari

May 2023

1 Multi-omic analysis and prostate cancer subtype

Prostate cancer, the second most common cancer in men globally, is affected by factors such as age, family history, genetics, and race. It exhibits a spectrum of behaviors, from aggressive to indolent forms, in localized cases. Risk stratification systems have been established to predict disease progression and inform treatment choices using clinical and pathological factors. Incorporating molecular features is critical in enhancing risk stratification and optimizing treatment strategies for prostate cancer.

An omic is a suffix used to refer to different fields of study that involve comprehensive analysis of a specific biological component or aspect. It 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. The omic sciences aim to understand the complex interactions and functions of these components to gain insights into biological processes.

There are three main approaches to multi-omics clustering: early, middle and late integration. Early integration combines all omics data into one matrix and applies clustering, but it overlooks the different distributions of values across omics. Middle integration builds a single model that considers all omics, incorporating joint dimension reduction and similarity-based analyses. Feature selection is essential due to the high dimensionality and complexity of the data, but similarity-based methods offer improved runtime and reduced reliance on feature selection. Late integration clusters each omic separately and then integrates the results, ignoring consistent interactions across omics.

Our goal is to discover disease subtypes in the prostate adenocarcinoma TCGA dataset and compare our clustering results with theirs, which used an integrative clustering model named iCluster on multi-omics data. The omics we are going to consider are mRNA, miRNA and protein expression data and the approach involves using the PAM algorithm on various matrices obtained by integrating the omic data with different techniques.

The original results were computed using iCluster, which is a clustering method specifically designed for analyzing data from multiple omics from the same set of samples. It employs a regularized latent variable model to perform joint inference across these data types, resulting in an integrated cluster assignment. By considering the interdependencies among the different omics, iCluster generates a comprehensive and cohesive clustering solution.

2 Multi-omic integration and clustering algorithms

Our approach involved three cases of early integration (one for each involved omic), all using the PAM algorithm, and five cases of middle integration, three of which using the PAM algorithm and the remaining two using spectral clustering. The first step was preprocessing the data and retaining from the multi-omics dataset only samples having an associated subtype. Using the squared euclidean distance we computed the similarity matrices for each omic, we then integrated the matrices using the following techniques: Similarity network fusion or SNF is a data integration method used to combine information from multiple data sources to construct a similarity network and it is particularly useful when dealing with heterogeneous data, where each data source provides complementary information about the underlying system. In our case SNF proved to be very effective. We used SNF to fuse together the similarity matrices to construct an overall similarity matrix following two main steps:

1. The similarity matrices are normalized to avoid scaling issues and an initial transition matrix is computed.

2. To enhance the integration an iterative network diffusion process is applied to the fused similarity matrices for a fixed number or rounds (in our case 20).

The last step allows the information to propagate across the network, refining the similarity estimates. By leveraging SNF, multiple data sources can be effectively combined, allowing for a more comprehensive understanding of complex systems. SNF provides a powerful framework for data integration, enabling the utilization of complementary information from diverse sources to improve the accuracy and reliability of subsequent analyses.

Another approach we used is Neighborhood-based multi-omics clustering or NEMO, it is a straightforward clustering approach for multi-omics data that utilizes similarities and builds upon existing methods. Initially, NEMO creates a similarity matrix based on Euclidean distances for each omic data set, capturing similarities between samples. These individual similarity matrices are then combined into a unified matrix, which we clustered using PAM and spectral clustering. Compared to other methods, this approach enables efficient and simple integration and clustering of multi-omics data. The significant advantage of NEMO is its ability to handle partial data sets, where certain samples are

measured only for a subset of omics data. This last approach is computationally efficient, does not require iterative optimization, is easy to use and can be adjusted to various situations. All it needs is a definition of the distance between two samples in a specific data set.

We also computed the mean across the similarity matrices of the different omics, this naive solution to integration can be considered a baseline and it is generally not recommended.

To test early integration methods we also considered the single similarity matrices computed from the omics. Once the integrated matrices were computed we used two clustering algorithms:

1. The PAM (Partitioning Around Medoids) algorithm is a clustering algorithm used to group data points into clusters based on their similarity. It is an extension of the k-medoid algorithm and it is often employed in data mining and machine learning applications. Instead of centroids PAM employs actual data points within the cluster called medoid, which are less sensitive to outliers and can provide better cluster representations when dealing with non-linear or asymmetric data. In summary, the PAM algorithm iteratively optimizes the selection of medoids to minimize the dissimilarity between data points within a cluster, aiming to create meaningful and cohesive clusters in the data set.

2. Spectral clustering is the other approach we employed, it is a technique used for clustering data points based on the spectral properties of a similarity matrix derived from the data. Spectral properties are characteristics or information extracted from the eigenvalues and eigenvectors of a matrix. Spectral clustering is particularly useful for identifying non-linear and complex structures within the data.

3 Experiments

To evaluate the efficacy of our techniques in identifying disease subtypes from the provided multi-omic samples we used the following indices:

- The Rand index is a statistical measure used in data clustering to assess the similarity between two clusterings of data. This index ranges from 0 to 1, where 0 means that there is no agreement between the two data clusterings on any pair of points, and 1 means that the data clusterings are identical. It counts how many pairs of objects are in the same clusters in both C_1 and C_2 (represented by n_{11}), and how many pairs are in different clusters in both C_1 and C_2 (represented by n_{00}), considering all the possible pairs. The Rand Index R is calculated as:

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

- The adjusted Rand index is a modified version of the Rand index that accounts for chance. It corrects for chance by establishing a baseline through

the expected similarity of pairwise comparisons between clustering defined by a random model. While the Rand Index is limited to values between 0 and +1, the adjusted Rand index can generate negative values if the index falls below the expected value.

- 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 $C_j \in C_2$. 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 can have values from 0 to 1, with the highest value of NMI achieved when C_1 is the same as C_2.

- The Jaccard index measures the similarity between two clusters by comparing 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.

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

4 Results

To evaluate the clusterings obtained with all the different techniques we computed the Rand index, the adjusted Rand index, the normalized mutual information and the Jaccard index on the clusters. Overall the similarity network fusion approach together with the PAM algorithm achieved the best results. Considering the early integration solutions we see very low scores across the board, as expected, focusing especially on the reverse-phase protein array we see an adjusted Rand index very close to zero and Rand index close to 0.5, which by the definition tells us that this clustering is equivalent to a random labeling with respect to the iCluster results. The RNA gene expression values achieved the best results with a Rand score above 0.5 and with a 20% overlap with the iCluster results according to the Jaccard index, furthermore it has the highest mutual information index among all the omics and even higher results than the simple integration based on the mean of the similarity matrices.

	Cluster #1	Cluster #2	Cluster #3
micro RNA gene expr	91	94	63
RNA gene expr 2	76	71	101
RP Protein Array	45	120	83
Mean	79	97	72
SNF (PAM)	78	92	78
SNF (spectral)	138	45	65
NEMO (PAM)	226	2	20
NEMO (spectral)	83	84	81
iCluster	105	60	83

Table 1: Amount of samples per cluster for each clustering

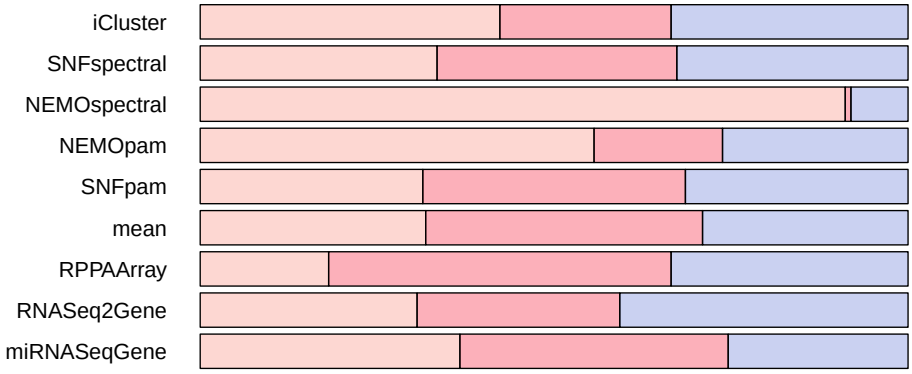

Figure 1: Distribution of the clusters in the computed clusterings

	Adjusted Rand	Jaccard	NMI	Rand
micro RNA gene expr	0.0270	0.2205	0.0426	0.5611
RNA gene expr	0.0621	0.2375	0.0581	0.5772
RP Protein Array	0.0003	0.2206	0.0144	0.5379
Mean	0.0419	0.2263	0.0704	0.5690
SNF (PAM)	**0.1795**	0.2973	**0.1567**	**0.6317**
SNF (spectral)	0.1191	0.2638	0.1172	0.6051
NEMO (PAM)	0.0272	0.2441	0.0568	0.5407
NEMO (spectral)	0.0424	**0.3443**	0.1071	0.4227

Table 2: Indices computed on the obtained clusterings w.r.t the iCluster results

Focusing on middle integration solutions we see that computing the average of the similarity matrices of the single omics does not lead to very high scores in any of the selected indices, but, surprisingly, it is in contention with NEMO,

achieving higher scores in the Rand index and similar results in terms of mutual information. The best results were obtained using similarity network fusion across the board, especially when paired with the PAM algorithm, opposed to the spectral solution. The similarity network fusion integration is a better approach to the Neighborhood-based multi-omics integration, even when paired with the spectral clustering algorithm (as the original paper suggests), this is evident in table 4. The disparity in terms of Jaccard index between the NEMO with spectral clustering approach and all the other solutions is explainable by looking at the Rand index, whose respective formulas are very similar, the only difference is that the Jaccard index does not consider the number of pairs not belonging to the same cluster in both analyzed clusterings, therefore we can conclude that the number n_{00} is very high in this case, which is also evident by looking at the size of the clusters in table 1 and figure 1.

We tried visualizing the data on a 2D plot to inspect the clusters using two different techniques:

- Principal component analysis: a technique used to reduce the dimensionality of the data keeping the maximum amount of information. Unfortunately the obtained representation does not contain enough information, which resulted in a very poor visualization. The total explained variance in the new dimensions was equal to 2% of the original variance.

- t-distributed stochastic neighbor embedding or t-SNE: another dimensionality reduction technique focused on the property that distant objects in the original space remain distant in the mapped space and likewise for close objects. Unfortunately we incurred the same problem and the visualization results were not good enough.

Despite progress, there is still much to uncover about the molecular basis of prostate cancer and its implications for risk stratification. Further research, using integrative multi-omics analysis and exploring novel clustering methods can enhance our understanding of prostate cancer and lead to more effective risk assessment and treatment strategies·