Bioinformatics exam project: Disease subtype discovery using multi-omics data integration

Francesca Ferraro - 16125A

September, 2023

Contents

Description : "The project regards the discovery of disease subtypes using a multi-omics dataset coming from TCGA. The dataset is the Prostate adenocarcinoma dataset (disease code "PRAD"). We will consider as disease subtypes the ones identified in a work performed by The Cancer Genome Atlas Research Network, where they used an integrative clustering model (called iCluster) on multi-omics data (somatic copy-number alterations, methylation, mRNA, microRNA, and protein levels) and discovered three disease subtypes."

1 Introduction

Cancer is a highly heterogeneous disease, with tumors differing widely in their molecular characteristics and clinical outcomes even within the same tissue of origin. This poses challenges in developing effective treatments, as a therapy that works for one tumor may not work for another of the same cancer type. An approach to tackle this heterogeneity is through identification of disease subtypes that capture common biological characteristics within a cancer type.

In recent years, multi-omics data integration has emerged as a powerful approach for disease subtype discovery. By combining multiple layers of molecular information (e.g. DNA mutations, mRNA, miRNA, proteins), we can obtain a more comprehensive view of the biological heterogeneity within a cancer type. A variety of unsupervised clustering techniques have been applied on integrated multi-omics data to identify disease subtypes, including consensus clustering, hierarchical clustering, and network-based approaches like Similarity Network Fusion (SNF).

Compared to single omics analysis, multi-omics integration provides a more holistic characterization of the molecular phenotypes within a disease. Different types of omics data can capture complementary aspects of the underlying biology. For example, genomic data identifies mutations driving tumor development, transcriptomic data reflects gene expression patterns, and proteomic data measures levels of functional protein products. Integrating these diverse data types allows for a more accurate and granular dissection of the heterogeneity in a cancer population.

Furthermore, subtype discovery from integrated multi-omics data has been shown to better recapitulate known biology compared to individual data types. In a pan-cancer analysis of 12 tumor types, multi-platform subtypes identified by iCluster more closely matched tissue of origin and known pathways compared to clustering on mRNA, miRNA, or DNA methylation alone. This demonstrates the benefits of data integration for extracting biologically meaningful disease subtypes.

A variety of computational methods have been developed for multi-omics data integration. Simple techniques like concatenation or averaging generate an integrated dataset but ignore complex interactions between data types. More advanced methods aim to model nonlinear relationships. These include multiple kernel learning, graph-based integration like SNF, and joint dimensionality reduction techniques like iCluster and moCluster.

In this project, I aim to discover disease subtypes for prostate adenocarcinoma, identified by Abeshouse et al. [1] using the iCluster model [2] to integrate multiple omics data types. Similarity Network Fusion (SNF) [3] is proposed in this work as an alternative approach for multi-omics data integration through network diffusion. For clustering, I used the PAM algorithm [4].

2 Methods

2.1 Data

I obtained multi-omics prostate cancer data from TCGA via the curatedTCGAData Bioconductor package. This package provides data from TCGA, a program that collected more than 11,000 cases across 33 tumor types, from multiple biological data sources.

This is a multi-omics dataset, i.e. a dataset composed of different biological data sources, where each of them is a different data modality that captures the state of a specific biological layer in the cells. In effect, each biological system is the result of the interplay of multiple different molecules, and for this reason considering many biological data sources at the same time gives us a more accurate point of view of the inside processes.

The data types I dowloaded are:

- **mRNA** (RNA-seq), that is a molecule that carries genetic information copied from DNA in the nucleus to the cytoplasm, where it is used as a template for making proteins.
- **miRNA**, a molecule that regulates mRNA availability and protein production.
- **Protein** (Reverse-phase protein array), that is the actual production of proteins based on the genetic code instructions in mRNA.

The structure that stores these data is the MultiAssayExperiment, designed to store and coordinately analyze multi-omics experiments.

In TCGA each patient is identified by a barcode with a specific structure, in which the first 12 characters identify a specific individual, while the other parts contain indications about the type of sample (i.e. primary, metastatic, solid, blood derived, etc), the type of genomic material extracted (i.e. DNA, RNA) and other information related to technical replicates.

Only primary solid tumor samples were retained by filtering based on the sample type "code" field, because they are identified by the 01 code. This removed samples from metastatic sites, normal tissue, recurrent tumors, and primary blood-derived cancers, and allowed to have a more homogeneous group.

In addition, FFPE (formalin-fixed, paraffin-embedded) samples were excluded as formalin fixation can degrade RNA. The FFPE status was obtained from the sample annotation data frame.

Samples having all 3 omics data types were identified via the intersectColumns function. This identifies samples with non-missing data across the data sets, enabling integrated analysis.

For each omics data type, the following pre-processing steps were applied:

- Removal of features having missing values (i.e. NA).

- Features with near zero variance were removed using the nearZeroVar function. This eliminates features that are unlikely to distinguish between subtypes, with the assumption that features that have more variance across samples bring more information and are the more relevant ones.
- The top 100 most variable features were retained via sorting on variance.
- Features were normalized to z-scores using the scale function. Z-score normalization centers and scales the features to enable combining datasets, by computing the standard deviation from the mean of the distribution.
- Barcodes were cleaned to retain only the first part specific for each individual ("Project-TSS-Participant").

Reducing to the 100 highest variance features removes noisy low information features while retaining markers that best distinguish heterogeneity between samples. Z-score normalization scales the features to enable combining datasets.

Disease subtypes were obtained from TCGA Biolinks, which includes curated analyses performed by the TCGA Research Network. Specifically, the "Subtype Integrative" column containing the iCluster multi-omics subtypes was extracted.

As not all subtype information was available for the filtered multi-omics samples, I subset the data to only samples having both molecular data and an assigned iCluster subtype. The sample identifiers were matched by IDs between the multi-omics and subtypes data to ensure alignment.

2.2 Data Integration

I integrated the multi-omics data using two approaches:

- **SNF**: implemented in the CRAN package SNFtool, constructs sample similarity networks for each data type and iteratively pushes samples with similar local network structure closer together, for 20 iterations with K=20 neighbors. This allows modeling of nonlinear relationships between the omics.

  The first step in the SNF algorithm is the construction of a similarity matrix among samples for each data source. I used scaled Euclidean distance to calculate these initial similarity matrices. The scaled Euclidean distance between two samples i and j is defined as:

  $$d_{ij} = \sqrt{\sum_{f=1}^{F} (\frac{x_{if} - \mu_f}{\sigma_f})^2}$$

  Where x_{if} is the value of feature f for sample i, μ_f and σ_f are the mean and standard deviation of feature f, computed over all samples. This transforms each feature to have mean 0 and variance 1 before computing distances.

- **Average**: Distance matrices were averaged across data types.

The first step in the SNF algorithm is the construction of a similarity matrix among samples for each data source. Then, other two matrices are derived. One is a "global" similarity matrix, which is needed to capture the overall relationships between patients, and the other one is a "local" similarity matrix, that captures the local structure of the network because considers only local similarities in the neighborhood of each individual, setting to zero all the others.

SNF iteratively propagates information between the networks. Samples that are highly ranked neighbors of each other in multiple networks get "fused" together, even if they are not nearest neighbors in any single network. This models nonlinear dependencies between the omics.

In contrast, simply averaging the distance matrices assumes an additive linear model. It does not account for interactions or dependencies between the different layers of omics data.

By modeling nonlinear relationships, SNF can potentially integrate complementary information across diverse data types. This enables a more holistic view of the underlying biology than linear averaging.

The iterative network fusion process allows SNF to preserve both shared and unique structure between the omics data types. This makes it a powerful technique for integrative multi-omics analysis.

2.3 Clustering

Disease subtype discovery was performed by clustering samples based on the distance matrices from integration. Clustering was done using:

- Distance matrices for individual data types
- Integrated distance matrix from SNF
- Integrated distance matrix from average

In all cases, I used the PAM (partitioning around medoids) algorithm with number of clusters set to 3 to match the 3 iCluster subtypes. PAM is more robust to outliers compared to k-means clustering.

The PAM algorithm initializes by selecting k medoid samples representing the cluster centers (BUILD PHASE). The first object in S is the one that has minimal distance with all the other objects, thus the most central data point. All other samples are assigned to their closest medoid. The medoids are iteratively updated to minimize the sum of distances between samples and their assigned medoid. This continues until the medioids stabilize.

The selection of representatives is done in the SWAP PHASE, where for each pair of representative $i \in S$ and non-representative $h \in U$:

- Swap i and h, as that h is a representative and i is not.

- Compute the contribution K_{jih} of each object $j \in U - \{h\}$ to the swap of i and h. We can have two main situations:

  1. $d(j,i) > D_j$, where D_j is the dissimilarity between j and the closest object in S. Then, $K_{jih} = min\{d(j,h) - D_j, 0\}$.

  2. $d(j,i) = D_j$. Then, $K_{jih} = min\{d(j,h), E_j\} - D_j$, where E_j is the dissimilarity between j and the second closest object in S.

- Compute the total results of the swap as $T_{ih} = \sum\{K_{jih} | j \in U\}$.

- Select the pair i, h that minimizes T_{ih}.

- If $T_{ih} < 0$ the swap is performed, D_j and E_j are recomputed and we return at the first step of the SWAP phase. Otherwise, the algorithm stops if all $T_{ih} > 0$.

2.4 Evaluation Metrics

To assess the biological relevance of the discovered subtypes, I compared the clusterings against the iCluster subtypes using the following evaluation metrics:

- **Rand Index**: Quantifies the similarity between two clusterings based on concordant and discordant pairs. It does not correct for chance grouping.

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

  Where TP and TN are the number of true positive and true negative pairs, and FP and FN are false positive and false negative pairs. Values range between 0 and 1 with higher indicating more concordance.

- **Adjusted Rand Index**: Compares the overlap between two clusterings. Values range from 0 (no agreement) to 1 (perfect agreement).

$$ARI = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex}$$

  The ARI adjusts the RI score to account for agreement that could occur by chance.

- **Normalized Mutual Information**: Measures the mutual dependence between two clusterings as the ratio of their shared information over their joint entropy. It is computed as:

$$NMI(X,Y) = \frac{I(X,Y)}{\max(H(X), H(Y))}$$

  Where $I(X,Y)$ is the mutual information between clusterings X and Y, and $H(X)$ and $H(Y)$ are the entropies. Values range from 0 to 1 with higher indicating greater similarity.

These metrics quantify the level of agreement between the discovered subtypes and the biology-driven iCluster subtypes from different perspectives. The ARI accounts for chance grouping, while NMI measures mutual information. RI is more intuitive but does not correct for randomness.

Together they provide a comprehensive evaluation of subtype reproducibility. Higher scores indicate the identified clusters better recapitulate known disease heterogeneity.

3 Results

The evaluation metrics comparing the discovered subtypes to the iCluster subtypes are shown in the following Table:

Approach	**AdjRand**	**NMI**	**Rand**
SNF integration	0.0871	0.1451	0.6837
Average integration	0.0592	0.1336	0.6704
miRNA	0.0452	0.1177	0.5552
mRNA	-0.0079	0.0259	0.5433
proteins	0.0272	0.0699	0.5860

Table 1: Evaluation metrics for each integration approach

3.1 Comparison of Integration Approaches

Among the single data source approaches, mRNA and proteins had the lowest performance with AdjRand values of -0.00789, NMI value of 0.0259 and Rand of 0.5433. miRNA performed better than the other single sources with an AdjRand of 0.0452, NMI of 0.1177 and Rand of 0.5860. However, all three sources had almost the same Rand Index. This suggests that no single omics contained clear subtype information according to the adjusted metrics, though miRNA provided some signal.

Comparing integration methods, **SNF** had higher AdjRand (0.0871), NMI (0.1451) and Rand (0.6837) values than average integration (AdjRand 0.0592, NMI 0.1336, Rand 0.6704) and all individual sources. These improved metric scores indicate that SNF integration was most effective at recapitulating the known subtypes.

The Rand Index values were the highest across approaches, followed by NMI and then AdjRand. The adjusted metrics account for chance grouping and provide a conservative assessment of clustering performance relative to the true labels of iCluster.

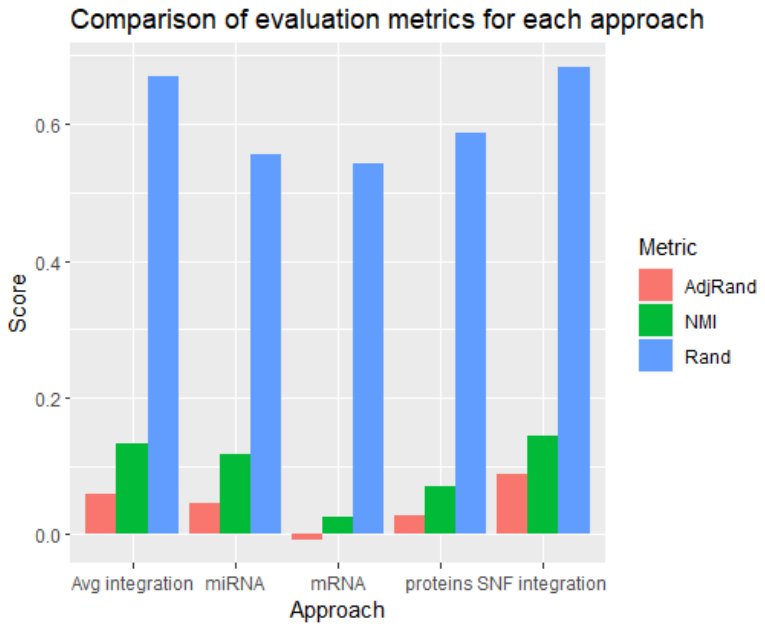

Figure 1: Plot of the results between each considered approach

4 Discussion

Based on the evaluation metrics, SNF integration outperformed both average integration and analysis of individual omics data sources, achieving the highest AdjRand, NMI and Rand Index scores. This suggests SNF was effective at integrating the multi-omics datasets and extracting subtype information that no single data type contained clearly.

However, the overall metric scores still indicate room for improvement in accurately recovering the original reference subtypes. Different clustering algorithms, parameters, preprocessing approaches or additional data could help refine the discovered subtypes.

Among individual sources, miRNA showed some signal, but lower than SNF integration. mRNA and proteins had very similar low performance, indicating little clear subtype information in any single omics modality. This reinforces the benefit of multi-omics integration to compensate for noise or limitations in individual datasets. Optimizing data preprocessing remains important to maximize integration accuracy.

In general, the difficulty of disease subtype discovery was highlighted by no approach perfectly reconstructing the reference subtypes. The analysis provides a stronger validation of SNF integration for this multi-omics prostate cancer problem.

References

[1] Abeshouse, A., et al. "The molecular taxonomy of primary prostate cancer." Cell 163.4 (2015).

[2] Shen, R., et al. "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis." Bioinformatics 25.22 (2009).

[3] Wang, B., et al. "Similarity network fusion for aggregating data types on a genomic scale." Nature methods 11.3 (2014).

[4] Kaufman, Leonard, and Peter J. Rousseeuw. "Partitioning around medoids (program PAM)." Finding groups in data: an introduction to cluster analysis 144 (1990).