Bioinformatics LAB 8

Multi Omics

Prof.ssa Elisa Ficarra
Prof.ssa Santa Di Cataldo
Eng. Marta Lovino
Eng. Alessio Mascolini

Politecnico di Torino
DAUIN
Dept. of Control and Computer Engineering

Important communications

- How projects will be organized (all details will be provided on Dec 22nd)
- Next consulting times: Dec 15th and Dec 22nd, then Christmas break. Only other two consulting times will be granted in January (I keep you updated on the dates).

LAB 8 - Objectives

Objectives

- Familiarize yourself with Multi-omics datasets
- Understand the main Multi-omics drawbacks
- Test differences between early and late integration approaches

Multi-omics. What is it?

- Reduction of costs for the sequencing of biological molecules
- An omic is a specific view on the biological problem (e.g. genome, proteome, transcriptome, epigenome, metabolome) and the name ‘omic’ refers to the end of the view.
- Many omic data can be taken into account: gene expression data (**mRNA**), microRNA expression data (**miRNA**), methylation data (**meth**), proteomics data (**prot**).
- Two strands are typically available in multi-omics analysis: first the **subdivision of the samples** into classes and second the **identification of specific pathways** and gene patterns

Multi-omics approaches

Early integration

Limitations:

- High dimensionality of the features
- Differences in the feature dynamics (range values)
- Feature selection/dimensionality reduction problems. Which is the suitable one?

Late integration

Limitations:

- How to properly merge the results?
- For classification: is the majority voting the best technique? Can we consider also the class membership probabilities?

Multi-omics classification

Given the clssification on a single omic, the consensus for a sample is built according to the next formula:

$$y_{consensus} = \begin{cases} Unknown, & \text{if } \max_i(S_m) < th \text{ or } \max_i(S_i)/S_a < tr \\ \arg\max_i(S_i), & \text{otherwise} \end{cases}$$

- n, m: the number of the omics and the number of the classes,
- th, tr: threshold on the omics and on the classes respectively,
- P_{ij}: the class membership probability for class i and omic j,
- $S_i = \sum_{j=1}^{n} P_{ij}$: the sum of the probabilities on all the omics for a single class,
- $S_a = \sum_{i=1}^{m} S_i$: the sum of the probabilities on all the omics and all the samples,
- $S_m = S_i/n$: the mean of the probabilities on all the omics for a single class.

Multi-omics classification

		Predicted					
		Healthy		Tumor			
		KIRC	KIRP	KICH	KIRC	KIRP	Unknown
Actual	healthy KIRC	5	0	0	0	0	0
	healthy KIRP	0	6	0	0	0	0
	tumor KICH	0	0	17	0	0	0
	tumor KIRC	1	0	5	119	2	1
	tumor KIRP	0	0	1	3	68	0

Table: Early Integration using MLP model

		Predicted					
		Healthy		Tumor			
		KIRC	KIRP	KICH	KIRC	KIRP	Unknown
Actual	healthy KIRC	3	0	0	0	0	2
	healthy KIRP	0	5	0	0	0	1
	tumor KICH	0	0	15	0	0	2
	tumor KIRC	0	0	2	114	1	11
	tumor KIRP	0	0	0	0	69	3

Table: Late Integration using MLP model

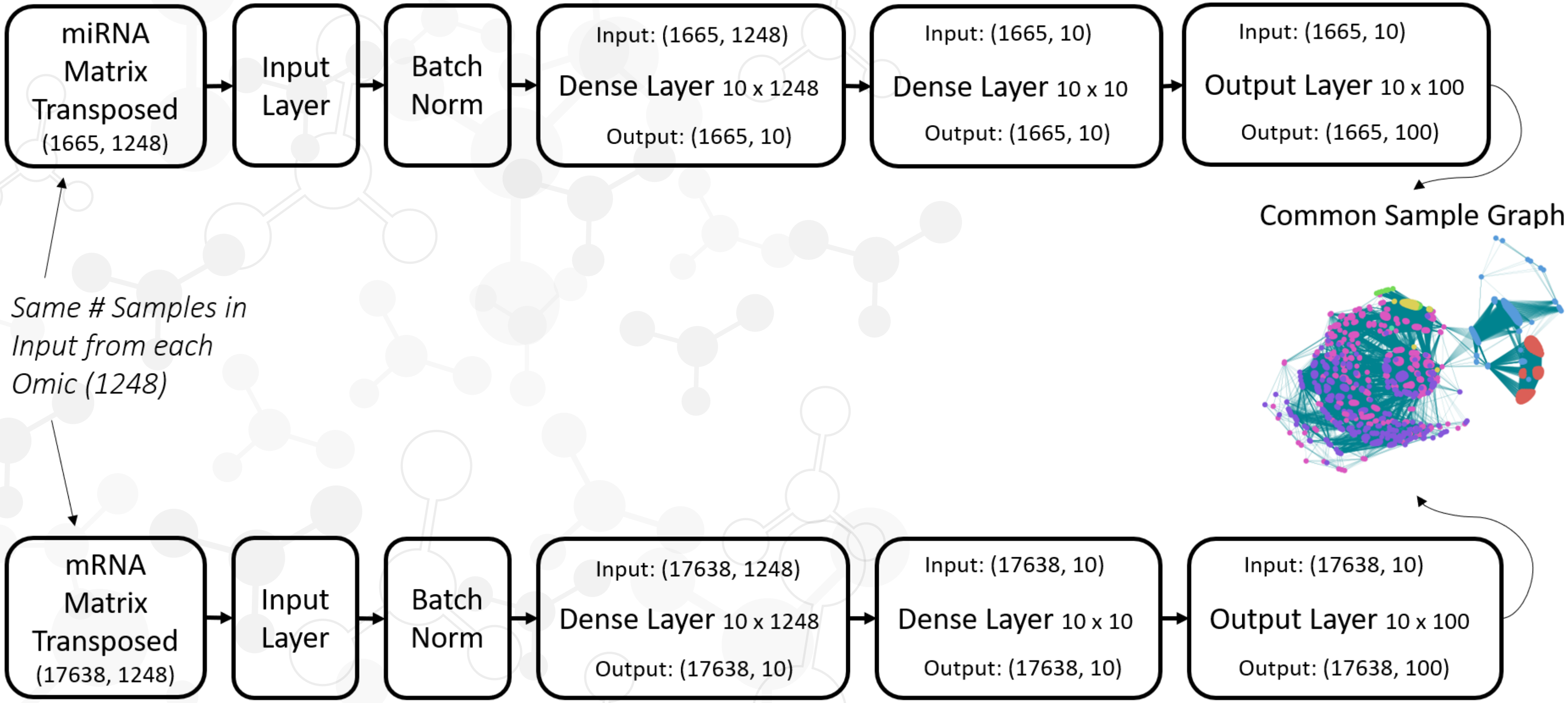
		Integration		
		Early	Late	Samples
Dataset	stomach	2.7%	97.3%	37
	lung	5.3%	29.7%	817

Table: *Unknown* in stomach and lung datasets

Main results

- Late integration is more precise than ealy integration approach
- Late integration is more effective in the identification of samples outside the training domain

Multi-omics clustering



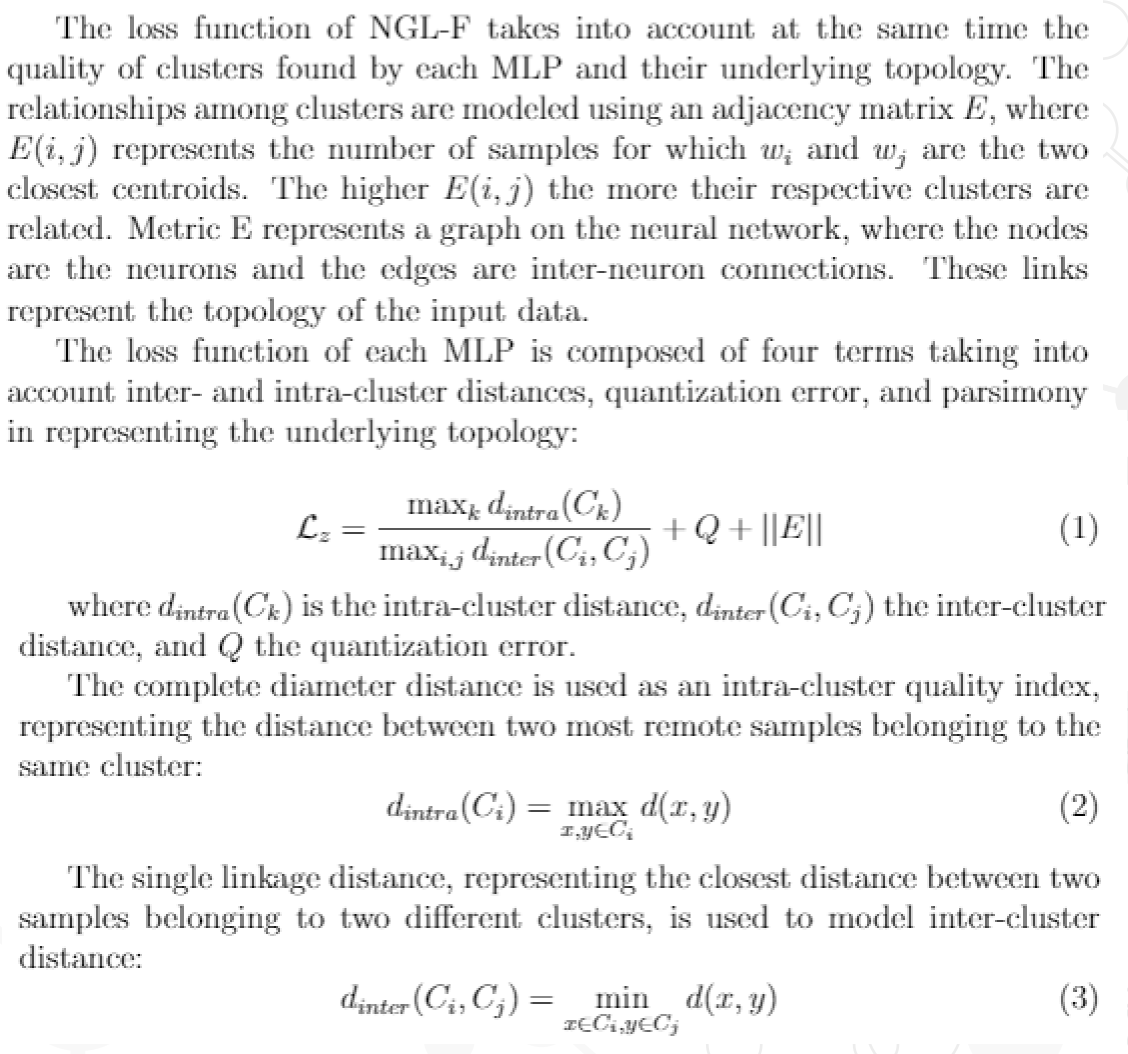
miRNA Matrix Transposed (1665, 1248) → Input Layer → Batch Norm → Input: (1665, 1248) Dense Layer 10 x 1248 Output: (1665, 10) → Input: (1665, 10) Dense Layer 10 x 10 Output: (1665, 10) → Input: (1665, 10) Output Layer 10 x 100 Output: (1665, 100)

Common Sample Graph

Same # Samples in Input from each Omic (1248)

mRNA Matrix Transposed (17638, 1248) → Input Layer → Batch Norm → Input: (17638, 1248) Dense Layer 10 x 1248 Output: (17638, 10) → Input: (17638, 10) Dense Layer 10 x 10 Output: (17638, 10) → Input: (17638, 10) Output Layer 10 x 100 Output: (17638, 100)

Multi-omics clustering



The loss function of NGL-F takes into account at the same time the quality of clusters found by each MLP and their underlying topology. The relationships among clusters are modeled using an adjacency matrix E, where $E(i,j)$ represents the number of samples for which w_i and w_j are the two closest centroids. The higher $E(i,j)$ the more their respective clusters are related. Metric E represents a graph on the neural network, where the nodes are the neurons and the edges are inter-neuron connections. These links represent the topology of the input data.

The loss function of each MLP is composed of four terms taking into account inter- and intra-cluster distances, quantization error, and parsimony in representing the underlying topology:

$$\mathcal{L}_z = \frac{\max_k d_{intra}(C_k)}{\max_{i,j} d_{inter}(C_i, C_j)} + Q + ||E|| \qquad (1)$$

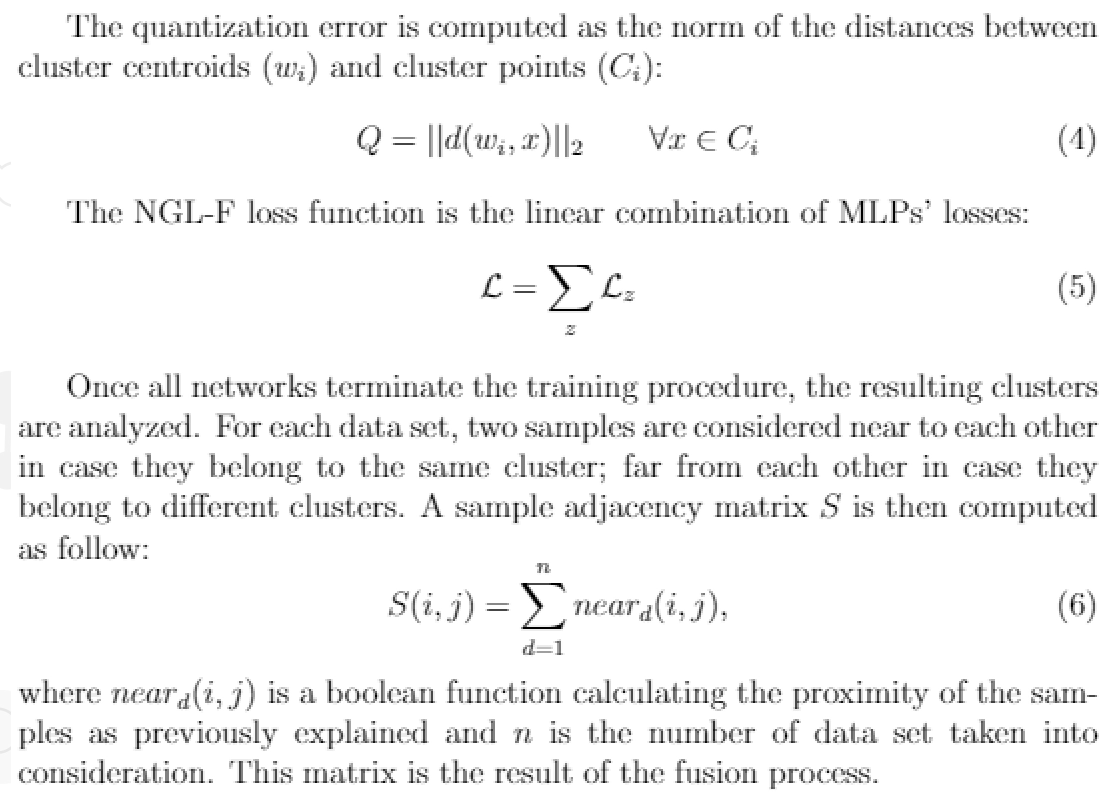
where $d_{intra}(C_k)$ is the intra-cluster distance, $d_{inter}(C_i, C_j)$ the inter-cluster distance, and Q the quantization error.

The complete diameter distance is used as an intra-cluster quality index, representing the distance between two most remote samples belonging to the same cluster:

$$d_{intra}(C_i) = \max_{x,y \in C_i} d(x,y) \qquad (2)$$

The single linkage distance, representing the closest distance between two samples belonging to two different clusters, is used to model inter-cluster distance:

$$d_{inter}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x,y) \qquad (3)$$



The quantization error is computed as the norm of the distances between cluster centroids (w_i) and cluster points (C_i):

$$Q = ||d(w_i, x)||_2 \qquad \forall x \in C_i \qquad (4)$$

The NGL-F loss function is the linear combination of MLPs' losses:

$$\mathcal{L} = \sum_z \mathcal{L}_z \qquad (5)$$

Once all networks terminate the training procedure, the resulting clusters are analyzed. For each data set, two samples are considered near to each other in case they belong to the same cluster; far from each other in case they belong to different clusters. A sample adjacency matrix S is then computed as follow:

$$S(i,j) = \sum_{d=1}^{n} near_d(i,j), \qquad (6)$$

where $near_d(i,j)$ is a boolean function calculating the proximity of the samples as previously explained and n is the number of data set taken into consideration. This matrix is the result of the fusion process.

Multi-omics clustering

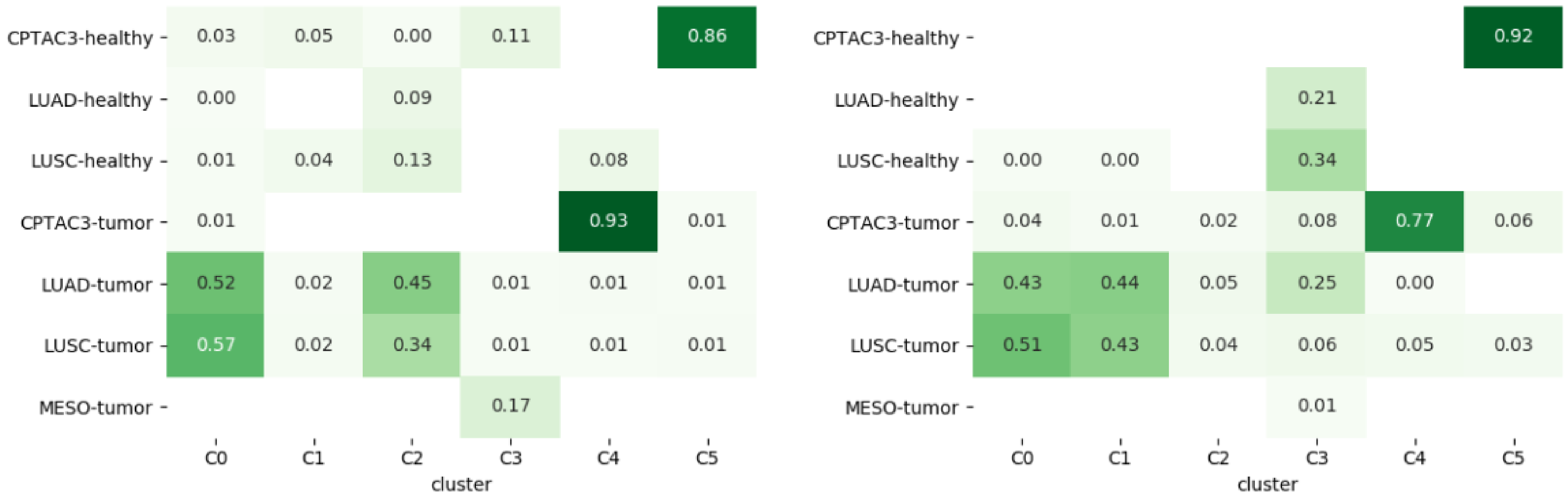
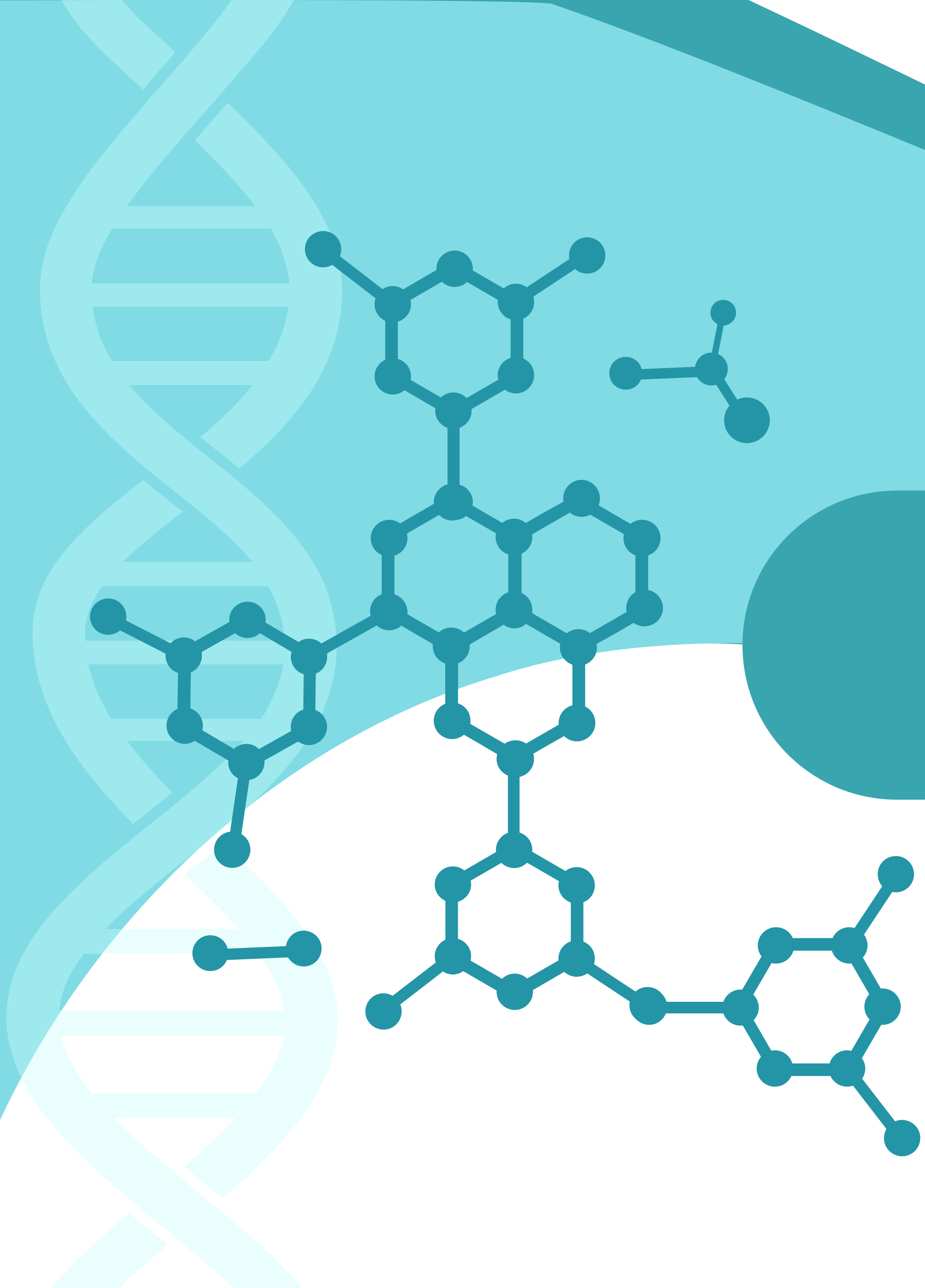


Figure: Harmonic mean of cluster efficiency and purity on SNF (left) and NGL-F (right) algorithms



LAB 8 - Assignments

Assignment 1: Implement a Multi-omics classifier

Download *simulated_dataset.rar* from the Teaching Portal. After decompressing the file, you will have: mRNA.*txt*, *meth.txt*, *prot.txt* which contain respectively transcriptome, genome and proteome of a simulated dataset. In *clusters.txt* you can find the label number for each sample.

Compare early and late integration approaches for sample classification.

- for both approaches implement 4 different classifiers. One of these must output not only the class label but also its probability
- Pay attention in balancing the classes and in the feature selection/dimensionality reduction process.

Which are the differences in early and late integration classifiers in terms of performances? The performances are affected by the type of classifier? Which differences can you notice looking at the probability?

Assignment 2: Implement a Multi-omics clustering

Download *simulated_dataset.rar* from the Teaching Portal. After decompressing the file, you will have: mRNA.*txt*, *meth.txt*, *prot.txt* which contain respectively transcriptome, genome and proteome of a simulated dataset. In *clusters.txt* you can find the label number for each sample.

Compare early and late integration approaches for sample clustering.

- for both approaches implement 2 different clustering techniques.
- Pay attention in balancing the classes and in the feature selection/dimensionality reduction process.

Which are the differences in early and late integration clustering in terms of performances? The performances are affected by the type of clustering?

Project extension 1: Multi-omics classifier

Develop your own integration method for multi-omics classification. You have to look at the classification methods in the literature and compare your own integration method with at least two existing ones.

- Which are the advantages of your approach?
- How much your approach is new?
- How much your approach is reliable?

Test your method on the real biological dataset from GDC data portal (details on that part will be provided in a lesson after Christmas break).

Project extension 2: Multi-omics clustering

Develop your own integration method for multi-omics clustering. You have to look at the clustering methods in the literature and compare your own integration method with at least two existing ones.

- Which are the advantages of your approach?
- How much your approach is new?
- How much your approach is reliable?

Test your method on the real biological dataset from GDC data portal (details on that part will be provided in a lesson after Christmas break).

LAB8 – Take home message

- Appropriately handling data spaces
- High feature dimensionality
- Class imbalances, and how to balance them
- Usually, low number of samples

Questions?

Remember:
no question is
stupid