Disease Subtype Discovery Using Multi-Omics Data Integration

Marco Barbaro, Valeria Stighezza
University of Milan
Department of Computer Science "Giovanni Degli Antoni"

***Abstract*—This study addresses the problem of prostate cancer subtype discovery through multi-omics data integration. Three omic layers, respectively mRNA, miRNA, and protein expression, are integrated using Similarity Network Fusion (SNF), matrix averaging and NEighborhood based Multi-Omics clustering (NEMO), to generate a unified representation of the data. Clustering approaches, including Partitioning Around Medoids (PAM) and Spectral Clustering are applied to identify potential subtypes, which are then compared with known reference prostate cancer classifications. The aim is to capture complex molecular patterns, enabling refined diagnostics and therapies in precision medicine while overcoming limitations of traditional methods like gene preselection and independent data analysis.**

I. Introduction

The identification and classification of disease subtypes through *multi-omics data analysis* has become a relevant technique in precision medicine, which aims to improve disease prevention and treatment by taking into account differences in people's genes, environments, and lifestyles.

Multi-omics analysis [1] combines various omic layers, including *geneomics, transcriptomics, epigenomics, proteomics* and *metabolicomics*; through an integrated approach that provides a holistic perspective in understanding disease processes. Each omic layer corresponds to a specific biological feature, therefore the integration of these layers can highlight relationships and patterns that single data type alone can't reveal. Although integrating these different data types is complicated, due to differences in their scale, noise, and biological relevance, it can be very useful for capturing the full spectrum of molecular variation.

Techniques such as *Similarity Network Fusion* (SNF), *Matrix Factorization* (e.g. *Multi-Omics Factor Analysis*, MOFA), Bayesian models (e.g. *iCluster*), *Neighborhood based multi-omics clustering* (NEMO) and deep learning frameworks (e.g. variational autoencoders), merge distinct omic layers to capture biologically meaningful patterns across data types.

This integrated representation is subsequently analyzed by clustering algorithms like *Partitioning Around Medoids* (PAM), *Spectral Clustering, Hierarchical Clustering*, and *k-means*. More recent advancements include graph-based clustering and density-based methods, which better handle high-dimensional data and complex relationships.

This study examines the potential of advanced integration and clustering methods to enhance our understanding of the molecular complexity of prostate cancer.

II. Methods

In this research, we investigate the discovery of the prostate cancer subtypes using a subset of omic layers formed by mRNA, miRNA, and protein expression data, extracted from the *Prostate Adenocarcinoma* (PRAD) *dataset*, which is part of *Cancer Genome Atlas* (TCGA).

A. Data Selection and Preprocessing

To ensure a reliable and accurate analysis, our research begins with the selection of relevant data from the PRAD data set. The data set is first filtered to include only primary solid tumors and no replicated samples. The next step is to exclude from the data set all samples that have been treated with formalin and embedded in paraffin (FFPE). This method is commonly used to preserve biological samples, but it can introduce errors or variations in the results. Therefore, these samples are excluded to avoid any potential distortion or inaccuracy that could affect the analysis. Subsequently, the dataset is filtered to retain only samples with complete data across the three omic layers (mRNA, miRNA, and protein expression), ensuring compatibility for multi-omics integration. Then, features with missing values are removed to eliminate incomplete variables, and, in order to focus on the most informative ones, only the top 100 features with the highest variance are selected for each omic layer. This threshold ensures that key biological signals are prioritized while maintaining a suitable dimensionality for integration and clustering. All features are then standardized using z-score normalization to ensure that the scales across the different omic layers are consistent.

The next step is downloading the disease subtypes identified by *The Cancer Genome Atlas (TCGA) Research Network* [2], which will serve as reference. TCGA used the *iCluster* integrative clustering model [3] to analyze multi-omics data (somatic copy number alterations, methylation, mRNA, microRNA, and protein levels) . It is important to note that not all samples have an associated subtype, so we retain only those with subtype information. We again ensure that only primary solid tumors are included in the analysis.

Once the two dataset are available and ready, we check for matching sample identifiers between the PRAD dataset and the new subtype information: only the samples that appear in both datasets are retained for analysis.

B. Multi-Omics Data Integration

This subsection describes the integration of the multi-omics preprocessed data using three distinct approaches: *Similarity Network Fusion* (SNF), a simple averaging method and *Neighborhood based multi-omics clustering* (NEMO).

SNF [4] is a network-based data fusion technique designed to combine diverse omics dataset, while preserving both shared and complementary information. For each data type, a pairwise similarity matrix is constructed using the scaled exponential Euclidean distance. Each similarity matrix is then interpreted as a network, where nodes represent samples and edges are weighted by the similarity values from the matrix. The SNF prcedure is then applied iteratively with the following parameters to integrate the similarity networks: $K = 20$, representing the number of neighbors used in constructing local similarity matrices, and $t = 20$, denoting the number of iterations for the fusion process. During the diffusion process, at each iteration, the similarity of a sample to its non-neighbors is recalculated indirectly via its neighbors, allowing the networks to propagate information locally within each dataset's network, while gradually integrating complementary information from the similarity networks of the other omics datasets. After t iterations, convergence results in a fused similarity network that combines shared patterns across all datasets while preserving unique contributions from each data type.

Considering a simpler fusion strategy, we calculated an integrated similarity matrix by averaging the individual similarity matrices from each data type. This trivial method serves as a baseline for comparison with more sophisticated approaches like SNF and NEMO.

The NEMO [5] process begins with the calculation of similarities between samples, using Euclidean distance and considering the $k = 20$ nearest neighbors within each omic dataset. Next, NEMO uses the Radial Basis Function (RBF) kernel to normalize the similarity scores, ensuring scaled values that are comparable across different omics types, making the results more robust and meaningful. After normalizing the data, NEMO refines the similarity values between samples by introducing the concept of *relative similarity*, where the similarity between two samples is measured in relation to their closest neighbors, making the comparison more context-sensitive. Once the relative similarity matrices are created for all omics datasets, NEMO combines them into a single matrix called the *Average Relative Similarity* (ARS) *matrix*, which represents the combined similarity between samples, taking into account the relationships captured in each individual omic dataset.

C. Multi-Omics Data Clustering

PAM [6] is a robust clustering algorithm which starts by selecting representative data points, called *medoids*, as the centers of clusters. Each medoid minimizes the total dissimilarity between itself and all other points in its cluster, making PAM less sensitive to noise and outliers. The algorithm iteratively refines the medoids by swapping them with non-medoid points until an optimal configuration is achieved. To identify disease subtypes we utilized the PAM algorithm setting the number of clusters equal to the number of subtypes discovered by iCluster. PAM can handle different types of dissimilarity measures, such as Euclidean or Manhattan distances. Moreover, the method can work with datasets represented by either measurement values or precomputed dissimilarity matrices, making it suitable for a wide range of applications. In our study, PAM clustering is performed on

- similarity matrices derived from individual omics layers (miRNA, mRNA, and protein expression) using the scaled exponential Euclidean distance. These matrices are then normalized and the corresponding distance matrices are computed
- integrated matrix obtained by averaging distance matrices
- integrated matrix produced using SNF integration technique
- integrated matrix produced using NEMO integration technique

In addition to using PAM, NEMO offers the flexibility to perform clustering through an alternative approach, called *Spectral Clustering* [7], which operates on the principles of graph theory and linear algebra, leveraging the structure of similarity graphs to identify clusters. The Average Relative Similarity matrix calculated by NEMO is a natural input for Spectral Clustering, as it encodes both local and global relationships among samples. The algorithm begins by constructing a similarity graph, where nodes are data samples and edge weights denote similarity scores derived from the ARS matrix. This graph is then represented by its Laplacian matrix, which encodes the graph's structural properties. The next step involves computing the eigenvalue decomposition on the Laplacian matrix to extract its principal eigenvectors, allowing for a low-dimensional representation of the data. This transformation enables the data to be analyzed in a spectral space, where clustering algorithms can be applied to identify clusters that correspond to complex and non-linear structures whithin the data.

To evaluate Spectral Clustering in the context of another integration approach, we applied the method to the matrix produced by SNF.

III. Results

In this section, we present the outcomes of our multi-omics data integration and clustering analysis for prostate cancer subtype discovery. The results are evaluated using both external and internal validation metrics to assess the performance of the proposed methods. External validation compares our clustering solutions with the one provided by TCGA Research Network which used the iCluster model, while internal validation focuses on the structural integrity of the clusters. The analysis highlights the comparative effectiveness of various integration and clustering techniques, offering insights into the ability of these methods to capture biologically meaningful patterns.

A. External Cluster Validation Metrics

To validate the performance of our clustering models, we conducted a comprehensive external validation by comparing the results of each model against the output of the iCluster method. The comparison was performed using several external validation metrics, each of which provides a specific measure of similarity and agreement between clustering methods.

One of the employed metrics is the *Fowlkes-Mallows Index (FMI)* [8], which compares the similarity between two clustering methods by measuring how well the clusters from one method correspond to those of another. The FMI uses a *matching matrix*, where each element represents the number of objects shared between clusters from the two methods. This matrix captures the overlap between clusters it serves as foundation for calculating the geometric mean of precision and recall, which together provide a single similarity score, which is the FMI itself. In this context, precision indicates how well the samples grouped in one method correspond to those in the other, and recall defines how well all the samples that should be grouped together in one method are also grouped together in the other. The FMI value ranges from 0 to 1, where value of 1 indicates perfect alignment, meaning the clusters from one method exactly match those of the other.

Another computed measure is the *Rand Index (RI)* [9], which assess the similarity between two data clusterings by comparing how pairs of elements are grouped in two different clustering solutions. The RI operates by considering all possible pairs of data points and categorizing them based on their relationship in the two clusterings. There are four possible outcomes for each pair of elements

- *True Positive* (TP): the pair is assigned to the same cluster in both clusterings
- *True Negative* (TN): the pair is assigned to different clusters in both clusterings
- *False Positive* (FP): the pair is assigned to the same cluster in the second clustering but to different clusters in the first clustering
- *False Negative* (FN): the pair is assigned to different clusters in the second clustering but to the same cluster in the first clustering.

The RI is then calculated as the proportion of pairs that are either TP or TN, relative to the total number of pairs.

The *Adjusted Rand Index* [10] (ARI) is the adjusted counterpart of the RI, as it accounts for the fact that two random clusterings of the same dataset can exhibit some level of similarity purely by chance. In practice, the ARI adjusts the RI by subtracting the expected similarity that would happen purely by chance, providing a more robust evaluation of clustering concordance, unaffected by chance alignments.

The RI is typical much higher than the ARI. Since the RI lies between 0 and 1, the expected value of the RI (although not a constant value) must be greater than or equal to 0. On the other hand, the expected value of the ARI is zero, and its range extends from -1 to 1: 1 means perfect agreement between the two clusterings, while negative values can occur when the agreement is less than expected by chance. Hence, there is a wider range of values that the adjusted Rand index can take on, thus increasing the sensitivity of the index.

An additional metric considered is the *Normalized Mutual Information* (NMI) [11], which quantifies how well two different clustering methods agree with each other. It is based on the concept of *Mutual Information* (MI), which measures the amount of information shared between two variables. MI tells us how much knowing the clusters from one method helps us predict the clusters from another. This is achieved by analyzing the reduction in the entropy of the class labels in our clustering methods when we know the cluster labels identified by iCluster. However, MI alone can be influenced by factors like the number of clusters or the size of the dataset, which can make comparisons between different clustering methods misleading [12]. To address this, NMI normalizes the MI score by taking into account the individual uncertainty (entropy) of each clustering method.

Another evaluated measure is the *Jaccard Coefficient* [13], which quantifies the similarity between two sets by dividing the size of their intersection by the size of their union. In our clustering context, it evaluates how well our predicted clusters align with clusters generated using iCluster, by focusing on data point pairs. It calculates the proportion of correctly clustered pairs to all possible pairs, ranging from 0 (no overlap) to 1 (perfect agreement).

Another evaluation technique we used to compare the clustering results of our models with those from iCluster is *Purity* [14], which measures how well clusters from one method align with those from another. It does so by assigning each cluster to the most frequent label from the comparison model, and calculating the fraction of correctly matched data points. This allows to quantify the overlap and consistency between the two clustering approaches.

Finally, *Variation of Information* (VI) [15] is an information-theoretic measure used to compare two clustering solutions for the same dataset. It evaluates the similarity by analyzing the amount of information gained or lost when transitioning from one clustering to another. VI combines two key concepts: *entropy*, which measures the randomness within each clustering, and *mutual information*, which quantifies the overlap or shared structure between the two. The VI is computed as a combination of the entropies of the two clustering solutions minus twice their mutual information. The VI ranges from 0 to $log(n)$ where n is the number of observations. Typically, a lower VI value indicates greater similarity, while a higher value suggests more divergence. However, in figure, the normalized inverted value of VI is plotted.

B. External Cluster Validation Results

The results produced by the considered external clustering validation metrics are provided in both tabular form (fig. 1) and histogram form (fig. 2). The analysis of these scores reveals that Spectral Clustering with NEMO integration is the best-performing method across most metrics. This consistent superiority underlines its ability to identify, better than other

methods, biologically meaningful clusters that align closely with the reference iCluster subtypes. While PAM-SNF ranks as the second-best clustering method, it does not achieve scores comparable to Spectral-NEMO. Additionally, PAM-NEMO, despite using the same integration method as Spectral-NEMO, shows weaker performance. This suggests that in this context, Spectral Clustering is better suited to exploit the full potential of NEMO's integration capabilities.

The performance among single-layer methods is not uniform. PAM applied to the mRNA layer alone achieves metrics that are comparable to some integration methods, including PAM-SNF, PAM-NEMO and Spectral-SNF; indicating that mRNA data alone is a strong contributor to clustering quality. Conversely, PAM-miRNA exhibits the worst performance across all metrics. These poor results suggest that the miRNA layer may provide less meaningful information or even introduce noise when integrated with other layers.

The poor performance of the miRNA layer warrants further investigation. Given its low standalone clustering quality, future studies should explore whether removing the miRNA layer from the integration process improves overall results. Alternatively, replacing miRNA with another omics layer, such as DNA methylation or metabolomics data, could potentially lead to better integration outcomes.

Clustering Method	FMI	rand	adjrand	nmi1	jaccard	purity	VI
PAM miRNA data	0.2024	0.6474	0.0037	0.0741	0.1026	0.2056	0.1470
PAM mRNA data	0.2330	0.6607	0.0394	0.1346	0.1196	0.2379	0.2000
PAM Protein expression data	0.2325	0.6550	0.0324	0.0843	0.1210	0.2298	0.1623
PAM AVG integration	0.2287	0.6543	0.0291	0.1168	0.1185	0.2339	0.1918
PAM SNF integration	0.2375	0.6670	0.0496	0.1288	0.1206	0.2419	0.1885
PAM NEMO integration	0.2325	0.6593	0.0373	0.1270	0.1197	0.2177	0.1955
Spectral NEMO integration	0.3051	0.6505	0.0791	0.1289	0.1752	0.3024	0.2804
Spectral SNF integration	0.2354	0.6651	0.0460	0.1304	0.1199	0.2298	0.1916

Figure 1. External Cluster Validation Results

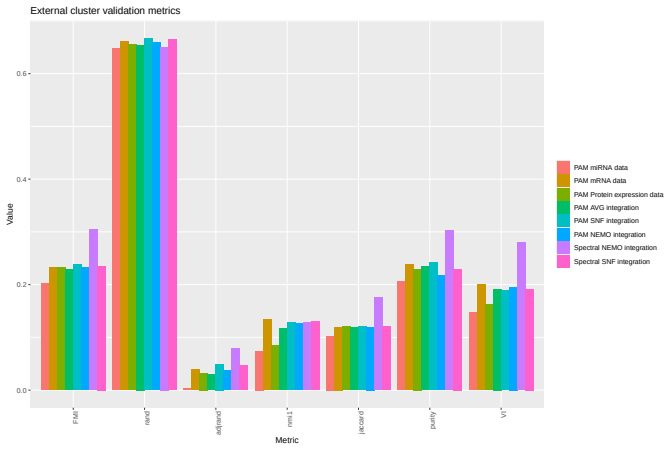

Figure 2. Histogram of External Cluster Validation Results

C. Internal Cluster Validation

In this subsection, we evaluate our clustering solutions based solely on their internal structure, independent of any external clustering solution. This approach focuses on assessing how well the clusters are formed in terms of *cohesion* (how similar data points within a cluster are) and *separation* (how distinct the clusters are from each other).

To achieve this, we used the *Silhouette Score* [16], which is based on the *Silhouette Width*, which measures how similar each data point is to its own cluster compared to the nearest neighboring cluster. A higher silhouette width indicates that the data point is well-matched to its cluster and far from others, while a lower or negative value suggests points near to the cluster boundaries or assigned to the wrong cluster.

The Silhouette Score can be calculated in two ways: the first approach (*Mean Individual Widths*) calculates the mean of the silhouette widths for each data point across all clusters. This gives an overall measure of how well the individual points are clustered. The second method (*Mean Cluster Widths Means*) averages the silhouette widths for each cluster and then computes the mean of these averages across all clusters, providing a summary of how well the clusters are defined as a whole. In either case, the Silhouette Score ranges from -1 to 1, where a value close to 1 indicates well-separated, cohesive clusters, a value around 0 suggests overlapping or ambiguous clusters, and negative values indicate that many points may be assigned to the wrong clusters or are in areas where clusters overlap significantly. The table in Fig. 3 summarizes the internal validation results of our clustering methods using both approaches.

Clustering Method	Mean individual widths	Mean Clusters widths means
PAM miRNA data	0.0080	0.0097
PAM mRNA data	0.0064	0.0071
PAM Protein expression data	0.0077	0.0097
PAM AVG integration	0.0040	0.0049
PAM SNF integration	0.0024	0.0025
PAM NEMO integration	0.0028	0.0034
Spectral NEMO integration	0.0043	0.0149
Spectral SNF integration	0.0018	0.0019

Figure 3. Internal Cluster Validation Results

In general, clustering methods applied to single data types, in particular PAM applied to miRNA Protein expression data, showed the best performance, achieving the highest Mean of Cluster Widths Means (0.0097). On the other hand, integration methods, which combine multiple data types, generally showed lower internal validation scores, suggesting challenges in achieving cohesive clusters when integrating heterogeneous datasets. Among the integration methods, Spectral-NEMO integration performed the best, slightly outperforming PAM-AVG integration and showing clear superiority over PAM-SNF

integration and Spectral-SNF integration, the latter being the worst-performing method in internal validation.

D. Comparing Internal and External Validation Results

When comparing internal validation with external validation, it becomes evident that Spectral Clustering using NEMO integration, which performed better than other methods in external validation, also stands out as the best-performing method in internal validation among those using multi-omics integration. This underlines its potential to achieve a better balance between cluster cohesion, clusters separation and biological interpretability compared to other integration approaches.

For single-data clustering methods, PAM miRNA achieved the best internal validation results, yet it was the worst method in external validation. This discrepancy suggests that while this method produce more tight and cohesive clusters, it may fail to capture the complex biological relationships spanning across datasets.

Overall, the internal validation results indicate generally weak cluster cohesion and separation across most methods, particularly for integration approaches. This aligns with external validation findings, where the results were similarly suboptimal, under the assumption that we are using the iCluster model as a reference (it is important to consider that this study is based on predicted labels, not on true labels). These observations suggest that none of the clustering methods offers a robust solution capable of excelling in both internal and external metrics. This highlights the need for further refinement of clustering algorithms designed specifically for multi-omics data integration to improve both statistical robustness and biological relevance.

REFERENCES

[1] Hasin, Y., Seldin, M., & Lusis, A., Multi-omics approaches to disease. Genome biology, 18, 1-15, 2017.

[2] A. Abeshouse et al., The molecular taxonomy of primary prostate cancer. Cell, vol. 163, no. 4, pp. 1011–1025, 2015.

[3] 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.

[4] B. Wang et al., Similarity network fusion for aggregating data types on a genomic scale. Nature methods, vol. 11, no. 3, pp. 333–337, 2014.

[5] N. Rappoport and R. Shamir, NEMO: Cancer subtyping by integration of partial multi-omic data. Bioinformatics, vol. 35, no. 18, pp. 3348–3356, https://doi.org/10.1093/bioinformatics/btz058, 2019.

[6] Kaufman, L. and Rousseeuw, P. J., Partitioning Around Medoids (Program PAM). In: Finding Groups in Data: An Introduction to Cluster Analysis, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., pp. 68–125, https://doi.org/10.1002/9780470316801.ch2, 1990.

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

[8] Fowlkes EB, Mallows CL (1983). A Method for Comparing Two Hierarchical Clusterings. Journal of the American Statistical Association, 78(383), 553–569. ISSN 0162-1459, 1537-274X.

[9] WARRENS, Matthijs J.; VAN DER HOEF, Hanneke. Understanding the rand index. In: Advanced studies in classification and data science. Springer Singapore, 2020. p. 301-313.

[10] YEUNG, Ka Yee; RUZZO, Walter L. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. Bioinformatics, 2001, 17.9: 763-774.

[11] KNOPS, Zeger F., et al. Normalized mutual information based registration using k-means clustering and shading correction. Medical image analysis, 2006, 10.3: 432-439.

[12] MCDAID, Aaron F.; GREENE, Derek; HURLEY, Neil. Normalized mutual information to evaluate overlapping community finding algorithms. arXiv preprint arXiv:1110.2515, 2011.

[13] VERMA, Vijay; AGGARWAL, Rajesh Kumar. A comparative analysis of similarity measures akin to the Jaccard index in collaborative recommendations: empirical and theoretical perspective. Social Network Analysis and Mining, 2020, 10.1: 43.

[14] FORESTIER, Germain; WEMMERT, Cédric; GANÇARSKI, Pierre. Background knowledge integration in clustering using purity indexes. In: Knowledge Science, Engineering and Management: 4th International Conference, KSEM 2010, Belfast, Northern Ireland, UK, September 1-3, 2010. Proceedings 4. Springer Berlin Heidelberg, 2010. p. 28-38.

[15] MEILĂ, Marina. Comparing clusterings—an information based distance. Journal of multivariate analysis, 2007, 98.5: 873-895.

[16] STARCZEWSKI, Artur; KRZYŻAK, Adam. Performance evaluation of the silhouette index. In: Artificial Intelligence and Soft Computing: 14th International Conference, ICAISC 2015, Zakopane, Poland, June 14-18, 2015, Proceedings, Part II 14. Springer International Publishing, 2015. p. 49-58.