Ein Mann, der recht zu wirken denkt,
Muß auf das beste Werkzeug halten
The man who seeks to be approved,
must stick to the best tools for it

1 Methodology

1.1 Working phases

Working phases, with the corresponding steps, followed in order to achieve the above objectives:

1. Application of integrative multi-omics methods to (I) the analysis of specific data sets provided by research units from our former affiliation center, VHIR, and other research institutions that we collaborate with [34, 36, 37] and (II) to the integrative analysis of larger data sets from public data bases, such as Breast Cancer samples from the TCGA project [18, 19].

2. Development of methods, either in terms of new algorithms or in terms of combinative workflows, which will be able to improve, and facilitate, the analysis and biological interpretation of those data sets to be integrated.

3. Implementation of the methods developed for this study in the appropriate bioinformatics tools, such as an R package or a web-based application, to facilitate their use in the context of biomedical research projects.

Here follows a brief description of these main five activities, the methods in which they are initially based, the objectives that they are related to, and the corresponding results:

1. Application of some state-of-the-art methods for integrative multi-omics data analysis to the study of human brain tissue samples, collected by the Neurovascular Diseases Laboratory at Vall d'Hebron Research Institute. This part is already finished, and led to publications in 2018 and 2021 [37, 38]. Researchers obtained different omics data from necropsies, which had been processed to obtain mRNA, microRNA and protein expression values. Each dataset had been first analyzed independently using standard bioinformatics protocols [20]. These analyses allowed selecting subsets of relevant features, for each type of data, to be used in the integrative analysis. Among all available options, we decided to use two distinct and complementary approaches: (I) Multiple Co-inertia Analysis implemented in Bioconductor packages made4 [21] and mogsa [22], and (II) Regularized Canonical Correlation Analysis with Sparse Partial Least Squares regression (sPLS), provided by mixomics R package [23]. This work had been presented at some meetings [39, 40, 41, 43] and in an already published extended abstract's series book [35]. This step had been obviously useful for the achievement of the objective number 3 explained in the previous section, which aims on the study of the regulome's response to ischemic stroke, but also useful for detecting the advantages and drawbacks of the methods applied, thus setting the basis for the work regarding to objective number 2.

2. Reproduction of the same analyses steps performed in point 1) above with publicly available databases, such as distinct omics data from 150 samples from the TCGA-BRCA collection. This data set contains the expression or abundance of mRNA, miRNA and proteomics for 150 breast cancer samples previously prefiltered, as explained in Rohart et al. [29], and allows identifying a good multi-omics signature to discriminate between Basal, Her2 and Luminal

A breast cancer subtypes. This work is already finished, and complies with objectives 3 and 2.

3. Use of all the data sets analyzed up to this point to make a comparison of results between the main implemented methods, and eventually some others, which is the aim of objective 1. This is based on quantitative and qualitative comparison and visualization methods, such as those explained by Thallinger [24] and Martin [25], going from simple Venn diagrams to more complex, network analysis, software such as some specific R packages [20] or Cytoscape [26]. The focus here is to use graphical visualization elements to compare the results of the analyses with and without the addition of biological information.

4. Development of new methods and/or workflows in order to improve and/or combine the benefits from the selected approaches, with focus in those allowing the addition of biological significance to the integration process. Here follows an overview of the methods developed to expand the original datasets (X, Y) with annotations (Ax, Ay) to obtain new blocks of data (Nx, Ny,and Nxy). And the workflow has been implemented adapting the integrative pipelines applied so far to the R targets package [33], a pipeline toolkit that improves reproducibility, skipping unnecessary steps already up to date and showing tangible evidence that the results match the underlying code and data. The development of this targets workflow is intended to comply with the objective number 2 of this working plan.

5. Implementation of the methods resulting from 4) as a new R package to be submitted to Bioconductor repository [27], and, finally, to complete objective 4 of this thesis plan, as a web application [28] to be used in further steps of the current biomedical research projects in which our collaborators are implied, as well as in future studies.

1.2 Method's overview

The addition of biological annotations to the data sets being integrated, prior to the integrative analysis itself, can be useful to improve the integration/analysis outcomes as well as their biological interpretability.

Passos principals explicats aqui:

A. Pre process omics datasets in order to include biological information before the joint analysis –> Expanded datasets

B. Analysis of the expanded datasets by the use of contrasted joint Dimensionallity Reduction techniques

C. Process semi automation in ease to use tools

Start the process already having a couple [punt de millora: admetre 3 o + inputs] of data sets from distinct 'omics sources, mapped to gene ids (if GO annotation has to be performed), containing the results from a selection of differentially expressed genes or most relevant proteins analysis, or similar.

[explicar aquí els requeriments de format dels data sets d'entrada!!]

For each input data set, if annotations are not already provided, two distinct basic annotation methods can be performed:

(i) a basic GO mapping, returning annotations to those GO entities for which we find more than a certain number of features (gene ids coming from our data set) annotated to them,

[mostrar formula] [mostrar exemple]

(ii) a Gene Enrichment Analysis (based on Hypergeometric tests against all GO categories, with FDR correction[ref clusterProfiler]) is performed in order to retrieve the most relevant annotations to that set of genes/features.

[mostrar exemple de llista de gens]

```
 [1] "RTN2"     "NDRG2"    "CCDC113"  "FAM63A"   "ACADS"    "GMDS"     "HLA.H"    "SEMA4A"   "ETS2"     "LIMD2"    "NME3"
[12] "ZEB1"     "CDCP1"    "GIYD2"    "RTKN2"    "MANSC1"   "TAGLN"    "IFIT3"    "ARL4C"    "HTRA1"    "KIF13B"   "CPPED1"
[23] "SKAP2"    "ASPM"     "KDM4B"    "TBXAS1"   "MT1X"     "MED13L"   "SNORA8"   "RGS1"     "CBX6"     "WWC2"     "TNFRSF12A"
[34] "ZNF552"   "MAPRE2"   "SEMA5A"   "STAT5A"   "FLI1"     "COL15A1"  "C7orf55"  "ASF1B"    "FUT8"     "LASS4"    "SQLE"
[45] "GPC4"     "AKAP12"   "AGL"      "ADAMTS4"  "EPHB3"    "MAP3K1"   "PRNP"     "PROM2"    "SLCO3A1"  "SNHG1"    "PRKCDBP"
[56] "MXI1"     "CSF1R"    "TANC2"    "SLC19A2"  "RHOU"     "C4orf34"  "LRIG1"    "DOCK8"    "BOC"      "C11orf52"  "S100A16"
[67] "NRARP"    "TTC23"    "TBC1D4"   "DEPDC6"   "ILDR1"    "SDC1"     "STC2"     "DTWD2"    "TCF4"     "ITPR2"    "DPYD"
[78] "NME1"     "EGLN3"    "CD302"    "AHR"      "LAPTM4B"  "OCLN"     "HIST1H2BK" "HDAC11"  "C18orf1"  "C6orf192"  "AMPD3"
[89] "COL6A1"   "RAB3IL1"  "APBB1IP"  "PSIP1"    "EIF2AK2"  "CSRP2"    "EIF4EBP3" "LYN"      "WDR76"    "SAMD9L"   "ASPH"
[100] "RBL1"    "SLC43A3"  "HN1"      "TTC39A"   "MTL5"     "NES"      "APOD"     "RIN3"     "ALCAM"    "C1orf38"  "PLCD3"
[111] "BSPRY"   "NTN4"     "IL1R1"    "EMP3"     "ZKSCAN1"  "FMNL2"    "OGFRL1"   "IRF5"     "IGSF3"    "DBP"      "CNN2"
[122] "CAMK2D"  "SIGIRR"   "AKAP9"    "ICA1"     "FGD5"     "DSG2"     "E2F1"     "QSOX1"    "TOB1"     "CSF3R"    "SHROOM3"
[133] "CCDC80"  "FRMD6"    "CXCL12"   "CCNA2"    "TIGD5"    "ALDH6A1"  "POSTN"    "FZD4"     "NCAPG2"   "SDC4"     "SNED1"
[144] "PLEKHA4" "KCNAB2"   "SH3KBP1"  "IGSF9"    "DNLZ"     "S1PR3"    "PTPRE"    "FLJ23867" "PLSCR1"   "LMO4"     "IFITM2"
[155] "LRRC25"  "TST"      "NCF4"     "NCOA7"    "IL4R"     "CCDC64B"  "SGPP1"    "RUNX3"    "SLC5A6"   "IFIH1"    "PREX1"
[166] "PLAUR"   "CDK18"    "SLC43A2"  "GK"       "ICAM2"    "YPEL2"    "CBR1"     "MEX3A"    "ZNRF3"    "PTPRM"    "C1orf162"
[177] "GAS6"    "C1QB"     "PVRL4"    "CTSK"     "MRVI1"    "LEF1"     "PLCD4"    "ZNF37B"   "MEGF9"    "GINS2"    "FAM13A"
[188] "CPT1A"   "SNX10"    "TRIM45"   "ELP2"     "ALOX5"    "AMN1"     "CERCAM"   "SEMA3C"   "KRT8"     "TP53INP2"  "JAM3"
[199] "ZNF680"  "PBX1"
```

Figure 1.1: List of gene symbols used as example

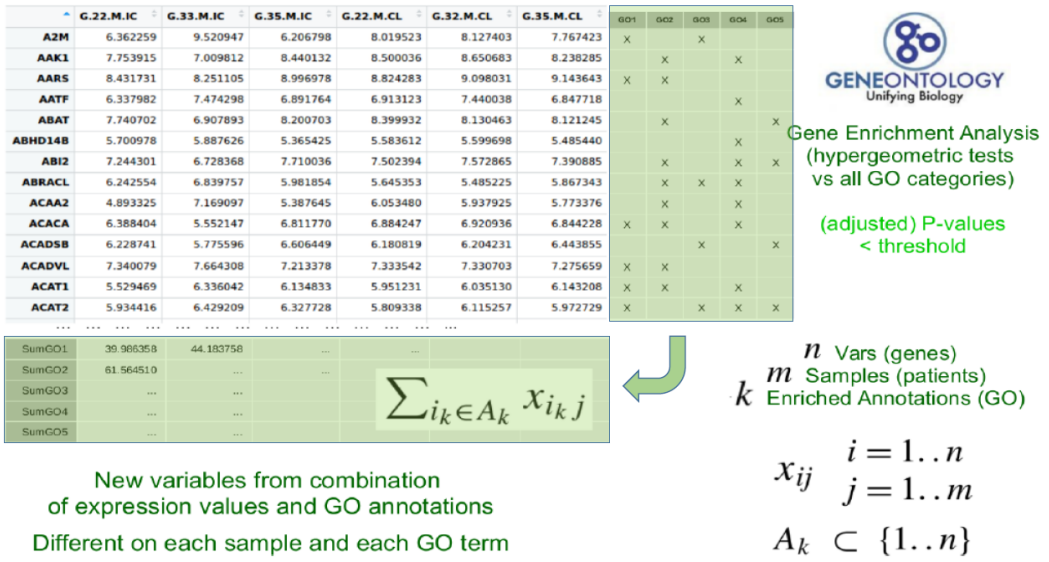

Figure 1.2: Addition of GO terms

[afegir aquí la opció d'afegir les anotacions com a individus suplementaris enlloc de variables]

Figure 1.2 is an example.

Alternatively, manual annotations can be provided (eg. GO terms, canonical pathways, or even annotation to custom entities) as an optional input file.

[mostrar el format requerit].

Other annotation methods can be implemented, as functions to be used by the main pipeline, if more complex methods for biological information addition are required.

[Mostrar el format final de les anotacions, com a matrius dels data sets amb anotacions binàries 1/0 com a columnes extra]

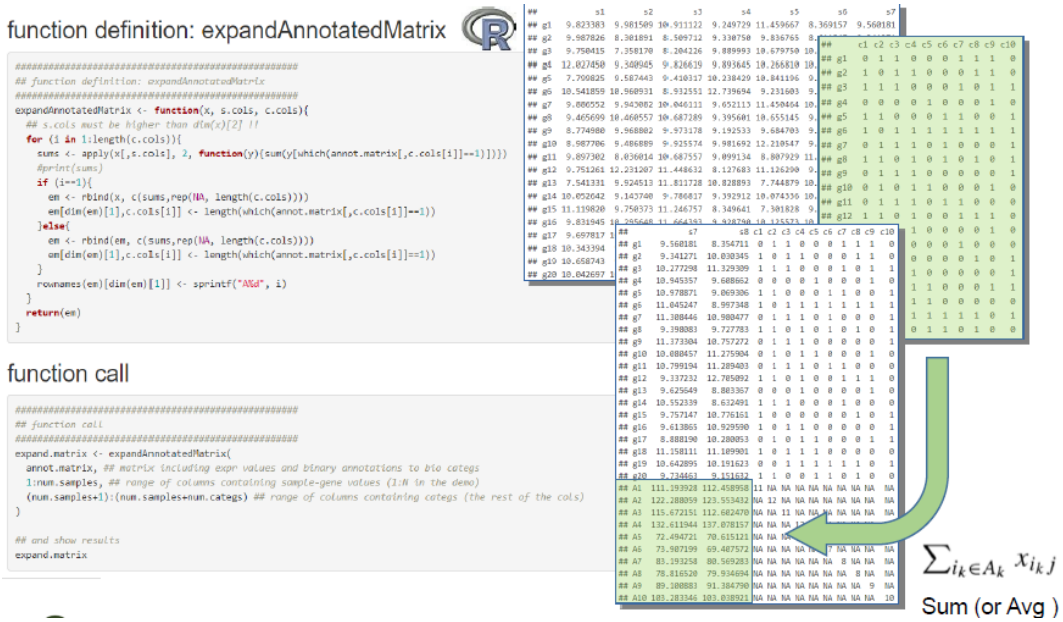

Figure 1.3: Addition of news feats

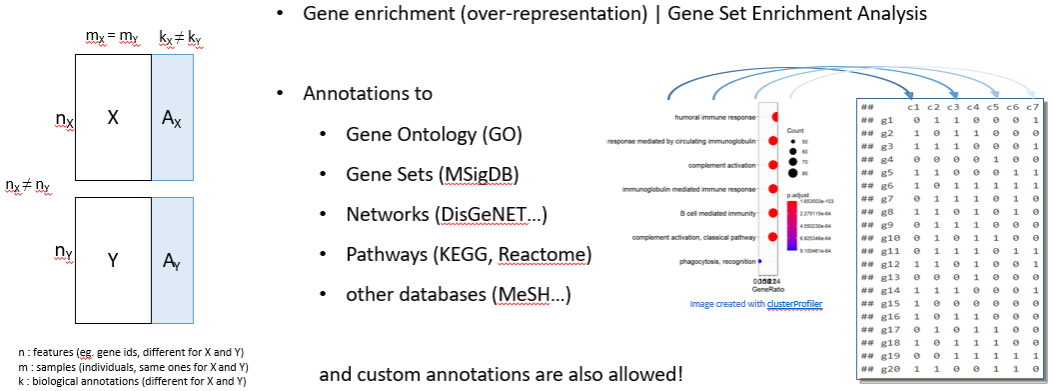

Figure 1.4: Gene enrichment diagram

Once the annotations are already computed, mapping each feature of the input data set to the corresponding biological entity, they can be used to generate new features (as new rows), computing the average value [punt de millora: funció de ponderació] of the expression/intensity values from all original features being mapped to the annotated biological entities.

Once we have the annotated matrices (Figure 1.5, highlighted in blue) we proceed to generate the Expanded matrices (in green) by casting these annotations as numerical values, that is, calculating the average of the numerical expressions of each individual for the variables annotated to each category. This is done with the matrix

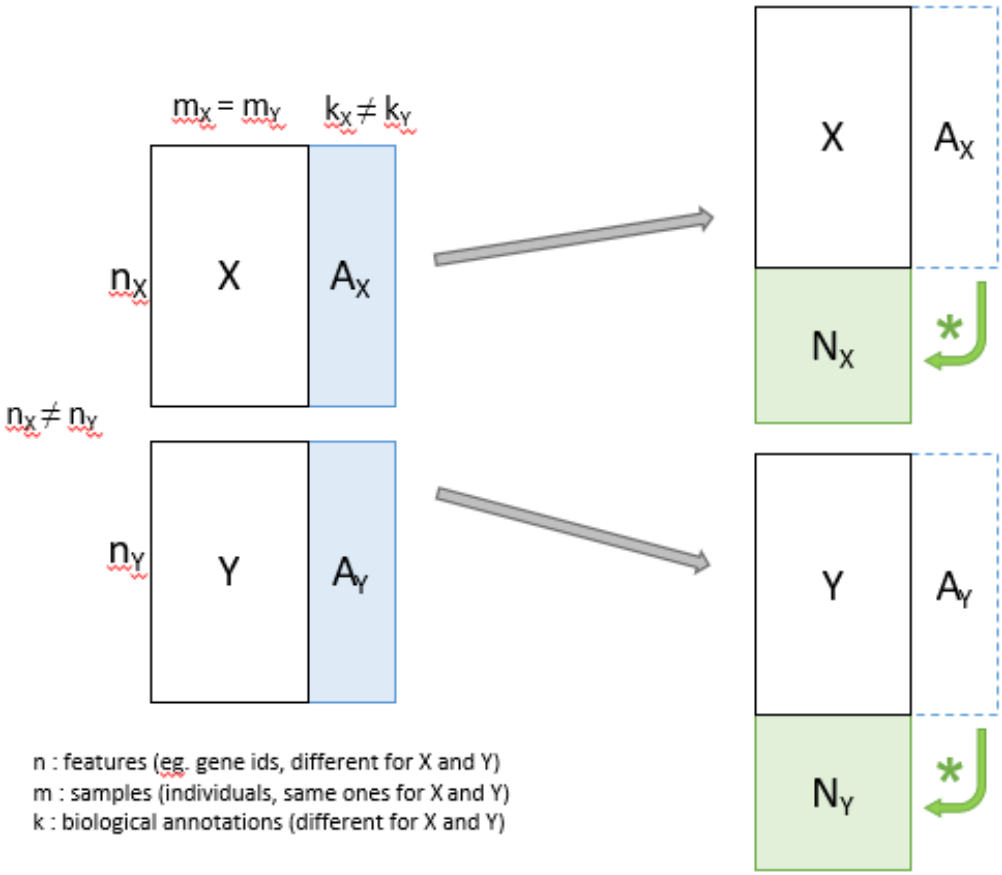

Figure 1.5: Matrix expansion diagram

product of the initial numerical values (expression, proteins...) with the transposed matrices of their annotations, and then with the inverse matrix of a diagonal matrix of the count of how many annotations each category or entity annotated has had.

1.2.1 Detail of the integrative data analyses applied...

Mètodes:

1- Significació biològica, com faig les anotacions

2- Expansió de les matrius (creació de noves vars a partir de les anotacions)

3- Anàlisi factorial en detall, + MCIA + RGCCA

4- TFM sobre workflows i automatització -> Paquet targets en general

$$N_G = \phi(G, A_G), \quad G \in X, Y, \ldots,$$

$$N_X = \frac{1}{r_X}(X \times A'_X),$$

$$N_Y = \frac{1}{r_Y}(Y \times A'_Y),$$

Figure 1.6: Addition of new feats (2)

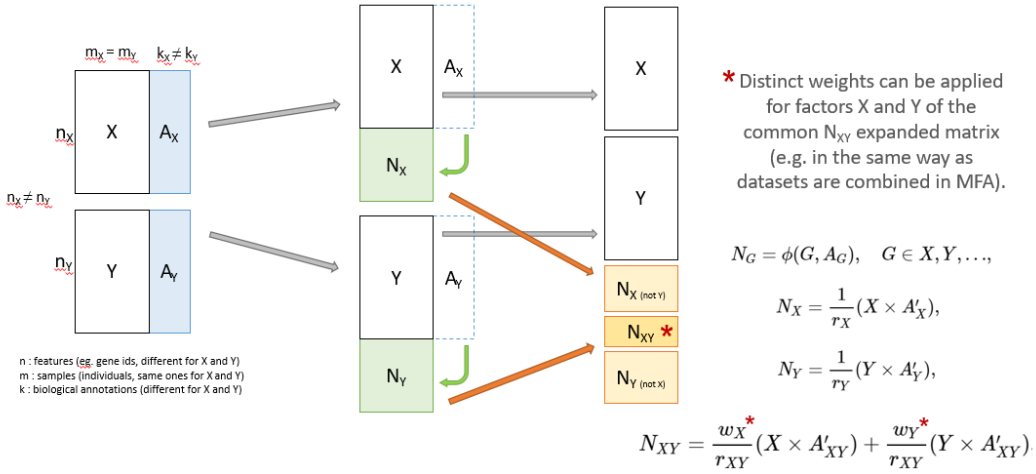

Figure 1.7: Matrix expansion diagram (2)

1.3 Comparison of ODA results

1.3.1 Numeric measurement

% variabilitat explicat segons la estructura de la intersecció de les 2 taules

1.4 Biological interpretation

1.5 Targets PIPELINE

R package creation...

Sistema que hem aplicat per crear el pipeline amb Targets...

The R 'targets' package is a powerful tool for building and managing data science and data analysis pipelines. It is primarily designed for workflow automation, dependency management, and parallel processing in R projects. This package is useful for the following purposes:

1. Define and Manage Workflows: You can create a directed acyclic graph (DAG) that represents the workflow of your data analysis or machine learning project. Each node in the graph corresponds to a target, which can be a data file, an R script, or any other computational task.

2. Manage Dependencies: 'targets' allows you to specify dependencies between targets, ensuring that tasks are executed in the correct order. If a target depends on another target, it won't be executed until its dependencies are up-to-date.

3. Parallel Processing: One of the strengths of 'targets' is its ability to parallelize tasks. It can automatically determine which targets can be executed concurrently, improving the efficiency of your workflows, especially when working with large datasets or computationally intensive tasks.

4. Incremental Builds: When you make changes to your code or data, 'targets' can identify the minimal set of targets that need to be recomputed, saving time and computational resources. This is particularly useful for iterative development and experimentation.

5. Reports and Logging: 'targets' provides tools for generating reports and logging the progress of your workflow, making it easier to track and document your work.

6. Integration: It can be seamlessly integrated with other R packages and tools, such as 'drake' for more advanced data workflow management.

So, the 'targets' package is especially valuable for projects where data processing is a significant component, and you need a structured way to manage the various steps of your analysis or modeling pipeline. It helps ensure that your analyses are reproducible, efficient, and well-documented.

Targets workflow diagram (Figure 1.8) showing the steps corresponding with the complete process: The pipeline starts from (A) a couple of 'omics-derived input data sets (e.g. pre-processed gene expression and protein abundance matrices). These are converted to R data frames with features in rows and samples in columns. Then, a data frame containing related annotations (B) is created, or loaded, for each given input matrix, and used to expand these original data, in order to end up with a pair of data frames (C) containing the original values plus the average

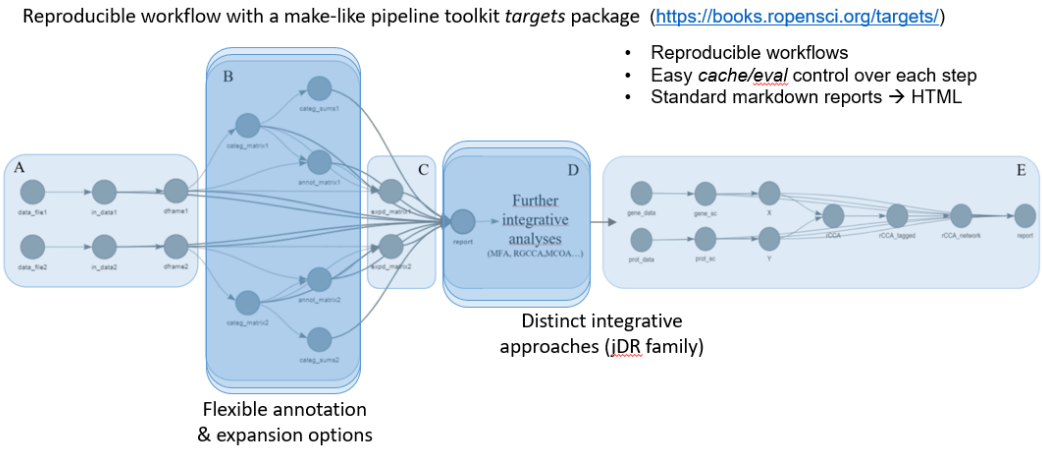

Figure 1.8: Workflow overview

expression/abundance values of the features related to each annotation as new features in additional rows. After that, distinct Dimension Reduction Methods are applied to perform the integrative analysis (D), and finally, an R markdown report (E) is rendered to show steps and main results of the full process.