UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Integrated cluster analysis of high dimensional data

Anastasia Volkova

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
Prof. dr. Ziv SHKEDY

CO-SUPERVISOR :
De heer Bernard OSANGIR

2023
2024

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Integrated cluster analysis of high dimensional data

Anastasia Volkova

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

CO-SUPERVISOR :

De heer Bernard OSANGIR

Abstract

Multi-omics data analysis have received much attention in recent years due to their potential for use in clinical practice, primarily for the development of personalized medicine. One of possible ways of processing multi-omics data is integrative clustering - finding coherent groups of samples based on data of different types. This study focuses on application of four approaches of integrative clustering: ADC, multi-source ABC, CEC and weighted similarity-based clustering. The purpose of the study is to illustrate the application of methods on simulated data with a known and pronounced cluster structure, and also to apply methods to search for structure in real experimental data. In the latter case, an integrative analysis of two types of omics data was performed - transcriptomic and proteomic data. The ability of the selected methods to determine the true cluster structure in simulation data is demonstrated. Using real data as an example, multi-source ABC and weighted similarity-based clustering methods were able to detect a biologically meaningful structure that was not found during single-source analysis.

Keywords: *integrative clustering, weighted similarity-based clustering, CEC, ADC, multi-source ABC, hierarchical clustering.*

List of Tables

List of Figures

Contents

1 Introduction

Advances of recent years in the quantitative analysis of biological samples, such as sequencing and mass spectrometry, have enabled researchers to collect high dimensional omics data of different types from the same set of biological samples. The number of the molecular features in high dimensional omics data is much higher than the sample number. Such data have enormous potential for clinical use and therefore are subject to comprehensive analysis. Multi-omics data include the data generated from genome, proteome, transcriptome, metabolome and epigenome, as well as lipidome, phosphoproteome and glycol-proteome analysis [1]. Macro- and micro molecules, such as genes, RNA, proteins, lipids and other metabolites, collaborate in cells to perform complex biological processes, therefore integrative analysis of multiple omics data types may reveal systems-level insights.

Integrative analysis of clinical or biological data can facilitate more personalized treatment of patients. Analyzing multi-omics profiles can help to address such questions as disease subtyping and classification, identification of diagnostic biomarkers and driver genes for diseases. Such a heterogeneous disease as cancer is the most common object of study using an integrative multi-omics approach. This is due to the fact that different subtypes of cancer, on the one hand, are characterized by different survival rates and prognosis of response to therapy, and on the other hand, they differ significantly from each other in the molecular processes occurring in the cancer cell. For example, four subtypes of Glioblastoma with different survival outcomes were identified with help of the Pathway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM), using gene expression and copy number data [2]. An entirely new subtype of prostate cancer with extremely poor survival outcome was identified with usage of a nonparametric Bayesian model for integrating patient-specific gene expression and copy number variation data [3]. The given examples of clinically significant discoveries, which were made using integrative analysis of multi-omics data, indicate the importance of improving integrated analysis methods and finding new, more reliable and effective ones.

One of the ways to analyze multiple-source high dimentional data is clustering - finding coherent groups of samples in the data, such that samples within a group are similar, and samples in different groups are dissimilar. Clustering is a highly researched computational problem, investigated by multiple scientific communities [4]. Integrative cluster analysis is especially relevant for processing multi-omics data: it can help to reduce the effect of experimental and biological noise in the data and to reveal different cellular aspects, such as effects manifest at the genomic and epigenomic levels [4].

The simplest way of searching for clusters in multi-omics data is concatenating all omics matrices and application of single-omics clustering to the resulting matrix. This method increases the dimensionality of the data, therefore low-rank approximation of the resulting matrix can be applicable [5]. This approach is often termed as the early integration. The low-rank assumption is used for low-dimensional representation of the original data, which is used to identify clusters of samples.

In late integration, each omics is clustered separately using a single-source clustering algorithm and the clustering solutions are integrated to obtain a single clustering solution. This approach is also called ensemble clustering. The advantage of late integration lies in that any clustering algorithm can be used for each omics, which makes possible usage of algorithms that are known to work well on a particular omics only. Method of consensus clustering is widely used to combine multiple clustering solutions. This method uses different resampling schemes to simulate perturbations of the original data set, so as to assess the stability of the clustering results with respect to sampling variability. The underlying assumption is that the more stable the results are with respect to the simulated perturbations, the more these results are to be trusted [6]. The clustering algorithm of choice is then applied to each of the perturbed data sets from each data source, and the consensus matrix is built, which stores, for each pair of samples, the proportion of clustering runs in which two items are clustered together. The consensus matrix can be used as a visualization tool to help assess the composition and number of clusters [6].

Consensus clustering is used in such multi-source clustering technique as aggregating bundles of clusters [7], which implies resampling from each data set separately and usage of resulting clustering solutions to build consensus matrix. On the other hand, cluster of cluster assignments (COCA) methodology computes a binary matrix of n by c dimensionality, where n is a number of samples and c is a sum of numbers of clusters in each omics data. This binary matrix, indicating whether a sample belong to a certain cluster in a certain data source, is then used as an input into consensus clustering [8].

A special class of multi-omics data clustering algorithms is the class of similarity based methods. The shortcoming of the late integration approach, described above, is that due to clustering of samples using data sets separately, concordant but weak signals may be vanished during the initial clustering. Similarity-based approach implies data integration before clustering procedure, which helps to make use of weak but concordant structures in different data sources.

Technology of the similarity network fusion (SNF) performs integration of data of different type by constructing networks of samples or patients for each available data type and then fusing these into one network that represents the full spectrum of underlying data. Fusion is performed by iterative updating these networks to increase their similarity until they converge to a single network, which is then partitioned using clustering algorithm [9]. The method of Neighborhood based Multi-Omics clustering (NEMO) implies building of inter-sample relative similarity matrix for each omics data set, which expresses the similarity between samples i and j relative to i's k nearest neighbors and to j's k nearest neighbors [10]. Since different omics have different data distributions, the relative similarity is more comparable between omics than the original similarity matrix. As in SNF approach, all relative similarity matrix are integrated into one matrix and that network is clustered.

Similarity-based clustering gives possibility to weight the importance of the different types of information. The weighted similarity-based approach was developed to solve the problem of discovering compounds that are similar with respect to structure and induced activity [11]. The integrative clustering of compounds uses the weighted average of similarity matrices as an input.

Integration of data at similarity level also can be accompanied by dimensionality reduction. For example, regularized multiple kernel learning for dimensionality reduction (rMKL-LPP) method uses dimensionality reduction, such that similarities between neighboring samples is maintained in lower dimension [12]. By this approach samples are projected into a lower dimensional, integrated subspace where they can be analyzed further. Group of similarity-based methods, together with methods that use statistical modeling of the data, is often referred to as intermediate integration.

Figure 1 schematically illustrates the multi-source data clustering approaches, described above.

1.1 Research question and organization of the report

The aim of the study is to illustrate application of methods of multi-source clustering to two omics data sets: transcriptomic and proteomic data from the radiation experiment. In order to investigate the performance of the multi-source clustering algorithms on the data with a well-known and pronounced clustering structure, the simulated data were used. In order to compare performance of methods of different types with each other, the work considers early integration (aggregated data clustering), intermediate integration (weighted similarity-based clustering), late integration (integration of clustering solutions using complementary ensemble clustering) and late integration with usage of consensus clustering (integration of clustering solutions using aggregated bundles of clusters) methods.

The report is organized as follows: section 2 is devoted to description of data for both experimental and simulated data sets. Section 3 provides details about methodology, which was used in the project. Sections 4 and 5 presents the results obtained for both real and simulated data. A brief discussion on software used in the project is given in section 6. Societal relevance, ethical thinking and stakeholder awareness are subjects of discussion in section 7. Section 8 includes discussion of results obtained, notes about limitation of the research, description of prospects for further research and the conclusion.

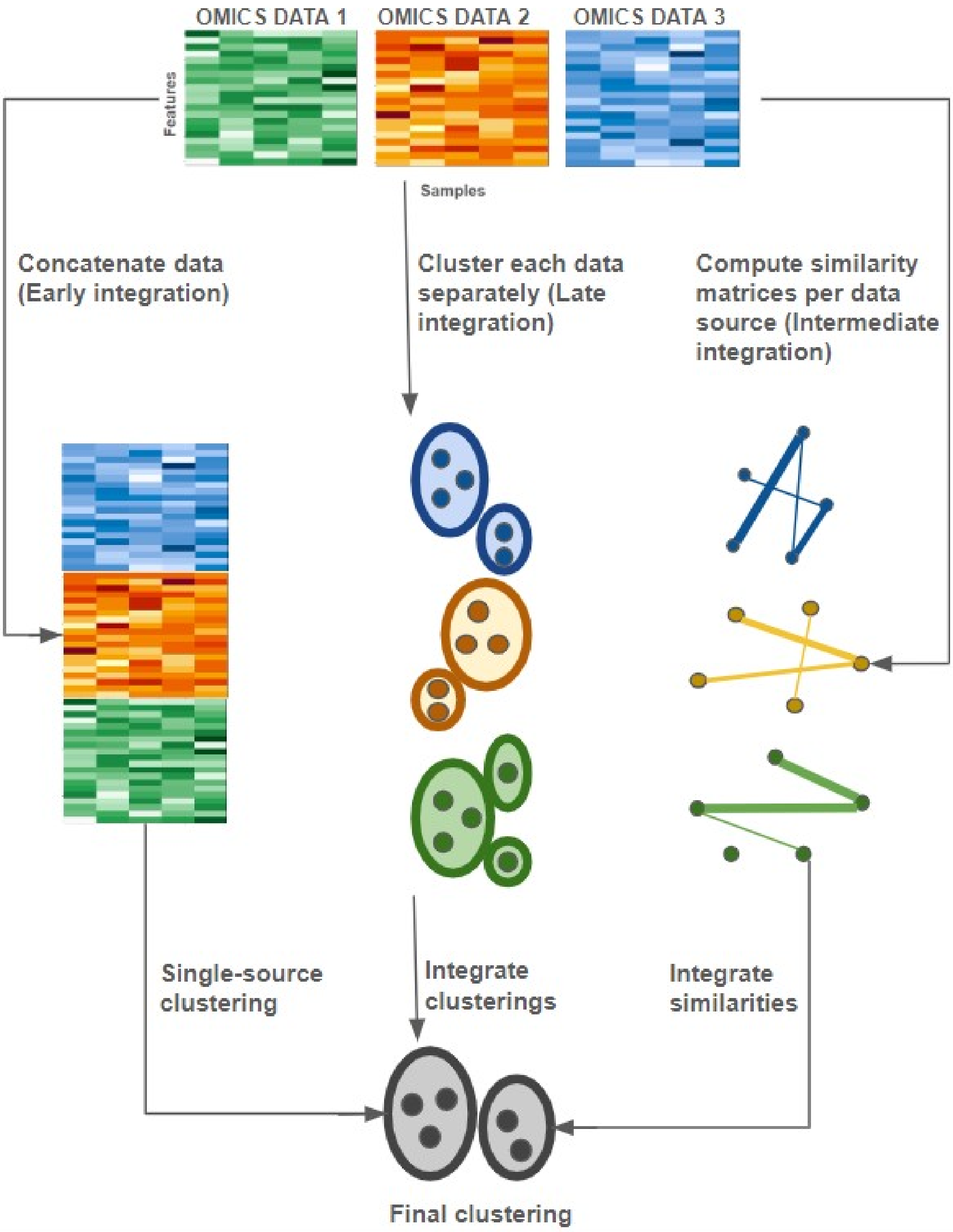

OMICS DATA 1
OMICS DATA 2
OMICS DATA 3
Features
Samples
Concatenate data (Early integration)
Cluster each data separately (Late integration)
Compute similarity matrices per data source (Intermediate integration)
Single-source clustering
Integrate clusterings
Integrate similarities
Final clustering

Figure 1: Framework of multi-source data clustering.

2 Data

Two sources of data were used for the analysis in this research project. First source is a simulated data, which were generated specifically for the purpose of illustration and investigation of the performance of clustering techniques. Second source of data is a radiation experiment in which newborn male mice were exposed to ionizing radiation.

2.1 Simulated data

Two data matrices were created, matrix X and matrix Y. Both matrices have 30 columns, corresponding to 30 samples. Number of rows, corresponding to the number of observations, is equal to 1000 in matrix X and 3000 in matrix Y. In the context of multi-omics data, rows may correspond to biological features such as genes, proteins, etc. In matrix X observations are correlated such that 30 samples form 6 groups of 5 samples in each group. In matrix Y samples form 3 groups including 15, 10 and 5 samples. Each group of samples was generated in the following way:

$$G_i \sim N_m(\mu, \Sigma), \tag{1}$$

where $N_m(\mu, \Sigma)$ denotes the multivariate normal distribution, G_i denotes a matrix, containing values of features for all samples in group i, m is a number of samples in the group i, μ is a vector of zeroes of length m, Σ is a covariance matrix with m rows and m columns:

$$\Sigma = \begin{bmatrix} 1 & 0.75 & 0.75 & 0.75 & 0.75 \\ 0.75 & 1 & 0.75 & 0.75 & 0.75 \\ 0.75 & 0.75 & 1 & 0.75 & 0.75 \\ 0.75 & 0.75 & 0.75 & 1 & 0.75 \\ 0/75 & 0.75 & 0.75 & 0.75 & 1 \end{bmatrix}. \tag{2}$$

Structure of matrices X and Y is shown in equations 3 and 4.

$$X = \begin{bmatrix} x_{1,1} & \cdot & x_{1,30} \\ . & . & . \\ . & x_{ij} & . \\ . & . & . \\ x_{1000,1} & \cdot & x_{1000,30} \end{bmatrix}. \tag{3}$$

$$Y = \begin{bmatrix} y_{1,1} & \cdot & y_{1,30} \\ . & . & . \\ . & y_{ij} & . \\ . & . & . \\ y_{3000,1} & \cdot & y_{3000,30} \end{bmatrix}. \tag{4}$$

Matrix X was generated using following algorithm:

1. Create covariance matrix with 5 rows and 5 columns, value of 1 on diagonal and 0.75 in all other cells.
2. Create a vector of zeros of length 5. This vector will be further used as a vector of mean values.
3. Sample 1000 vectors of random values from multivariate normal distribution with a vector of mean values from p. 2 and correlation matrix from p. 1. This results in a matrix of 1000 rows and 5 columns.
4. Repeat previous action 5 times, which gives 5 different matrices of the same dimensionality and with same correlation between columns.
5. Combine six matrices into one by columns.

Matrix Y was generated using the same algorithm, except that it requires separate generation of three vectors of means and three correlation matrices of different dimensionality. In matrix Y, the first 15 columns represent

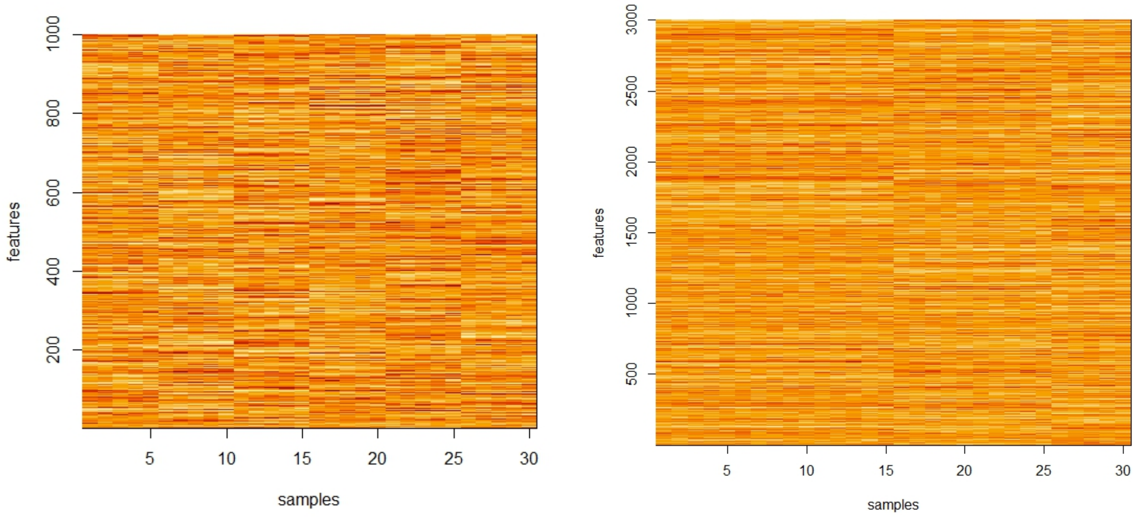


Figure 2: Color scheme of simulated data: matrix X (left) and matrix Y (right).

samples of the first cluster, columns from 16 to 25 - second cluster, and the last five columns represent samples belonging to the last cluster.

Characteristics of matrices X and Y are summarized in Table 1. Code for data simulation can be found in Appendix.

Table 1: Characteristics of simulated data matrices.

	Rows	Columns	Groups of clustered samples	Coefficient of correlation
Matrix X	1000	30	6	0.75
Matrix Y	3000	30	3	0.75

Generated matrices X and Y are shown schematically on Figure 2.

Similarity matrices of X and Y are visualized on Figure 3. In this setting, similarity matrix is a square matrix with 30 rows and 30 columns, representing 30 samples in the data sets. A score at intersection of the ith row and the jth column expresses the similarity between samples i and j (details are explained in the Methodology section). On Figure 3 one can clearly observe grouping of samples, whereby samples from the same group are characterized with higher similarity. Six non-overlapping blocks on the left heat map on Figure 3 represent six clusters in matrix X and three blocks on the right heat map correspond to three clusters in matrix Y.

2.2 Radiation data

Experimental data, used in the project, were generated by the Belgian Nuclear Research Centre [13] in a multi omics radiation study in which newborn male mice were exposed to 2 Gy Gamma radiation immediately after their birth. Mothers of these mice were treated with two different diets: a rich in folic acid (FA) diet and a standard diet. Tissues of brain of mice, such as cortex and hippocampus, were extracted at day 11 and day 17 after exposure.

Distribution of animals between groups regarding day of tissue selection, radiation exposure status and diet of mother is shown in the Table 2. Mothers of animals in groups $FA+, Radiation+$ and $FA+, Radiation-$ received the folic acid-rich diet, while mothers of animals in groups $FA-, Radiation+$ and $FA-, Radiation-$ did not. Animals in groups $FA+, Radiation+$ and $FA-, Radiation+$ were exposed to ionizing radiation, while animals in groups $FA+, Radiation-$ and $FA-, Radiation-$ were not. In total there were 31 samples of cortex and 32 samples of hippocampus collected.

Each sample was a subject to transcriptomic and proteomic analysis. Cortex data include information on 3102 proteins and 18380 genes. Hippocampus data include information on 3011 proteins and 18329 genes.

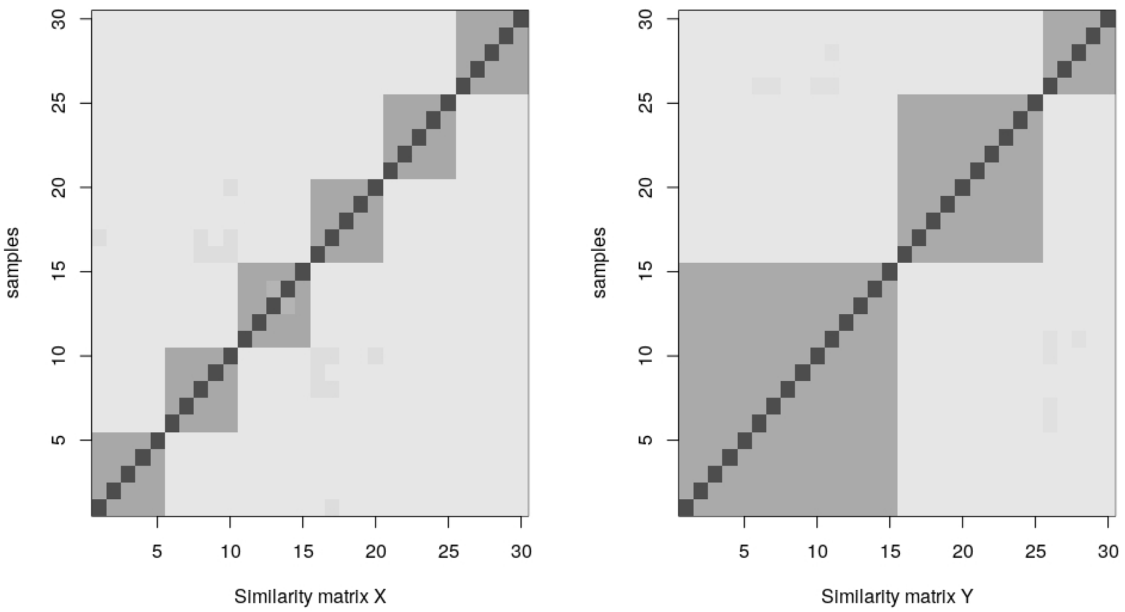


Figure 3: Color-coded heat maps of similarity matrices X (left) and Y (right).

Table 2: Distribution of animals between groups.

	Day 11	Day 17	Day 11	Day 17
	Cortex		Hippocampus	
$FA+, Radiation+$	4	4	4	4
$FA+, Radiation-$	3	4	4	4
$FA-, Radiation+$	4	5	4	5
$FA-, Radiation-$	4	3	4	3

The data are stored in a matrix format with rows indicating the features and columns indicating the samples. Structure of data is shown in equation 5. Let $X_{k,ij}$ denote the measurement of the jth feature, $j = 1, ..., m_k$, for the ith sample, $i = 1, ..., n$, where n and m denote number of samples and number of features respectively, k=1 for proteins data and k=2 for gene expression data.

$$X_k = \begin{bmatrix} x_{k,11} & \cdot & x_{k,1n} \\ \cdot & \cdot & \cdot \\ \cdot & x_{k,ij} & \cdot \\ \cdot & \cdot & \cdot \\ x_{k,m_k 1} & \cdot & x_{k,m_k n} \end{bmatrix}. \tag{5}$$

Different days of tissue collection, type of analysis and region of brain are given in separate matrices, which results in 8 data files. Information on treatment status of mothers and exposure of animals to radiation is given in 4 (separately for two days and two brain regions) additional data matrices. For illustrative purposes, only result of analysis for two settings are presented in this report: cortex data, day 11 and hippocampus data, day 11.

Similarity matrices of radiation data of two experimental settings - cortex, day 11 and hippocampus, day 11 - are presented on Figures 4 and 5 respectively. It can be clearly observed, that, in contrast with heat maps for simulated data (Figure 3), there is no obvious grouping of samples and the signal in general is much weaker.

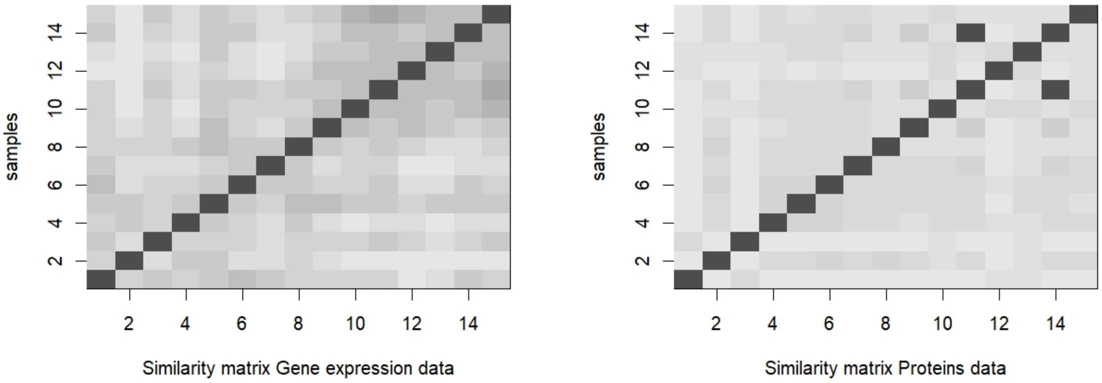


Figure 4: Color-coded heat maps of similarity matrices of gene expression (left) and proteins (right) data, cortex, day 11.

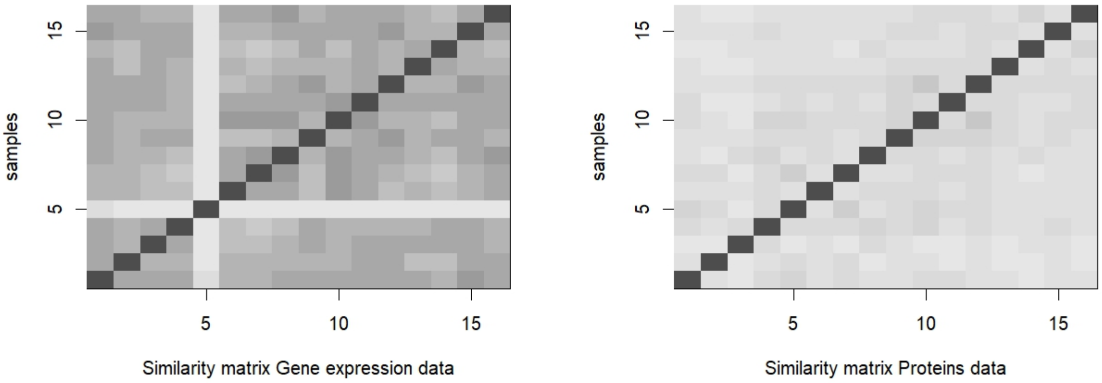


Figure 5: Color-coded heat maps of similarity matrices of gene expression (left) and proteins (right) data, hippocampus, day 11.

3 Methodology

The aim of integrative clustering is to cluster samples using two or more data sources. As a reference, single-source clustering was also performed in this project. The simplest way of integrative clustering, applied to radiation data, is the clustering on merged data (aggregated data clustering). Ensemble clustering approach, which implies aggregation of single-source clustering solutions in order to produce an overall clustering structure, was also used (methods of ABC, multi-source ABC and complementary ensemble clustering). Method of weighted similarity-based clustering combines data of two or more sources at the level of similarity matrices and assign a weight to each similarity matrix (corresponding to each data source). Hierarchical clustering is used in this work to find groups in the data, due to its popularity and advantages, described in further sections. This method results in a dendrogram which can be cut at any point to obtain a number of clusters in a range from 1 to N, where N is a number of samples in the data. Optimal number of clusters was defined using the 'gap' method, based on minimizing the within-cluster dispersion.

3.1 Cluster analysis of single-source data

3.1.1 Similarity matrix

Similarity matrix is a matrix which contains information about pairwise similarity between observations/samples [14]. Cells of matrix contain scores, measuring how similar a pair of samples are to each other. Number of rows and columns in similarity matrix is equal to number of observations. The smaller the value of the score, the greater the similarity. Euclidean distance was used in this research project as a measure of similarity of samples. Euclidean distance between observations or samples i and i', for which p features are measured, is defined as follows [14]:

$$d_{ii'} = \sqrt{\Sigma_j (x_{ij} - x_{i'j})^2}, \tag{6}$$

where $d_{ii'}$ is the Euclidean distance between samples i and i' and j=1,2,...,p. The values of Euclidean distance between ith and i'th samples are located in the similarity matrix at the intersection of the ith column and the i'th row.

3.1.2 Hierarchical clustering

Hierarchical clustering is a technique for grouping observations or samples based on the pairwise similarities between them [14]. Agglomerative hierarchical clustering is a hierarchical clustering procedure, which starts from considering each sample as its own cluster. At each step of the procedure the pair of most similar clusters is defined and these two clusters are fused. After that the pairwise similarities among remaining clusters are calculated. The algorithm is repeated until only one cluster remains. Similarity of clusters in this project was measured by the average linkage: in clusters A and B the average of all pairwise similarities was calculated and these values were used to assess similarity between clusters A and B.

One way to present results of Hierarchical clustering is a dendrogram —a graphical representation of clustering solution, depicted as an upside-down tree. An example of a dendrogram is shown on Figure 6, which illustrates the result of cluster analysis of cortex gene expression data, collected at day 11. Samples are indicated by four colors on the Figure, which corresponds to four clusters in the data. All samples, belonging to mice whose mothers received folic acid, are grouped into cluster (colored in blue) and thus separated from other animals. On a dendrogram each observation is represented as a leaf. Leaves fuse with each other such that there are $N-1$ points of fusion, where N is a number of observations. By agglomerative hierarchical clustering, observations group together starting from the bottom of the dendrogram and points of fusion are located at different heights. The height of fusion indicates how similar clusters or observations are. The most similar samples fuse at the lowest part of dendrogram, whereas clusters or samples that fuse close to the top of the dendrogram are less similar.

Number of clusters on a dendrogram is defined by height, at which horizontal cut is being made. 3-clusters and 6-clusters structure of matrix X are shown on Figure 7. It is observed that samples which are grouped together in 6-cluster structure remain in one cluster when K=3. This illustrates the hierarchical assumption:

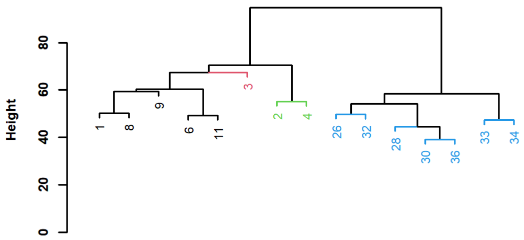


Figure 6: Dendrogram and 4-clusters structure of gene expression analysis data for cortex samples, collected at day 11.

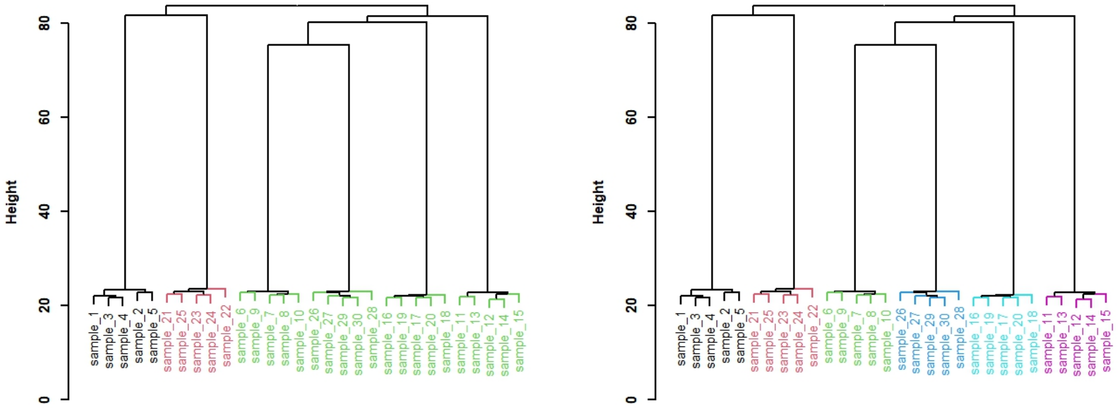


Figure 7: 3-clusters and 6-clusters structure of matrix X.

for any values H and h of height of fusion on a dendrogram, such that $h < H$, clusters obtained by cutting a dendrogram at height h are always nested within clusters, obtained by cutting at H.

3.1.3 Single-source ABC clustering

The aggregating bundles of clusters (ABC) is an iterative procedure, developed for analysis of gene expression data [7]. Method is based on aggregating results obtained from an ensemble of randomly resampled data (both samples and features). Assume that data are organized as a matrix $X = [x_{ij}]$, such that $i = 1, ..., N$, and $j = 1, ..., F$ where N is number of samples and F is number of features. A weight ω_j is assigned to jth feature:

$$\omega_j = \frac{1}{R_j + c}, \tag{7}$$

where R_j is the rank of the variance of jth feature and c is such that the $0.01F$ features with the highest variance have a combined probability of 0.2 of being selected. At each of 1000 iterations a random sample of 100 features was selected using weighted random sampling and N samples were selected with replacement, discarding any replicate samples. Selected samples and features form a matrix which is used as an input into the hierarchical clustering procedure with Ward's linkage [15]. For a given value of k on each iteration each selected sample is assigned to a cluster from range $1, ..., k$ and an incidence matrix is set up. Incidence matrix, in this context, is a square matrix where one row and one column correspond to each sample. The entry $I_{ii'}$ is equal to 1 if samples i and i' belong to the same cluster, and 0 otherwise. After 1000 iterations, all incidence matrices are summed up and divided by number of times two objects were selected simultaneously. The resulting similarity matrix is then transformed into dissimilarity matrix, expressing the number of times the samples were not clustered together. This dissimilarity matrix is taken as an input for a clustering algorithm.

Due to weighted random sampling of features, less informative features are less likely to be selected. Informative features are favored to be included into clustering at each step. Randomness of samples selection leads to increasing of diversity of clustering solutions and reduces dependence among clusters defined at different iterations.

3.2 Integrative clustering

In this section we present several methods of integrative multi-source clustering that were applied in this thesis in order to find structure in the data.

3.2.1 Aggregated data clustering

Under this approach two data sets were merged into one by rows with subsequent analysis of the single resulting matrix by applying hierarchical clustering with Euclidean distance and average linkage. The approach is described in [16] and implies computation of ensemble clusters of the merged data. For simplification, the algorithm was modified such that the merged data set itself was used as an input to final clustering procedure, which results in only one clustering solution instead of an ensemble of them. Two main limitations of this approach are increasing of the dimensionality of the data and ignoring of the different distributions of values in different omics.

In case with simulated data, merging of matrices results in a data set with 30 columns and 4000 rows, which represent 30 observations and 4000 features, of which 1000 come from matrix X and 3000 come from matrix Y. In case of radiation experiment, proteomic and gene expression data for the same day (11 or 17) were aggregated together, since these data come from the same animals.

3.2.2 Weighted similarity-based clustering

This method, introduces by Perualila-Nan et al. [11], combines similarity matrices and assigns weights to them, producing a weighted similarity matrix, which can be further a subject to a clustering algorithm. A remarkable feature of this approach is the possibility to change weights gradually, thus observing the effect of each data source on clustering structure.

For M data sources, the weighted similarity matrix S_N^W for N samples is given by

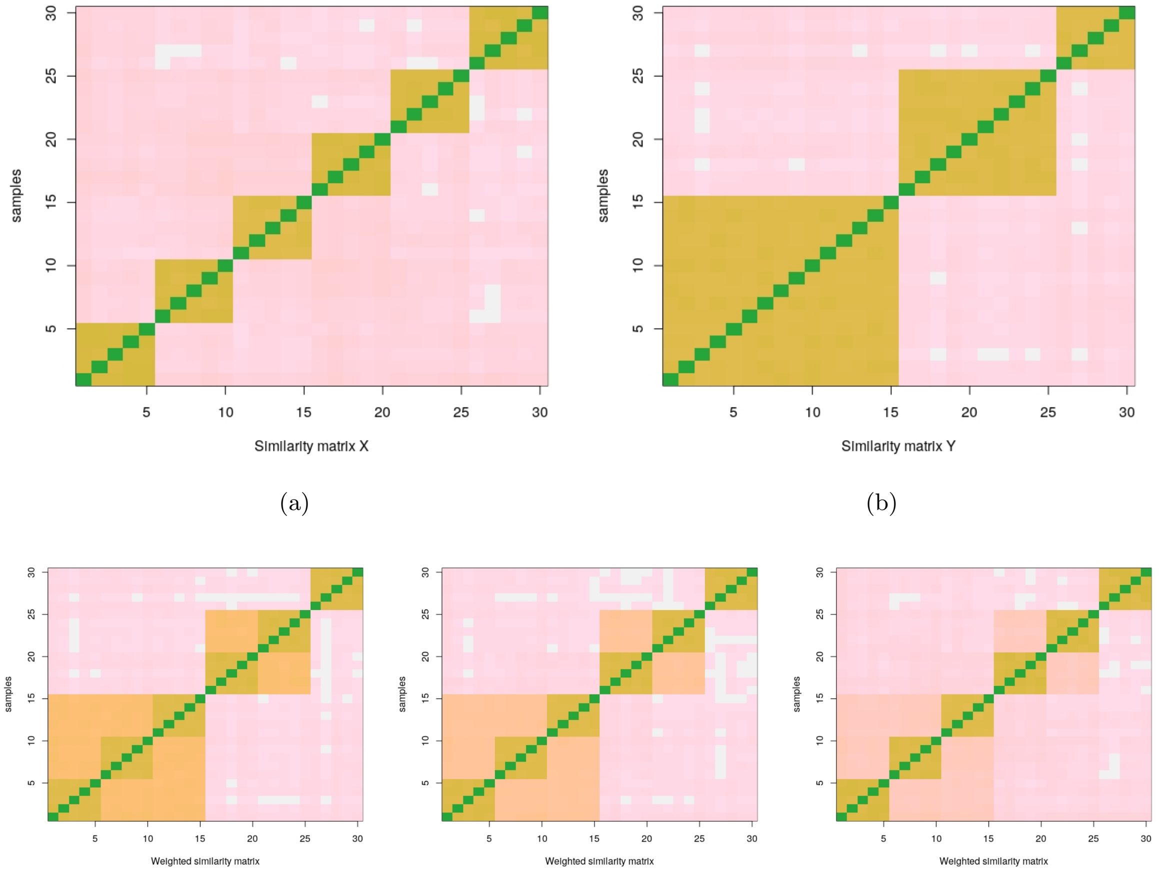
$$S_N^W = \sum_{m=1}^{M} \omega_m S_N^m, \tag{8}$$

$$0 \leq \omega_m \leq 1, \tag{9}$$

$$\sum_{m=1}^{M} \omega_m = 1, \tag{10}$$

where S_N^m is the similarity matrix for the mth data source, $m = 1, ..., M$ and ω_m is the weight associated with the mth source. A weight, assigned to a data source, is a number between 0 and 1. The sum of the weights must be equal to 1. Matrix S_N^W is used as an input for hierarchical clustering procedure.

Weighted summation of similarity matrices is illustrated on Figure 8. Two color-coded heat maps at the top of the Figure 8 (a, b) represent similarity matrices of simulated data sets X and Y. Clustering of samples can be clearly seen: six groups of samples in matrix X and three groups in matrix Y are characterized by a high degree of similarity. Three heat maps at the bottom of Figure 8 (c, d, e) depict a weighted sum of similarity matrices X and Y, with weights of 0.25, 0.5 and 0.75 assigned to similarity matrix X. It can be observed that after summation the similarity between samples which are grouped together both in X and Y remained high. On the other side, similarity of samples which were grouped only in matrix Y depends on a weight, assigned to matrices X and Y. When a weight of similarity matrix Y decreases, similarity between these samples in the weighted similarity matrix also goes down.



(a)

(b)

(c)

(d)

(e)

Figure 8: (a, b) - similarity matrices X and Y. (c, d, e) - summation of similarity matrices X and Y with weight of 0.25 (c), 0.5 (d) and 0.75 (e) assigned to data source X.

In this project only weighted combinations of two matrices were considered, in other words, M in equations 8 and 10 is equal to 2. Equation 8 takes the form:

$$S_N^W = \omega_1 S_N^1 + \omega_2 S_N^2. \tag{11}$$

If $M = 2$, a weight of 0 assigned to the first data source means that the second data source has weight of 1. In this case the second data source is the only one data source which clustering is based on.

In case of radiation experiment data, one of these two matrices contains proteomic data and the other - gene expression data. In case of simulated data, weighted combination of matrices X and Y was considered.

3.2.3 Complementary ensemble clustering

The complementary ensemble clustering approach (CEC) is introduced by Fodeh et al. [16]. In this method clustering algorithm is applied to a weighted linear combination of the coassociation matrices obtained from separate ensemble clustering of different data sources.

Each data matrix first is taken as an input to the iterative hierarchical clustering procedure. Let T denote a number of iterations and $t = 1, 2, ..., T$. T must be specified as a parameter for the clustering procedure, as well as number of clusters K to cut the dendrogram in. Let n denote number of data sources and $i = 1, 2..., n$. For each iteration of the ensemble, the number of features is randomly set between $(m/2)$ and $(m - 1)$, where m is the total number of features that are randomly sampled. At every iteration a binary incidence matrix is set up. All the resulting incidence matrices for data source i are aggregated into the coassociation S_i as follows:

$$S_i^{(t+1)} = S_i^{(t)} + C_i^{(t)} C_i^{(t)T}, \tag{12}$$

where the matrix product $C_i^{(t)} C_i^{(t)T}$ is a binary 0/1 matrix, also called an incidence matrix, that indicates whether a pair of objects belongs to the same cluster during the tth iteration of the ensemble.

The resulting final coassociation matrix S_i again is taken as an input to the hierarchical clustering procedure resulting in a final clustering solution for the data source i.

CEC approach combines the coassociation matrices of ensemble clusters from different data sources into one aggregate coassociation matrix that is subsequently used for obtaining the consensus clusterings. The joint coassociation matrix is computed by adding the coassociation matrices of the different data sources by following way:

$$S_{combined} = \alpha_1 S_1 + \alpha_2 S_2 + ... + \alpha_n S_n, \tag{13}$$

where α is a parameter that governs the weight of each data source. Applying hierarchical clustering algorithm to the combined coassociation matrix $S_{combined}$ yields the final clusters.

Method was applied to integrate proteomic and gene expression data for each day (day 11 or day 17) and for each brain region (cortex or hippocampus).

3.2.4 Multi-source ABC clustering

ABC method, described in section 3.1.3, can be adapted to incorporate two or more data sets [7]. On each of 1000 iterations a subsets including 100 features were selected from each data set to apply clustering algorithm. As was described in section 3.1.3, each iteration for each data source results in an incidence matrix. The difference from single-source ABC method is that these incidence matrices are summed up not only over number of iterations, but also over number of data sources. This results in a single similarity matrix, which, after transforming into dissimilarity matrix, is used as an input into a clustering algorithm. In this project only two sources were used: proteomic and gene expression data for radiation experiment (separately for day 11 and day 17 and for two brain regions), or X and Y matrices for simulated data.

3.3 The Gap statistic

For estimating the number of clusters the 'gap' method was used, as proposed by Tibshirani et al. [17]. Let $d_{ii'}$ denote the Euclidean distance between samples i and i', as shown in equation 6. let also C_r denote the indices of samples in cluster r, such that the sum of the squared pairwise distances for all samples in cluster r is defined as follows:

$$D_r = \sum_{i,i' \in C_r} d_{ii'}^2. \tag{14}$$

Estimate of the optimal number of clusters is the value of k, which maximizes the following statistic:

$$Gap_n(k) = E_n[log(W_k)] - log(W_k), \tag{15}$$

where W_k is the pooled within-cluster sum of squares around the cluster mean and n_r denotes the number of samples in cluster r.

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r. \tag{16}$$

$E_n[log(W_k)]$ denotes expectation of $log(W_k)$ under sample size of n from the reference distribution - a uniform distribution over a box aligned with the principal components of the data. For each value of k 100

Monte Carlo samples were generated from this distribution and for each sample $log(W_k)$ was computed. Average of these values is an estimate of $E[log(W_k)]$. The standard deviation of 100 Monte Carlo replicates of $log(W_k)$ is denoted by sd_k. Preferable number of clusters is equal to smallest k such that

$$Gap(k) \geq Gap(k+1) - s_{k+1}, \tag{17}$$

where

$$s_k = sd_k \sqrt{1 + 1/100}, \tag{18}$$

$$sd_k = \sqrt{\frac{1}{100} \sum_{i=1}^{100} (log(W_k)_i - \overline{log(W_k)})}. \tag{19}$$

4 Application of the methods to the simulated data

Figure 9 illustrates the dendrograms of simulated data matrices X and Y. By simulation, data matrix X includes 6 clusters of samples and data matrix Y includes 3 clusters, which are highlighted by colors on the Figure. It can be observed that samples in each data matrix fuse into clusters at the same height, which implies the same within-cluster similarity of samples.

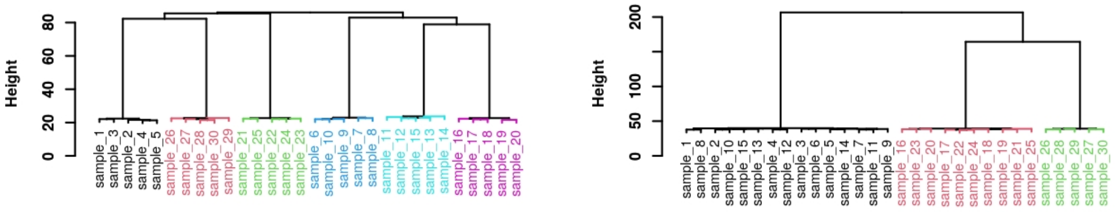


Figure 9: Dendrograms of simulated data matrices X (left) and Y (right).

Application of the 'gap' method to the simulated data resulted in values of optimal number of clusters K, given in Table 3. Value of K for matrices X and Y corresponds to grouping structure of samples, determined by simulation.

Table 3: Definition of the optimal number of clusters in simulated data with the 'gap' method.

Data set	K
Matrix X	6
Matrix Y	3
Merged matrices X and Y	3

4.1 Weighted similarity-based clustering

Formula 8 for weighted similarity matrix S_N^W in this setting takes the form:

$$S_{30}^W = \omega_X S_{30}^X + \omega_Y S_{30}^Y, \tag{20}$$

where ω_X and ω_Y denotes weights, and S_{30}^X and S_{30}^Y - similarity matrices of X and Y respectively.

Figure 10 shows, how clustering of simulated data changes when a weight of similarity matrix X (ω_X in equation 20) changes from 0 to 1 with step 0.1. Each row represent the clustering solution for a certain value of weight. Three colors on the Figure highlight three different clusters. The row 'Only Y' on Figure 10 corresponds to situation of weight of 0 assigned to matrix X and weight of 1 by matrix Y. This implies clustering based on data source Y only. It can be observed, that clustering on line 'Only Y' is exactly the same as on the right dendrogram on Figure 9: first 15 samples belong to the first cluster (colored in brown), samples 16 to 25 form the second cluster, colored in red, and the last five sample belong to a yellow-colored cluster. This structure is maintained with increasing weight of matrix X from 0 to 0.9. Clustering changes only when weight of X is larger, than 0.9 (upper row on Figure 10). This change of structure is shown explicitly on Figure 11. It is shown, that the 3-clusters structure of integrated data is still preserved as it is in matrix Y until the weight of matrix X reaches value of 0.93. On the left dendrogram (Figure 11) first three groups of samples in data set X cluster together, which corresponds to the first cluster of data set Y. Fourth and fifth groups of samples from X form the second cluster of Y, and the last group of X coincide with the last group in Y. For values of weight of X equal to 0.93 and higher, the 3-clusters structure of matrix X, which is determined by random sampling from normal distribution during data generation, affects integrated clustering: clustering on the right dendrogram does not coincide with clustering in data set Y.

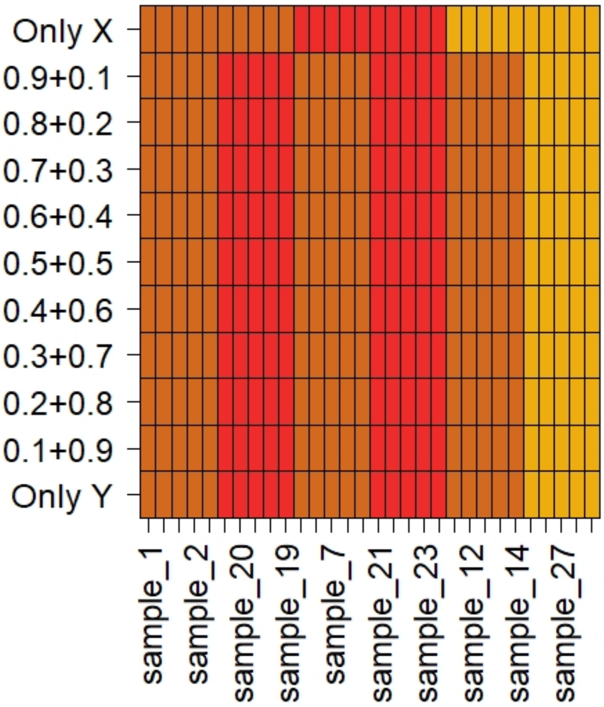


Figure 10: Clustering of simulated data with gradual changing weight of X. The colors denote the cluster membership of the sample per clustering solution.

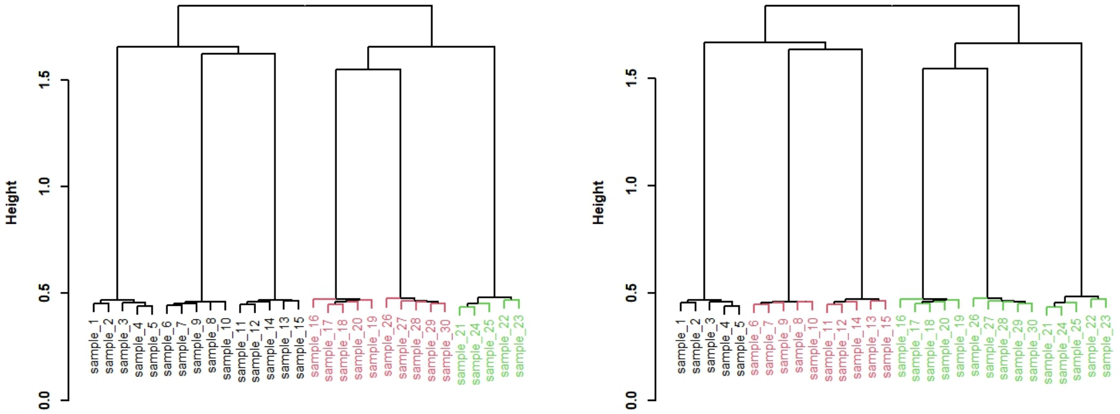


Figure 11: Change in clustering of integrated data when weight of similarity matrix X changes from 0.92 (left) to 0.93 (right).

As can be seen on Figure 10, the 3-clusters structure of weighted integrated data is quite stable and is predominantly determined by clustering in matrix Y. The stability of the 3-clusters structure of integrated data with increasing weight of data source X is due to the strength of intragroup correlation of samples in matrix Y. It is illustrated on Figure 12. The graph shows how the maximal weight of X, at which 3-clusters structure is preserved as it is in Y, depends on intragroup correlation in Y. As it was expected, lower intragroup correlation in Y leads to lower stability of 3-clusters structure of integrated data when the weight of similarity matrix X increases. As has been shown previously, at correlation value of 0.75 the structure is preserved until weight of X reaches a value as high as 0.93. In a contrast, correlation of 0.1 decreases maximal possible weight of X to approximately 0.6, as can be seen on Figure 12.

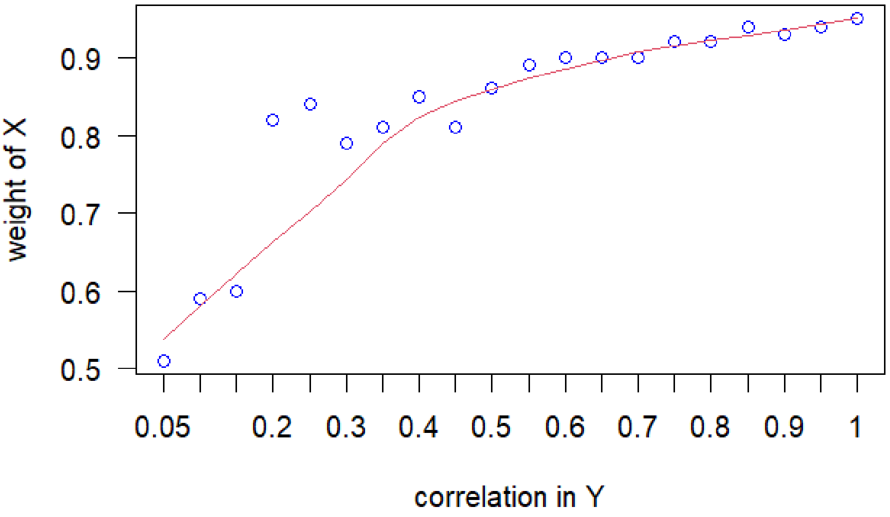


Figure 12: Dependency between the intragroup correlation in matrix Y and maximal weiht of X, at which clustering structure of integrated data is preserved.

4.2 ABC clustering

Single-source clustering solution for simulated data, obtained using ABC method, is represented on Figure 13.

Groups of samples are identified in data set X as they are determined by simulation. There are six clusters of five samples in each and equal height of fusion, which implies equal intragroup similarity.

Clustering solution for the Y data set also coincides with the true structure in data: there are three clusters, including 15, 10 and 5 samples. However, the fusion height is clearly uneven: the biggest cluster, which includes 15 samples, has the lowest point, while the smallest cluster has the highest point of fusion. This result is invariably repeated in different runs of the procedure and therefore cannot be explained by random fluctuation.

4.3 Multi-source ABC clustering

Multi-source clustering solution for number of clusters K=3 and K=6 is presented on Figure 14. It can be observed, that grouping of samples in integrative clustering into six clusters correspond to grouping in data set

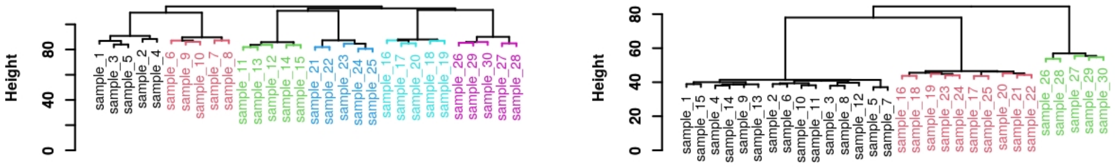


Figure 13: Clustering solution for simulated data by ABC approach: clustering in data set X (left) and Y (right).

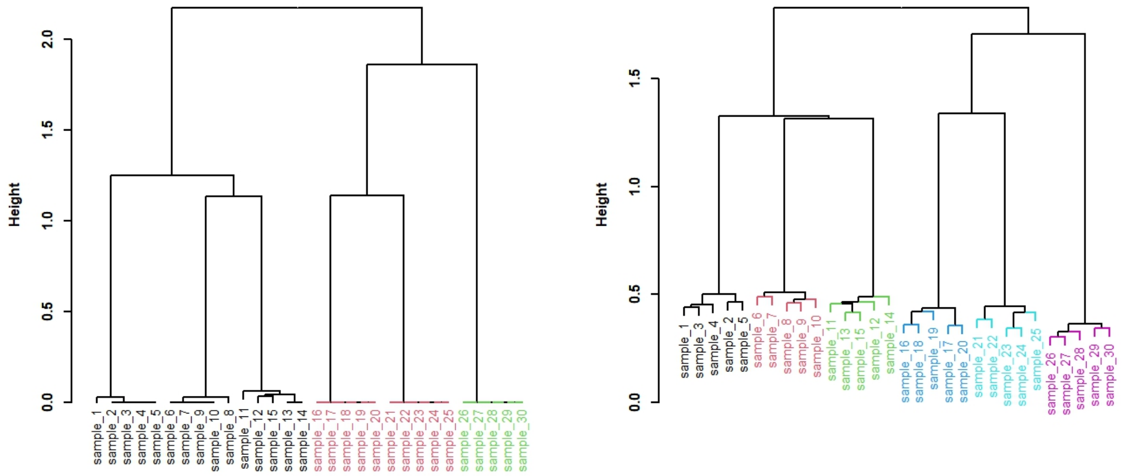


Figure 14: 3-clusters (left) and 6-clusters (right) structure for simulated data by multi-source ABC approach.

X. When value of K changes from 6 to 3, first 3 groups cluster together and groups 4 and 5 form the second cluster. This 3-clusters structure fully corresponds to the structure in the Y data set.

5 Application of the methods to the radiation experiment data

Samples in radiation experiment data consist of four groups, depending on day of tissue selection (11 or 17) and region of brain, where tissue was taken from (cortex or hippocampus). Animals in each of these four groups can be further subdivided by four, with regard to whether they were exposed to radiation or not, and whether their mothers received a diet, rich of folic acid. In the following, 'Folic acid +' denotes the group of animals whose mothers were fed a folic acid-rich diet, and 'Folic acid -' denotes the group of animals whose mothers were fed a standard diet. 'Radiation +' denotes the group of animals, that have been exposed to radiation, and 'Radiation -' denotes intact animals. Identification numbers (ID) of all animals are given in Table 4.

Table 4: Groups of animals with regard to radiation status and treatment of mother.

	Radiation+	*Radiation−*
	Cortex, day 11	
Folic acid +	26, 28, 33, 34	30, 32, 36
Folic acid -	1, 2, 6, 8	3, 4, 9, 11
	Cortex, day 17	
Folic acid +	39, 40, 43, 44	38, 45, 46, 48
Folic acid -	18, 19, 20, 23, 24	13, 15, 21
	Hippocampus, day 11	
Folic acid +	74, 76, 81, 82	78, 80, 83, 84
Folic acid -	50, 53, 54, 56	51, 52, 57, 59
	Hippocampus, day 17	
Folic acid +	87, 88, 91, 92	86, 93, 94, 96
Folic acid -	65, 66, 68, 71, 72	63, 69, 61

Table 5 presents the results of application of the 'gap' method to radiation data to define the optimal number of clusters K. The obtained results indicate K=1 for all data sets, as well as for all merged data. This value of K implies no clustering in the data. Since optimal value of K can not be defined for these data, the number of groups of animals in the experiment will be used as a number of clusters in further analysis. One can suppose that clustering of animals may be due to their radiation exposure status. Groups of mice whose mothers received or not received a folic acid-rich diet may also form clusters. Finally, both of these divisions can simultaneously lead to clustering of animals. Based on these considerations, a values of K equal to 2 and 4 were used in the subsequent analysis.

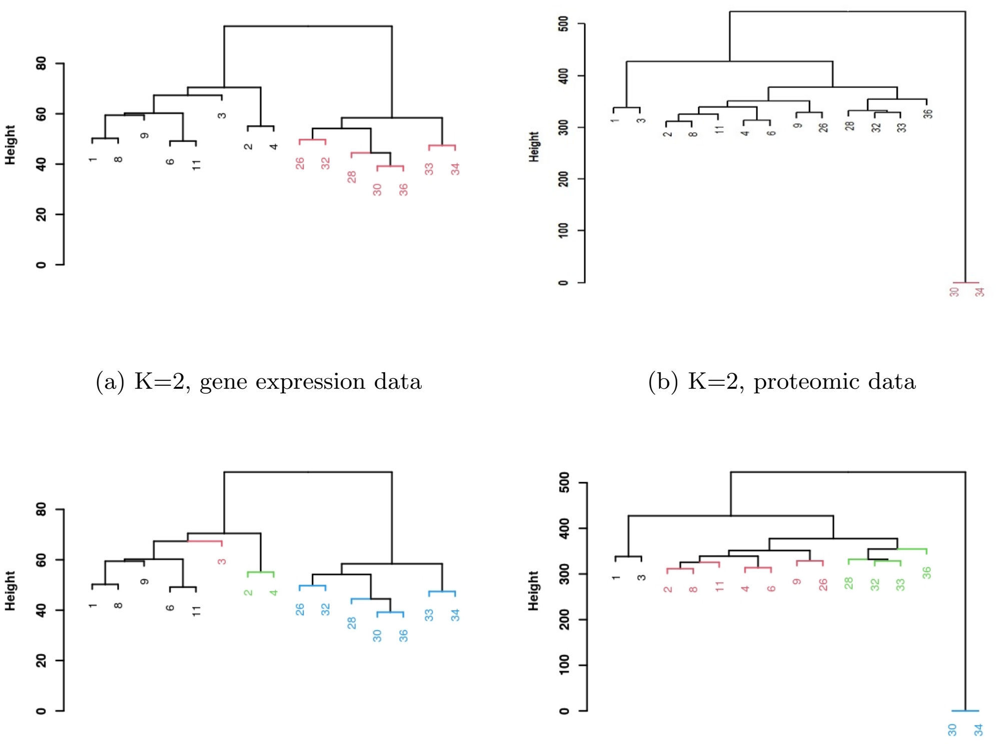
Table 5: Definition of the optimal number of clusters in radiation data with the 'gap' method.

Data set	K
Cortex, day 11	
Gene expression data	1
Proteins data	1
Merged gene expression and proteins data	1
Cortex, day 17	
Gene expression data	1
Proteins data	1
Merged gene expression and proteins data	1
Hippocampus, day 11	
Gene expression data	1
Proteins data	1
Merged gene expression and proteins data	1
Hippocampus, day 17	
Gene expression data	1
Proteins data	1
Merged gene expression and proteins data	1

In this report only result of analysis for two settings are presented: cortex data, day 11 and hippocampus data, day 11.

5.1 Gene expression and proteomic data from cortex samples collected at day 11

Hierarchical clustering of gene expression and proteomic data obtained from cortex samples, collected at day 11 after exposure of mice to radiation, is presented on Figure 15. The samples on the Figure are indicated by the numbers of the animals from which they were taken. The animal numbers in the Figure correspond to the identification numbers in the Table 4. Clusters of samples are highlighted by different colors. It can be observed, that the 2-clusters structure, identified in gene expression data (Figure 15 (a)), perfectly reflects the difference in folic acid treatment status of mothers of mice. All samples from the group 'Folic acid +' are clustered in one cluster, while all animals from the group 'Folic acid -' form the second cluster. In the proteomic data one of the two clusters includes only two samples, with ID 30 and 34, (Figure 15 (b)), which obviously do not exhaust any group of experimental animals. However, both samples belong to the same group 'Folic acid +'.



(a) K=2, gene expression data

(b) K=2, proteomic data

(c) K=4, gene expression data

(d) K=4, proteomic data

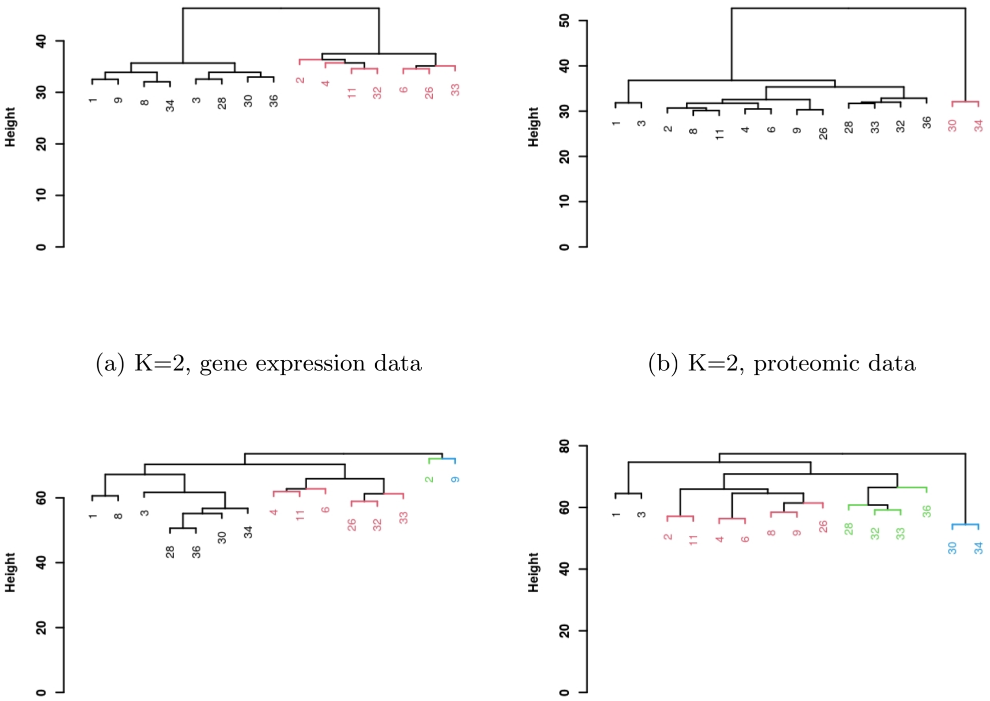
Figure 15: Single-source hierarchical clustering in gene expression and proteomic data for cortex samples, day 11.

In 4-clusters structure of the gene expression data all samples, obtained from animals of group 'Folic acid +' still form a separate cluster and all remaining samples are divided by three clusters (Figure 15 (c)). Clustering solution for proteomic data with K=4 does not match neither treatment nor radiation status of animals.

5.1.1 ABC and multi-source ABC methods

Single-source ABC clustering solutions for cortex data, day 11, are presented on Figure 16. This result differs from the simple (non-consensus) hierarchical single-source clustering, discussed in the previous section. For a number of clusters K=2 clustering in gene expression and proteomic data do not display any reflection of treatment or radiation exposure status of animals (Figure 16 (a,b)). However, two samples, which form separate cluster in proteomic data, both belong to the 'Folic acid +' group. These are again the samples with ID 30 and 34.

For K=4, single-source ABC clustering of gene expression data does not show any meaningful structure (Figure 16 (c)). However, in proteomic data (Figure 16 (d)) two out of four defined clusters include only



(a) K=2, gene expression data

(b) K=2, proteomic data

(c) K=4, gene expression data

(d) K=4, proteomic data

Figure 16: ABC method applied to gene expression and proteomic data for cortex samples, day 11.

samples from group 'Folic acid +', third cluster includes only samples from group 'Folic acid -' and the fourth cluster includes one (ID 26) sample from the 'Folic acid +' group and six samples from the 'Folic acid -' group.

Multi-source ABC clustering of cortex data, day 11, is shown on Figure 17. Results for K=4 are close to a true reflection of the division of animals by maternal diet status (Figure 17, (b)). There are two clusters defined, which include only samples from the 'Folic acid -' group, one cluster which includes only samples from the 'Folic acid +' group and one cluster, which includes one sample from the 'Folic acid -' group and three samples from the 'Folic acid +' group.

For both single-source and multi-source ABC clustering, there is no connection between clustering structure and radiation status observed.

5.1.2 Aggregated data clustering

Aggregation scheme and dimensionality of resulting matrices are shown in the Table 6.

Table 6: Aggregating of radiation experiment cortex data.

Dimension	Value
Number of samples	15
Number of proteomic features	3102
Number of gene expression features	18380
Total number of features	21482

Clustering of the aggregated gene expression and proteomic data is shown on Figure 18. For number of clusters K=2 there are two samples, ID 1 and 3, which form the separate cluster and both belong to the group 'Folic acid -' (Figure 18, (a)). These two samples also form the separate cluster in the 4-clusters structure (Figure 18, (b)).

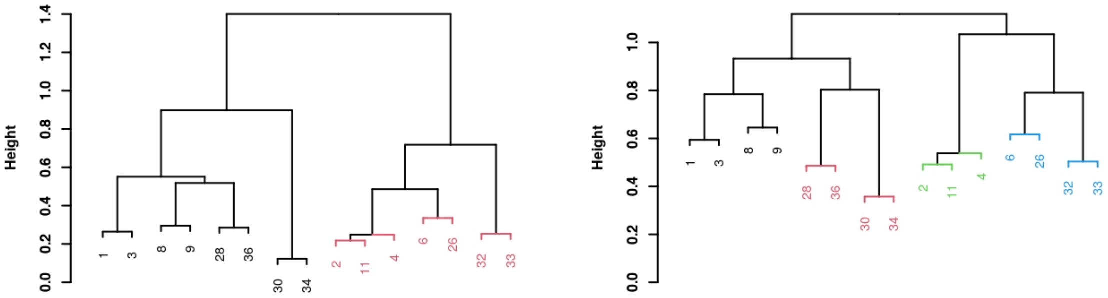


Figure 17: Multi-source ABC method applied to data for cortex samples, day 11, with K=2 (left) and K=4 (right).

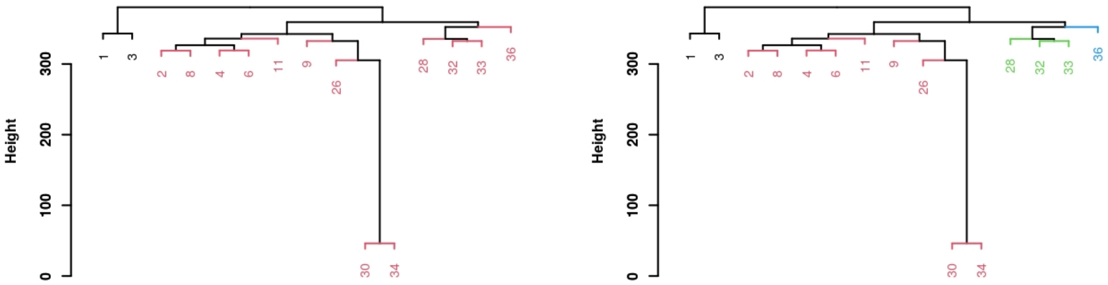


Figure 18: Aggregated data clustering for cortex samples collected at day 11 with K=2 (left) and K=4 (right).

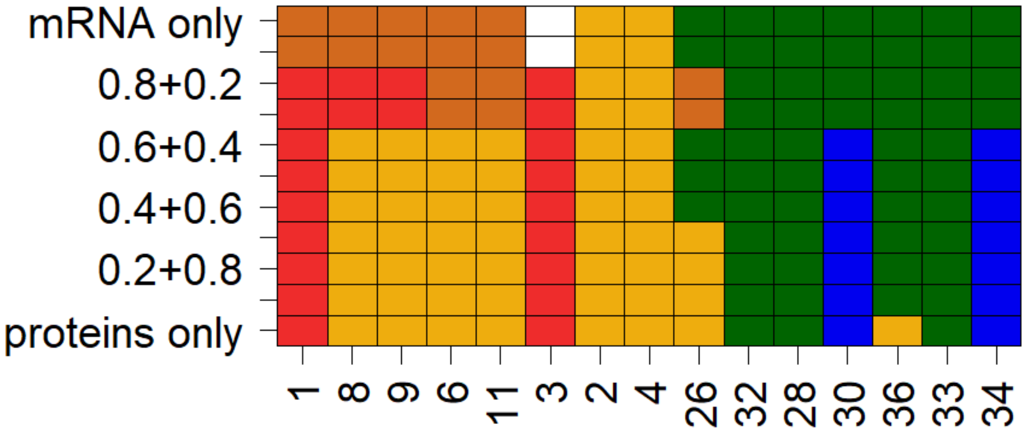


Figure 19: Weighted similarity-based clustering in cortex data, day 11.

5.1.3 Weighted similarity-based clustering

In this approach the weighted similarity matrix was calculated by assigning weights to similarity matrices of the gene expression data set and the proteins data set. Figure 19 illustrates how clustering changes when a weight of similarity matrix of proteins data set changes from 0 to 1 when K=4. It can be observed the stable grouping of samples with regard to the folic acid-treatment status of mothers. This grouping is largely preserved over all range of weights, assigned to similarity matrices. Pair of samples 1 and 3, as well as pair 30 and 34, cluster together into a separate cluster, and fuse with other samples only when a weight of the gene expression data reaches value of 0.7.

5.1.4 Complementary ensemble clustering

Results of application of the complementary ensemble clustering method to the radiation data from cortex, day 11, are presented on Figure 20. In case of partition by 4 clusters, all samples from animals, whose mothers received a folic acid-rich diet, are grouped together into one cluster (this cluster is highlighted by blue on the Figure 20). Samples from the 'Folic acid -' group are distributed between remaining three clusters. In case of partition by 2 clusters, samples 1 and 3 form the separate cluster, which is consistent with results of other methods application. For both K=2 and K=4 there is no influence of radiation exposure status on the clustering structure observed.

5.2 Gene expression and proteomic data from hippocampus samples collected at day 11

Hierarchical clustering of gene expression and proteomic data obtained from hippocampus samples, collected at day 11 after exposure of mice to radiation, is presented on Figure 21. The samples on the Figure are indicated by the numbers of the animals from which they were taken. The animal numbers on the Figure 21 correspond to the identification numbers in the Table 4. Clusters of samples are highlighted by different colors. Single-source clustering solution for proteomic data (Figure 21, (b)) for K=2 implies a separate cluster for four samples (ID 50, 54, 57 and 51), all of which belong to the 'Folic acid -' group. Grouping of animals by folic acid treatment of mothers is captured by a single-source clustering of proteomic data with K=4 (Figure 21, (d)): two clusters, highlighted by green and blue, include all samples from the group 'Folic acid +' and two cluster, highlighted by red and black, include all samples from the group 'Folic acid -'.

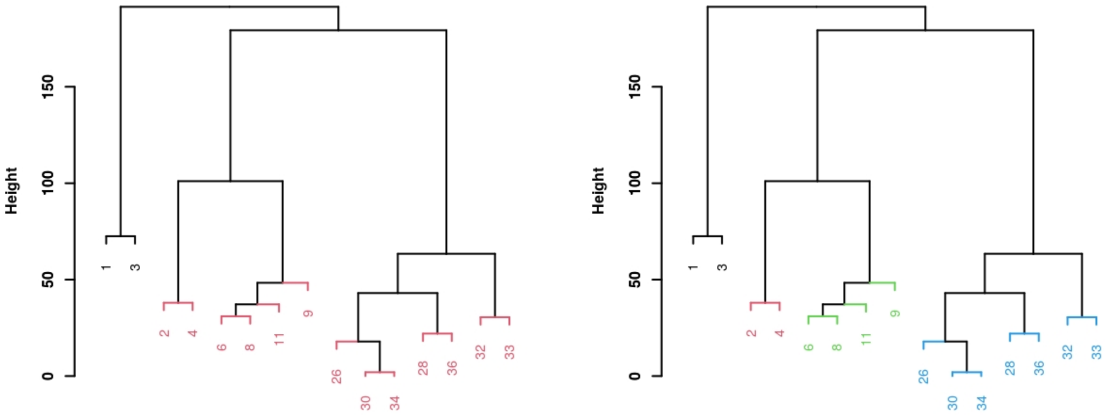
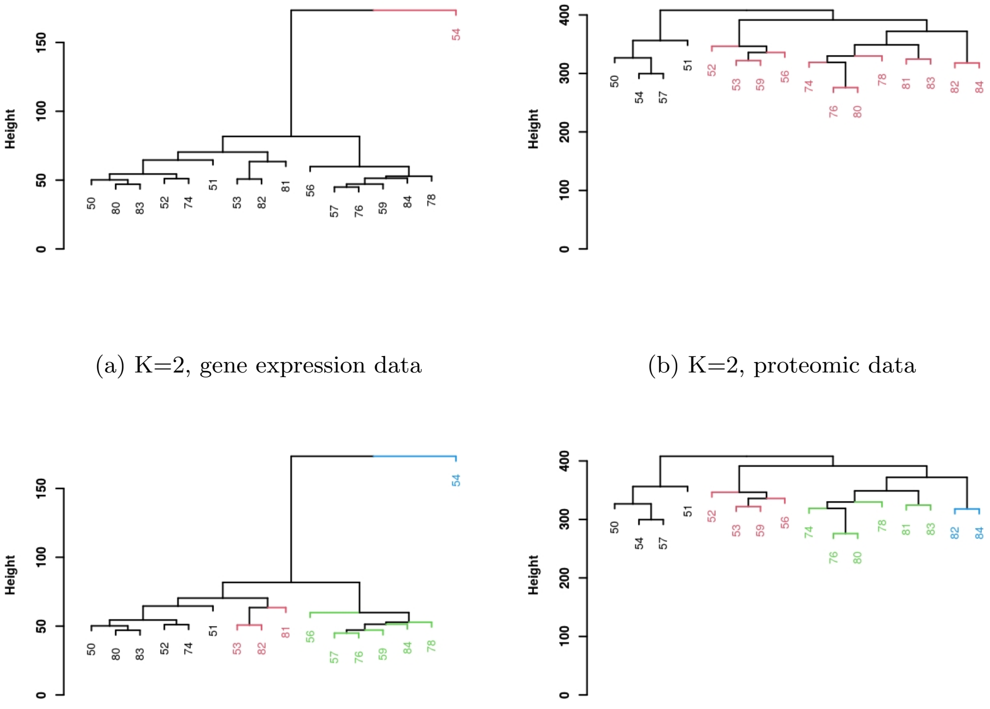


Figure 20: CEC for cortex samples collected at day 11 with K=2 (left) and K=4 (right).



(a) K=2, gene expression data

(b) K=2, proteomic data

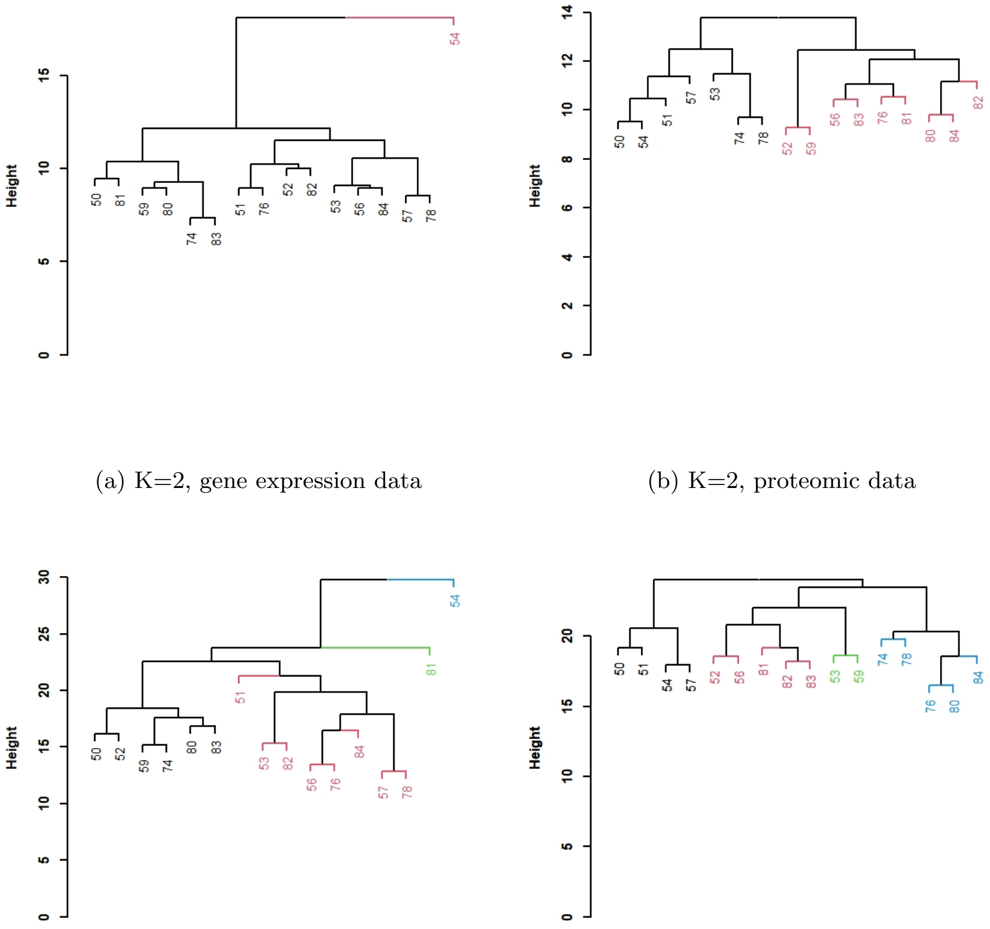
(c) K=4, gene expression data

(d) K=4, proteomic data

Figure 21: Single-source hierarchical clustering in gene expression and proteomic data for hippocampus samples, day 11.

5.2.1 ABC and multi-source ABC methods

For K=2 the single-source ABC clustering of the gene expression data coincide with solution, obtained by single-source simple (non-consensus) hierarchical clustering (Figure 22, (a)). In both solutions, there is only one sample (ID 54), allocated to a separate cluster. However, ABC clustering for K=2 and K=4 does not reflect grouping of animals, neither with regard to treatment, nor to radiation status.



(a) K=2, gene expression data

(b) K=2, proteomic data

(c) K=4, gene expression data

(d) K=4, proteomic data

Figure 22: ABC method applied to gene expression and proteomic data for hippocampus samples, day 11.

Multi-source ABC clustering for K=2 and K=4 is shown on Figure 23. The dendrogram on the left, on which two clusters are highlighted, display exact division of samples with regard to treatment status of mothers. All animals, whose mothers received folic acid-rich diet, are gathered in a cluster highlighted by red. This structure is not preserved with K=2 (dendrogram on the right of Figure 23). Three out of four clusters include samples obtained from animals with different folic acid-treatment and radiation status. Only one cluster, colored in black, includes samples from the same treatment group ('Folic acid -').

5.2.2 Aggregated data clustering

Aggregation scheme and dimensionality of resulting matrices are shown in the Table 7.

Table 7: Aggregating of radiation experiment hippocampus data.

Dimension	Value
Number of samples	16
Number of proteomic features	3011
Number of gene expression features	18329
Total number of features	21340

Clustering of the merged gene expression and proteomic data is shown on Figure 24. Two clusters, presented

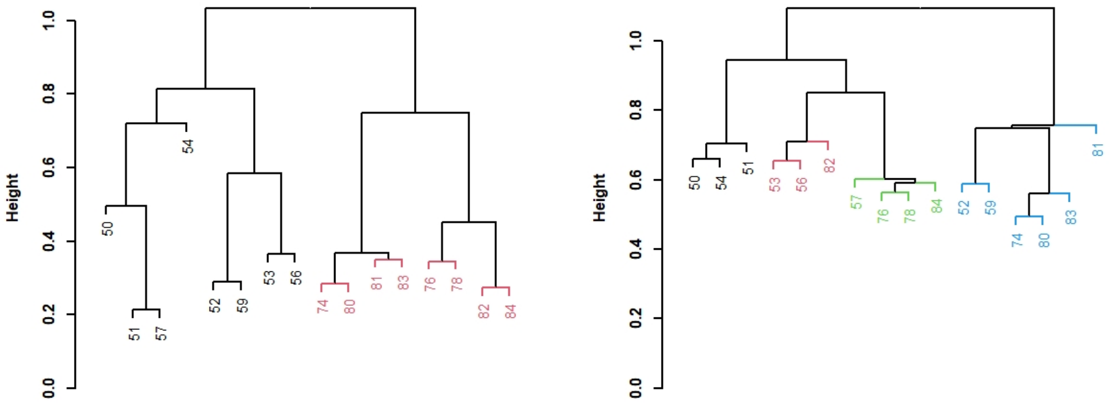


Figure 23: Multi-source ABC method applied to data for hippocampus samples, day 11, with K=2 (left) and K=4 (right).

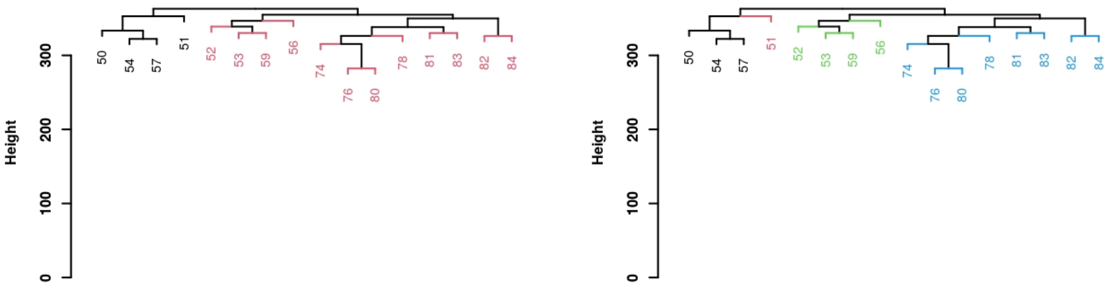


Figure 24: Aggregated data clustering for hippocampus samples collected at day 11 with K=2 (left) and K=4 (right).

on the left dendrogram, do not capture grouping of experimental animals by design. However, cluster colored by black includes samples of the same treatment group. Better reflection of grouping of animals is obtained for K=4: all animals, whose mothers received a folic acid-rich diet, are group together into a cluster, highlighted by blue. Animals, whose mothers received usual diet, are distributed between remaining three clusters (black, red and green).

5.2.3 Weighted similarity-based clustering

Weights are assigned to similarity matrices of the gene expression data set and the proteins data set with further summation of similarity matrices. Figure 25 illustrates change of clustering when a weight of similarity matrix of proteins data set changes from 0 to 1 with K=4. It can be observed, that integrative clustering differs from a single-source clustering only when assigned weights are not lower than 0.3 (or no higher that 0.7). Structure, reflecting treatment status of animals, can be observed on line ’0.3+0.7’ which implies following expression for the weighted similarity matrix S_{16}^{W}:

$$S_{16}^{W} = 0.3S_{16}^{GE} + 0.7S_{16}^{Proteins}, \tag{21}$$

where S_{16}^{GE} denotes a similarity matrix of the gene expression data, $S_{16}^{Proteins}$ denotes a similarity matrix of

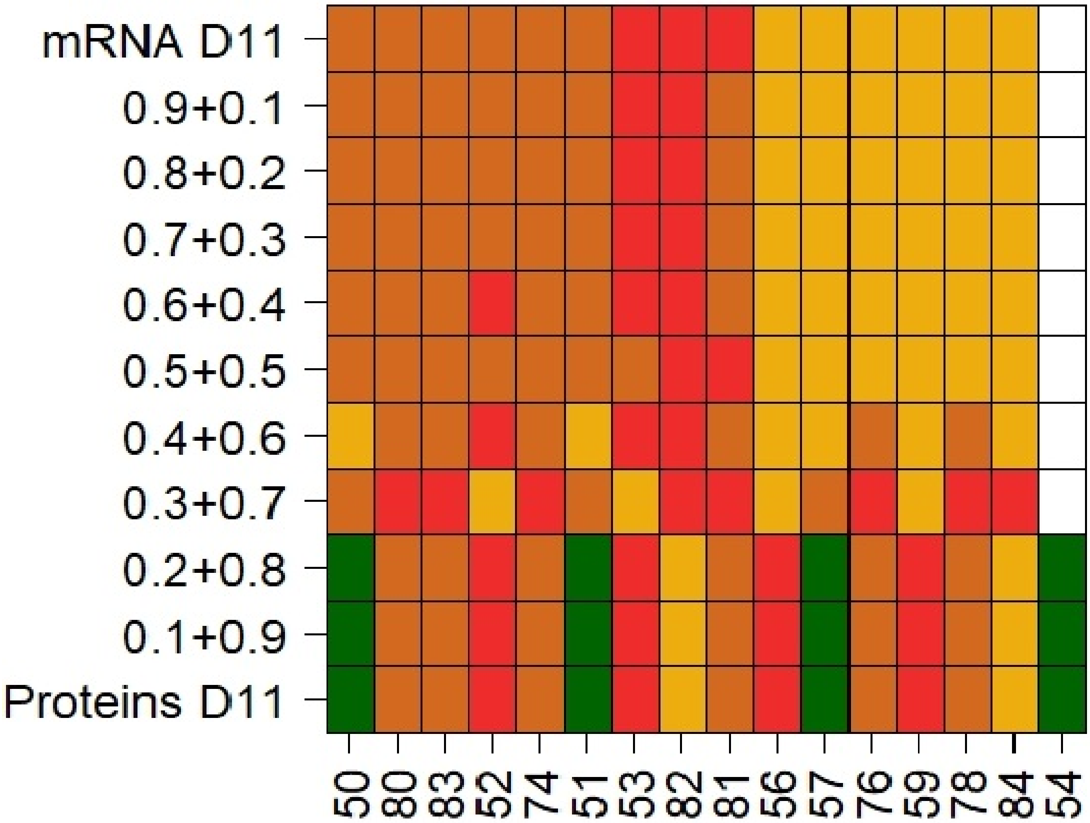


Figure 25: Weighted similarity-based clustering in hippocampus data, day 11.

the proteins data and the index '16' refers to a number of samples. In clustering solution for these weights all samples from the group 'Folic acid +' are gathered into one cluster, colored in red. Samples from the group 'Folic acid -' are distributed between remaining three clusters, colored in yellow, orange and white. For another values of weights no meaningful clustering structure can be observed.

5.2.4 Complementary ensemble clustering

Results of application of the complementary ensemble clustering method to the radiation data from hippocampus, day 11, are presented on Figure 26. As can be observed on the dendrogram, the sample with ID 54 fuses with others samples at large value of height, which indicates low similarity of this sample to all the others. This result is consistent with single-source clustering solutions for gene expression data and with weighted similarity-based clustering solution where a weight, assigned to the gene expression data, is equal to 0.3 or higher. For K=2 the sample 54 form the separate cluster, while all the remaining samples belong to the second cluster. For K=4 sample 54 remains the single sample in the separate cluster. Neither influence of radiation status of animals, nor of diet of mothers, can be noticed in this clustering structure.

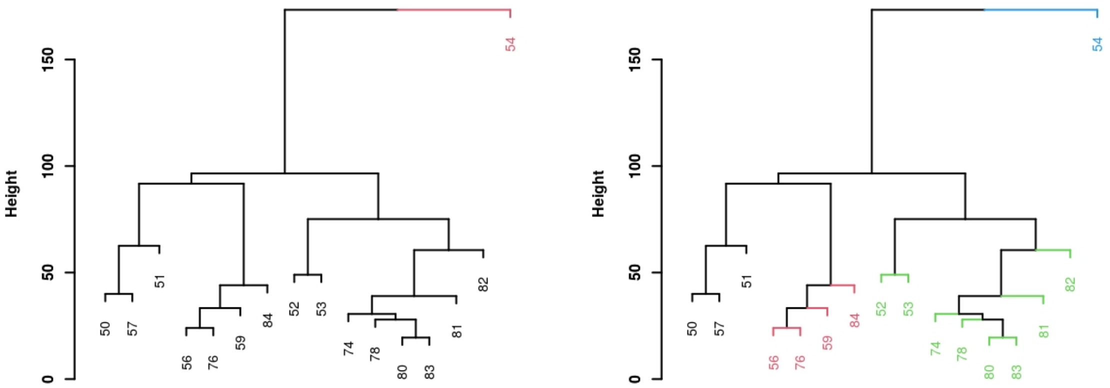


Figure 26: CEC for hippocampus samples collected at day 11 with K=2 (left) and K=4 (right).

6 Software

The analysis was conducted with the R software (version 4.3.3). Function Gap from the package clusterGenomics (verision 1.0) was used to define optimal number of clusters in the data sets. Cluster analysis was performed with usage of functions from the package IntClust (version 0.1.0): ADC (aggregated data clustering), ABC.SingleInMultiple (single-source ABC clustering), CEC (complementary ensemble clustering), Cluster (single-source clustering), M_ABC (multi-source ABC clustering) and WeightedClust (weighted similarity-based clustering). Listed fuctions may take as an input both the data itself and similarity matrices of the data sets. Functions ClusterPlot and ComparePlot from the package IntClust were used for visualization of results of cluster analysis.

While using the listed above functions from the IntClust package for cluster analysis, a few bugs were discovered. When working with function WeightClust, an error occurs when trying to set list of different weight combinations for the data sets consisting of a number of values other than ten. List of ten values is set by default.

According to the documentation, the output of the function ABC is a list of two elements: the resulting distance matrix and the resulting clustering structure. However, in practice the output of the function is a single matrix, consisting of numbers of clusters, which each sample was assigned to on each iteration. This output should be further processed to obtain a final clustering picture. In this project the output of the ABC function was further used as an input into the Cluster function.

7 Ethical thinking, societal relevance and stakeholder awareness

This research project does not include work with biological materials or experimental work with animals. Therefore, ethical considerations regarding these areas of research work are not applicable to this project. The study also does not involve working with personal data. The confidentiality of data used in the work was strictly maintained and was never compromised.

Only licensed and open-source software was used in the project. All methods of data analysis that were used in the work were published in peer-reviewed publications of the relevant field. The cluster analysis methods used are relevant, modern and known in the scientific community.

Modern approaches to the analysis of biological materials make it possible to obtain different types of omics data from one sample in one experiment. The demand for the accurate and efficient approaches to the analysis of multi-omics data is high and it can be assumed that it will remain high in the near future. The results of the integrative clustering methodology application presented in this report demonstrate the variety of approaches to this problem, as well as the advantages and disadvantages of individual methods. The work also touches on the question of the influence of folic acid content in the mother's diet on the proteome and transcriptome of the offspring. This issue, although it has been studied for a long time, does not lose relevance and remains the topic of many scientific works [18], [19].

The results of the study will be reported and defended at an open meeting. The results will be separately reported to the organization providing the data —the Belgian Nuclear Research Centre (SCK CEN), which is the main stakeholder. SCK CEN, which is located in Mol, Belgium, is a global leader in the field of nuclear research and innovations.

8 Discussion and conclusion

8.1 Discussion

Simulated data sets were used in this project in order to test the performance of selected methodology. Results of application of the 'gap' method to the simulated data correspond to the true underlying clustering structure in matrices X and Y. In the real experimental data, the 'gap' method defined no clustering structure (K=1 in all settings) both in single-source and aggregated data. Further cluster analysis of radiation data confirmed the absence of a strongly expressed clustering structure. Thus, the results obtained confirm the feasibility of using the 'gap' statistic to determine the optimal number of clusters in high-dimensional data.

Two approaches of multi-source data clustering were illustrated using the simulated data sets: weighted similarity-based clustering and Multi-source ABC clustering. Single-source ABC method also was applied, which resulted in correct classification of all samples both in X and Y data sets. Both weighted similarity-based and multi-source ABC methods identified the true underlying 3-clusters structure of the integrated simulated data. However, when a weight, assigned to the X data source using weighted similarity-based method, is higher than 0.92, the integrative clustering structure changes to incorrect. In case of zero weight assigned to the matrix Y, clustering of integrated data coincide with single-source clustering in X, which is expected.

The feasibility and usefulness of integrative analysis is clearly shown on the example of radiation experiment data for hippocampus, day 11. When analyzing the proteomic and gene expression data separately, no meaningful clustering structure was found. However, when the multi-source ABC method was applied to these data with parameter K=2, samples were divided by two clusters in full accordance with treatment status of mothers: all samples belonging to the 'Folic acid -'group are assigned to the first cluster, and all samples from the 'Folic acid +' group are assigned to the second cluster. Another example of finding meaningful structure in the integrated data, which was not detected in the single-source data analysis, is the results of weighted similarity-based clustering of data sets from hippocampus, day 11. For value of K=4 and weight of 0.3 assigned to gene expression data, clustering solution involves grouping all samples from animals, whose mothers received folic acid-rich diet, into one separate cluster. These results make weighted similarity-based and multi-source ABC clustering techniques the best performers among approaches that have been tried on the hippocampus, day 11 data set.

From the dendrograms in Figures 18 and 24 it can be seen that clustering solutions for aggregated data (ADC method) are shifted towards proteomic data single-source clustering. This result is counter-intuitive, since the size of proteomic data is almost 6 times smaller than the size of gene expression data. Clustering solution for concatenated data for obvious reasons is expected to be biased toward the omics with the most numerous features. However, the observed shift towards proteomic data source can be explained by difference in measurement scale and differences in variability in proteomic and transcriptomic data. It can be assumed that the results of the analysis with the ADC approach will be more objective if a step of data standardization is introduced after merging the two matrices.

The results obtained illustrate that simple hierarchical clustering may outperform the consensus-based ensemble approach. Using simple non-ensemble hierarchical clustering analysis, a cluster structure was identified in the cortex data, day 11, separating animals whose mothers received a folic acid-rich diet from all other animals. These animals are grouped in one cluster both when K=2 and when K=4. However, single-source ABC approach, applied to these data, failed to separate animals depending on treatment status of mothers. This structure was also not found by multi-source ABC method and by ADC method. CEC approach made it possible to distinguish animals from the 'Folic acid +' into a separate group, but only for K=4. This structure was also defined by the weighted similarity-based approach, but only when a weight, assigned to gene expression data, is 0.9 or higher.

For the sake of simplicity, this project ignores the possibility of animals belonging to the same litter. A group of animals from the same litter is likely to have less variability in proteomic and transcriptomic features compared to the overall variability. This can cause animals from the same litter to cluster together [21]. It can be observed, that some pairs of samples are often clustered together regardless of the method used and of the omics data type. For example, samples 30 and 34 in cortex data, day 11, always belong to the same cluster and often form the separate cluster with only two members. The similarity of these samples is most pronounced in the results of the ADC approach, where the point of fusion of samples 30 and 34 is far lower than points of fusion of all the other samples. Although the belonging of animals to one litter or another, as well as the protocol for randomizing animals into groups, are unknown, looking at these results it can be assumed that animals with ID 30 and 34, as well as some other pairs of animals, are related.

8.2 Limitations and future research

One of possible drawbacks of the study is using of only one clustering technique - hierarchical clustering. It was shown in the literature, that a method of integrated clustering, which uses several algorithms of clustering to identify clusters which are robust against the choice of clustering algorithms, outperforms existing approaches in identifying known subtypes of breast cancer [20].

One of possible ways to improve the clustering algorithm is to add a step of introducing noise into data. In order to discover reliable groups of samples, one may estimate how often each pair of samples is grouped together when data are perturbed, for example, by adding Gaussian noise [20].

All data sets used in this work contained no missing data and the applied method require full data for correct performance. However, in real experimental settings, often for some samples or patients only a subset of the omics features is available. Solutions, which imply using only those samples with all omics features measured, have obvious disadvantages, as well as usage of imputation techniques. At the same time, it is crucial to be able to analyze partial data because of high cost of experiments and the unequal cost for acquiring data for different omics.

8.3 Conclusion

Based on simulations, it is shown that methods of weighted similarity-based clustering and multi-source ABC clustering are able to recover the true integrative clustering of samples. In real experimental data (hippocampus, day 17), joint analysis of multiple omics data with these two methods allowed the separation of samples from animals whose mothers consumed a diet rich in folic acid from those whose mothers did not. This information was only available at the integrative level, whereas at the single-source level no grouping of samples according to maternal folic acid treatment status was observed. However, in other experimental settings integrative clustering analysis by the selected methods did not provide new information not available at the single-source level.

References

[1] Indhupriya Subramanian et al. “Multi-omics Data Integration, Interpretation, and Its Application”. In: *Bioinform Biol Insights* 14 (2020). DOI: 10.1177/1177932219899051.

[2] Charles J. Vaske et al. “Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM”. In: *Bioinformatics* 26(12) (2010), pp. 237–245. DOI: 10.1093/bioinformatics/btq182.

[3] Yinyin Yuan et al. “Patient-Specific Data Fusion Defines Prognostic Cancer Subtypes”. In: *PLoS Comput Biol* 7(10) (2011). DOI: 10.1371/journal.pcbi.1002227.

[4] Nimrod Rappoport and Ron Shamir. “Multi-omic and multi-view clustering algorithms: review and cancer benchmark”. In: *Nucleic Acids Res* 46(20) (2018). DOI: 10.1093/nar/gky889.

[5] Dingming Wu et al. “Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification”. In: *BMC Genomics* 16 (2015). DOI: 10.1186/s12864-015-2223-8.

[6] Stefano Monti et al. “Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data”. In: *Machine Learning* 52 (2004). DOI: 10.1023/A:1023949509487.

[7] Dhammika Amaratunga et al. “Microarray learning with ABC”. In: *Biostatistics* 9(1) (2008), pp. 128–36. DOI: http://dx.doi.org/10.1093/biostatistics/kxm017.

[8] Katherine Hoadley et al. “Multi-platform analysis of 12 cancer types reveals molecular classification within and across tissues-of-origin”. In: *Cell* 158(4) (2015). DOI: 10.1016/j.cell.2014.06.049.

[9] Bo Wang et al. “Similarity network fusion for aggregating data types on a genomic scale”. In: *Nature methods* 11(3) (2014). DOI: 10.1038/nmeth.2810.

[10] Nimrod Rappoport and Ron Shamir. “NEMO: cancer subtyping by integration of partial multi-omic data”. In: *Bioinformatics* 35(18) (2019). DOI: 10.1093/bioinformatics/btz058.

[11] Nolen Perualila-Tan et al. “Weighted similarity-based clustering og chemical structures and bioactivity data in early drug discovery”. In: *Journal of Bioinformatics and Computational Biology* 14(4) (2016). DOI: http://dx.doi.org/10.1142/S0219720016500189.

[12] Nora K. Speicher and Nico Pfeifer. “Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery”. In: *Bioinformatics* 31(12) (2015). DOI: 10.1093/bioinformatics/btv244.

[13] SCK CEN. *Webcite*. URL: https://www.sckcen.be/en.

[14] Gareth James et al. *An introduction to statistical learning*. Springer US, 2021. ISBN: 9781071614174.

[15] Jr Ward J.H. “Hierarchical grouping to optimise an objective function”. In: *Journal od the American statistical association* 58 (1963), pp. 236–244. DOI: http://dx.doi.org/Ward,J.H.(1963).HierarchicalGroupingtoOptimizeanObjectiveFunction.JournaloftheAmericanStatisticalAssociation,58(301),236.doi:10.2307/2282967.

[16] Samah Jamal Fodeh et al. “Complementary ensemble clustering of biomedical data”. In: *Journal of Biomedical informatics* 46(3) (2013), pp. 436–443. DOI: http://dx.doi.org/10.1016/j.jbi.2013.02.001.

[17] Robert Tibshirani et al. “Estimating the number of clusters in a data set via the gap statistic”. In: *Journal of the Royal Statistical Society* 63 (2001), pp. 411–423. DOI: http://dx.doi.org/10.1111/1467-9868.00293.

[18] Taylor A Ricci et al. “Maternal nutrition and effects on offspring vascular function”. In: *Pflugers Arch* 475(7) (2023). DOI: 10.1007/s00424-023-02807-x.

[19] Xiguang Xu et al. “Risk of Excess Maternal Folic Acid Supplementation in Offspring”. In: *Nutrients* 16(5) (2024). DOI: 10.3390/nu16050755.

[20] Hung Nguyen et al. “PINSPlus: a tool for tumor subtype discovery in integrated genomic data”. In: *Bioinformatics* 35(16) (2019). DOI: 10.1093/bioinformatics/bty1049.

[21] Marc Aerts et al. *Topics in Modelling of Clustered Data*. Chapman Hall, 2002. ISBN: 9780429119224.

Appendix

```
library(clusterGenomics)
library(IntClust)

###################### Generation of data ######################

# XDAT.A
n <- 1000 # No. of features
nvar=5
R <- matrix(c(1, 0.75,0.75,0.75,0.75,
             0.75,1,0.75,0.75,0.75,
             0.75,0.75,1,0.75,0.75,
             0.75,0.75,0.75,1,0.75,
             0.75,0.75,0.75,0.75,1),
           nrow = nvar, ncol = nvar, byrow = TRUE)
dim(R)
mu <- rep(0,nvar)
x.dat1<-mvtnorm::rmvnorm(n, mean = mu, sigma = R)
dim(x.dat1)
cor(x.dat1)
x.dat2<-mvtnorm::rmvnorm(n, mean = mu, sigma = R)
x.dat3<-mvtnorm::rmvnorm(n, mean = mu, sigma = R)
x.dat4<-mvtnorm::rmvnorm(n, mean = mu, sigma = R)
x.dat5<-mvtnorm::rmvnorm(n, mean = mu, sigma = R)
x.dat6<-mvtnorm::rmvnorm(n, mean = mu, sigma = R)
xdat.a<-cbind(x.dat1,x.dat2,x.dat3,x.dat4,x.dat5,x.dat6)
sample_names <- paste0("sample_", rep(1:30))
colnames(xdat.a) <- sample_names # Changing the names samples/subject ID
dim(xdat.a) # should be a 1000X30 matrix
cor(xdat.a)
print(head(xdat.a))


# YDAT.A
n <- 3000 # Number of features
nvar=15
vec<-rep(0.75,nvar)
vec
R<-matrix(rep(vec,nvar),nvar,nvar)
dim(R)
for(i in 1:nvar)
{
  R[i,i]<-1
}
diag(R)
R[1,]
R[2,]
```

```
mu <- rep(0,nvar)
y.dat1<-mvtnorm::rmvnorm(n, mean = mu, sigma = R)
dim(y.dat1)
cor(y.dat1)

nvar=10
vec<-rep(0.75,nvar)
vec
R<-matrix(rep(vec,nvar),nvar,nvar)
dim(R)
for(i in 1:nvar)
{
  R[i,i]<-1
}
diag(R)
R[1,]
R[2,]
mu <- rep(0,nvar)
y.dat2<-mvtnorm::rmvnorm(n, mean = mu, sigma = R)
dim(y.dat2)
cor(y.dat2)


nvar=5
vec<-rep(0.75,nvar)
vec
R<-matrix(rep(vec,nvar),nvar,nvar)
dim(R)
for(i in 1:nvar)
{
  R[i,i]<-1
}
diag(R)
R[1,]
R[2,]
mu <- rep(0,nvar)
y.dat3<-mvtnorm::rmvnorm(n, mean = mu, sigma = R)
dim(y.dat3)
cor(y.dat3)

ydat.a<-cbind(y.dat1,y.dat2,y.dat3) # should be a 3000X30 matrix
sample_names <- paste0("sample_", rep(1:30))
colnames(ydat.a) <- sample_names
dim(ydat.a) # should be a 1000X30 matrix
cor(ydat.a)
print(head(ydat.a))

# Integrated data
```

```
xydat.a = rbind(xdat.a, ydat.a)
#-----------------------------------------------------------------------------------------------------------------
------#
# Visualization

# (a) xdat.a
# Image plot
random.data = t(xdat.a)
image(c(1:dim(random.data)[1]),c(1:dim(random.data)[2]),random.data,
ylab="features",xlab="samples")
# 3D
persp(c(1:dim(t(random.data))[2]), c(1:dim(t(random.data))[1]), random.data, theta = 325, phi
= 15, col = "yellow3", xlab = "Samples", ylab = "Features", zlab = " ")

# (b) ydat.a
# Image plot
random.data = t(ydat.a)
image(c(1:dim(random.data)[1]),c(1:dim(random.data)[2]),random.data,
ylab="features",xlab="samples")
# 3D
persp(c(1:dim(t(random.data))[2]), c(1:dim(t(random.data))[1]), random.data, theta = 325, phi
= 15, col = "yellow3", xlab = "Samples", ylab = "Features", zlab = " ")

# (c) Integrated data
xydat.a = rbind(xdat.a, ydat.a)


###################### Plots of similarity for "Data' section  #################

w=WeightedClust(list(t(xdat.a), t(ydat.a)),
type='data',distmeasure=c("euclidean","euclidean"))


image(c(1:dim(w$Dist[[1]])[1]),c(1:dim(w$Dist[[1]])[2]),w$Dist[[1]],col = gray.colors(12),
ylab="samples",xlab="Similarity matrix X")

image(c(1:dim(w$Dist[[2]])[1]),c(1:dim(w$Dist[[2]])[2]),w$Dist[[2]], col =
gray.colors(18),ylab="samples",xlab="Similarity matrix Y")

image(c(1:dim(w$Clust$DistM)[1]),c(1:dim(w$Clust$DistM)[2]),w$Clust$DistM, col =
hcl.colors(100, "terrain"),ylab="samples",xlab="Weighted similarity matrix")

image(c(1:dim(w$Dist[[1]])[1]),c(1:dim(w$Dist[[1]])[2]),0.25*w$Dist[[1]]+0.75*w$Dist[[2]],col =
hcl.colors(100, "terrain"), ylab="samples",xlab="Weighted similarity matrix")

image(c(1:dim(w$Dist[[1]])[1]),c(1:dim(w$Dist[[1]])[2]),0.75*w$Dist[[1]]+0.25*w$Dist[[2]],col =
hcl.colors(100, "terrain"), ylab="samples",xlab="Weighted similarity matrix")
```

```
#------------------------------------------------------------------------------------------------------------------------#
############## Gap method ####################
library(clusterGenomics)
gap(t(xydat.a),cl.method='hclust',dist.method='euclidean',linkage='ave',Kmax=15,B=1000)


############### Cluster Analysis of simulated data ##########################

library(readxl); library(dplyr); library(stringr); library(IntClust); library(clusterGenomics)
Colours <-
ColorPalette(colors=c("chocolate","firebrick2","darkgoldenrod2","darkgreen","blue2","darkorchid3","deeppink", "grey"),ncols=8)

# (a) xdat.a
xdat.a_clust=Cluster(Data=t(xdat.a), type='data', distmeasure='euclidean')
xdat.a_clust
ClusterPlot(Data1=xdat.a_clust, nrclusters=6,cols=Colours,main="xydat.a",ylim=c(-0.1,10))

# (b) ydat.a
ydat.a_clust=Cluster(Data=t(ydat.a), type='data', distmeasure='euclidean')
plot(ydat.a_clust)
ClusterPlot(Data1=ydat.a_clust, nrclusters=3,cols=Colours,main="xydat.a",ylim=c(-0.1,10))


# Weighted clustering
w=WeightedClust(list(t(xdat.a), t(ydat.a)),
type='data',distmeasure=c("euclidean","euclidean"))
ComparePlot(list(w),nrclusters=3,
cols=Colours,fusionsLog=TRUE,margins=c(8.1,10.1,4.1,4.1),names=c('X','0.9+0.1','0.8+0.2','0.7+0.3','0.6+0.4','0.5+0.5','0.4+0.6','0.3+0.7','0.2+0.8','0.1+0.9','Y'),
weightclust=F,plottype="new",location=NULL)


plot(w$Clust$Clust, main='weights=0.5')

rrr=DetermineWeight_SimClust(L, type='data',distmeasure=c("euclidean","euclidean"),
nrclusters=c(4,4))


d1=1*w$Dist[[1]]+0*w$Dist[[2]]
W_xy1=Cluster(Data=d1, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy1,nrclusters=3,cols=Colours)

d09=0.99*w$Dist[[1]]+0.01*w$Dist[[2]]
W_xy09=Cluster(Data=d09, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy09,nrclusters=3,cols=Colours)
```

```
d08=0.98*w$Dist[[1]]+0.02*w$Dist[[2]]
W_xy08=Cluster(Data=d08, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy08,nrclusters=3,cols=Colours)

d07=0.97*w$Dist[[1]]+0.03*w$Dist[[2]]
W_xy07=Cluster(Data=d07, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy07,nrclusters=3,cols=Colours)

d06=0.96*w$Dist[[1]]+0.04*w$Dist[[2]]
W_xy06=Cluster(Data=d06, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy06,nrclusters=3,cols=Colours)

d05=0.95*w$Dist[[1]]+0.05*w$Dist[[2]]
W_xy05=Cluster(Data=d05, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy05,nrclusters=3,cols=Colours)

d04=0.94*w$Dist[[1]]+0.06*w$Dist[[2]]
W_xy04=Cluster(Data=d04, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy04,nrclusters=3,cols=Colours)

d03=0.93*w$Dist[[1]]+0.07*w$Dist[[2]]
W_xy03=Cluster(Data=d03, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy03,nrclusters=3,cols=Colours)

d02=0.92*w$Dist[[1]]+0.08*w$Dist[[2]]
W_xy02=Cluster(Data=d02, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy02,nrclusters=3,cols=Colours)

d=0.92*w$Dist[[1]]+0.08*w$Dist[[2]]
W_xy=Cluster(Data=d, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy,nrclusters=6,cols=Colours)

d01=0.91*w$Dist[[1]]+0.09*w$Dist[[2]]
W_xy01=Cluster(Data=d01, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy01,nrclusters=3,cols=Colours)

d00=0.4*w$Dist[[1]]+0.6*w$Dist[[2]]
W_xy00=Cluster(Data=d00, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy00,nrclusters=3,cols=Colours)

L=list(W_xy1, W_xy09, W_xy08, W_xy07, W_xy06, W_xy05, W_xy04, W_xy03, W_xy02,
W_xy01, W_xy00)
ComparePlot(L,nrclusters=3, cols=Colours,margins=c(8.1,10.1,4.1,4.1),names=c('Only
X','0.99+0.01','0.98+0.02','0.97+0.03','0.96+0.04','0.95+0.05','0.94+0.06','0.93+0.07','0.92+0.
08','0.91+0.09','0.9+0.1'), weightclust=F,plottype="new",location=NULL)
```

```
## Dependence between correlation and weight of fusion
w=WeightedClust(list(t(xdat.a), t(ydat.a)),
type='data',distmeasure=c("euclidean","euclidean"))


d00=0*w$Dist[[1]]+1*w$Dist[[2]]
W_xy00=Cluster(Data=d00, type='dist', distmeasure='euclidean')
ClusterPlot(W_xy00,nrclusters=3,cols=Colours)

## ABC and mABC

X_data_ABC=ABC.SingleInMultiple(data=t(xdat.a),distmeasure="euclidean",
                              weighting=TRUE,stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=100,linkage="flexible",
                              alpha=0.625,NC=6, NC2=NULL, mds=FALSE)
X_data_ABC_clust=Cluster(Data=t(X_data_ABC), type='data', distmeasure='euclidean')
ClusterPlot(Data1=X_data_ABC_clust,
nrclusters=6,cols=Colours,main="ABC",ylim=c(-0.1,10))

ydat.a_ABC=ABC.SingleInMultiple(data=t(ydat.a),distmeasure="euclidean",
                        weighting=TRUE,stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=100,linkage="flexible",
                        alpha=0.625,NC=3, NC2=NULL, mds=FALSE)
ydat.a_ABC_clust=Cluster(Data=t(ydat.a_ABC), type='data', distmeasure='euclidean')
ClusterPlot(Data1=ydat.a_ABC_clust,
nrclusters=3,cols=Colours,main="ABC",ylim=c(-0.1,10))


L=list(t(xdat.a), t(ydat.a))
MABC=M_ABC(List=L,distmeasure=c("euclidean","euclidean"),
               weighting=c(TRUE, TRUE),stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=c(100,100),linkage=c("flexible","flexible"),
               alpha=0.625,NC=6, NC2=NULL, mds=FALSE)
ClusterPlot(Data1=MABC, nrclusters=6,cols=Colours,main="ABC",ylim=c(-0.1,10))


###############  Reading radiation  data for cortex, day 11 ###############
D11_mRNA_data = read.csv("D11_mRNA_Cor_normalized_tmm.csv", row.names = 1,
check.names = FALSE)
D11_Protein_data =read.csv("D11_Protein_Cor_normalized_quant.csv", row.names = 1,
check.names = FALSE)

# Finding optimal number of clusters with help of Gap statistic
res_clust_mRNA11=gap(t(D11_mRNA_data),
cl.method='hclust',dist.method='euclidean',linkage='ave',Kmax=15,B=100)
res_clust_mRNA11
```

```
res_clust_prot11=gap(t(D11_Protein_data),
cl.method='hclust',dist.method='euclidean',linkage='ave',Kmax=15,B=100)
res_clust_prot11

################### Cortex day 11 gene expression ###########################


D11Cg<-as.data.frame(D11_mRNA_data)

# Hierarchical clustering in gene expression data

D11Cg_clust=Cluster(Data=t(D11Cg), type='data', distmeasure='euclidean')
ClusterPlot(Data1=D11Cg_clust, nrclusters=4,cols=Colours,main="xydat.a",ylim=c(-0.1,10))


# Single-source ABC method and vizualization

D11Cg_ABC=ABC.SingleInMultiple(data=t(D11Cg),distmeasure="euclidean",
                    weighting=TRUE,stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=100,linkage="flexible",
                    alpha=0.625,NC=4, NC2=NULL, mds=FALSE)



D11Cg_ABC_clust=Cluster(Data=t(D11Cg_ABC), type='data', distmeasure='euclidean')
ClusterPlot(Data1=D11Cg_ABC_clust,
nrclusters=4,cols=Colours,main="ABC",ylim=c(-0.1,10))


######################### Cortex day 11 proteins ##########################

# Reading data

D11Cp<-as.data.frame(D11_Protein_data)


# Hierarchical clustering in proteins data


D11Cp_clust=Cluster(Data=t(D11Cp), type='data', distmeasure='euclidean')
ClusterPlot(Data1=D11Cp_clust, nrclusters=4,cols=Colours,main="xydat.a",ylim=c(-0.1,10))

# Single-source ABC clustering

D11Cp_ABC=ABC.SingleInMultiple(data=t(D11Cp),distmeasure="euclidean",
                   weighting=TRUE,stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=100,linkage="flexible",
                   alpha=0.625,NC=2, NC2=NULL, mds=FALSE)
```

```
D11Cp_ABC_clust=Cluster(Data=t(D11Cp_ABC), type='data', distmeasure='euclidean')
ClusterPlot(Data1=D11Cp_ABC_clust,
nrclusters=2,cols=Colours,main="ABC",ylim=c(-0.1,10))

################ Cortex day 11 integrative clustering #################
L=list(t(D11Cg), t(D11Cp))

#M_ABC method

MABC=M_ABC(List=L,distmeasure=c("euclidean","euclidean"),
         weighting=c(TRUE, TRUE),stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=c(100,100),linkage=c("flexible","flexible"),
         alpha=0.625,NC=4, NC2=NULL, mds=FALSE)
ClusterPlot(Data1=MABC, nrclusters=4,cols=Colours,main="ABC",ylim=c(-0.1,10))

#ADC method
D11C_ADC=ADC(L, distmeasure = "euclidean", normalize = FALSE, method = NULL,
   clust = "agnes", linkage = "ave", alpha = 0.625)
ClusterPlot(Data1=D11C_ADC, nrclusters=2,cols=Colours,main="ABC",ylim=c(-0.1,10))

# Weighted similarity-based clustering

w=WeightedClust(L, type='data',distmeasure=c("euclidean","euclidean"))
ComparePlot(list(w),nrclusters=4,
cols=Colours,fusionsLog=TRUE,margins=c(8.1,10.1,4.1,4.1),names=c('mRNA
only','0.9+0.1','0.8+0.2','0.7+0.3','0.6+0.4','0.5+0.5','0.4+0.6','0.3+0.7','0.2+0.8','0.1+0.9','prote
ins only'), weightclust=F,plottype="new",location=NULL)

plot(w$Clust$Clust, main='weights=0.5')

# CEC method

CEC_2_10_1=CEC(List=L,distmeasure=c("euclidean","euclidean"),normalize=c(F, F)
           ,t=100, r=c(100,100), nrclusters=list(seq(2,10,1),seq(2,10,1)),clust="agnes",linkage=
             c("ave","ave"),weightclust=0.5)


ClusterPlot(Data1=CEC_2_10_1$Clust, nrclusters=4,cols=Colours,ylim=c(-0.1,10))
ClusterPlot(Data1=CEC_2_10_1$Clust, nrclusters=2,cols=Colours,ylim=c(-0.1,10))



############   Reading data of hippocampus, day 11 ##################
```

```
hD11_mRNA_data = read.csv("D11_mRNA_Hippocampus_normalized_tmm.csv",
row.names = 1, check.names = FALSE)
hD11_Protein_data =read.csv("D11_Protein_Hippocampus_normalized_quant.csv",
row.names = 1,  check.names = FALSE)

# Finding optimal number of clusters with help of Gap statistic
res_clust_mRNA11=gap(t(hD11_mRNA_data),
cl.method='hclust',dist.method='euclidean',linkage='ave',Kmax=15,B=100)
res_clust_mRNA11

res_clust_prot11=gap(t(hD11_Protein_data),
cl.method='hclust',dist.method='euclidean',linkage='ave',Kmax=15,B=100)
res_clust_prot11


################# Hippocampus day 11 gene expression ########################


D11Hg<-as.data.frame(hD11_mRNA_data)

# Hierarchical clustering

D11Hg_clust=Cluster(Data=t(D11Hg), type='data', distmeasure='euclidean')
ClusterPlot(Data1=D11Hg_clust, nrclusters=4,cols=Colours,main="xydat.a",ylim=c(-0.1,10))

# Single-source ABC clustering

#K=4
hmRNA_11_ABC=ABC.SingleInMultiple(data=t(hD11_mRNA_data),distmeasure="euclidean
",
                         weighting=TRUE,stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=100,linkage="flexible",
                         alpha=0.625,NC=4, NC2=NULL, mds=FALSE)
hmRNA_11_ABC_clust=Cluster(Data=t(hmRNA_11_ABC), type='data',
distmeasure='euclidean')
ClusterPlot(Data1=hmRNA_11_ABC_clust,
nrclusters=4,cols=Colours,main="ABC",ylim=c(-0.1,10))

#K=2
hmRNA_11_ABC=ABC.SingleInMultiple(data=t(hD11_mRNA_data),distmeasure="euclidean
",
                         weighting=TRUE,stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=100,linkage="flexible",
                         alpha=0.625,NC=2, NC2=NULL, mds=FALSE)
hmRNA_11_ABC_clust=Cluster(Data=t(hmRNA_11_ABC), type='data',
distmeasure='euclidean')
ClusterPlot(Data1=hmRNA_11_ABC_clust,
nrclusters=2,cols=Colours,main="ABC",ylim=c(-0.1,10))
```

```
########################## Hippocampus day 11 proteins ########################

#Reading data

D11Hp<-as.data.frame(hD11_Protein_data)

# Hierarchical clustering

D11Hp_clust=Cluster(Data=t(D11Hp), type='data', distmeasure='euclidean')
ClusterPlot(Data1=D11Hp_clust, nrclusters=4,cols=Colours,main="xydat.a",ylim=c(-0.1,10))

# Single-source ABC clustering


#K=4
hProtein_11_ABC=ABC.SingleInMultiple(data=t(hD11_Protein_data),distmeasure="euclidea
n",
                    weighting=TRUE,stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=100,linkage="flexible",
                    alpha=0.625,NC=4, NC2=NULL, mds=FALSE)
hProtein_11_ABC_clust=Cluster(Data=t(hProtein_11_ABC), type='data',
distmeasure='euclidean')
ClusterPlot(Data1=hProtein_11_ABC_clust,
nrclusters=4,cols=Colours,main="ABC",ylim=c(-0.1,10))

#K=2
hProtein_11_ABC=ABC.SingleInMultiple(data=t(hD11_Protein_data),distmeasure="euclidea
n",
                    weighting=TRUE,stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=100,linkage="flexible",
                    alpha=0.625,NC=2, NC2=NULL, mds=FALSE)
hProtein_11_ABC_clust=Cluster(Data=t(hProtein_11_ABC), type='data',
distmeasure='euclidean')
ClusterPlot(Data1=hProtein_11_ABC_clust,
nrclusters=2,cols=Colours,main="ABC",ylim=c(-0.1,10))

#################### Hippocampus day 11 integrative ##########################
L=list(t(D11Hg), t(D11Hp))

# M_ABC method

set.seed(33)
MABC=M_ABC(List=L,distmeasure=c("euclidean","euclidean"),
      weighting=c(TRUE, TRUE),stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=c(100,100),linkage=c("flexible","flexible"),
      alpha=0.625,NC=2, NC2=NULL, mds=FALSE)
```

```
ClusterPlot(Data1=MABC, nrclusters=2,cols=Colours,main="M_ABC",ylim=c(-0.1,10))

set.seed(34)
MABC=M_ABC(List=L,distmeasure=c("euclidean","euclidean"),
        weighting=c(TRUE, TRUE),stat="var", gr=c(),bag=TRUE,
numsim=1000,numvar=c(100,100),linkage=c("flexible","flexible"),
        alpha=0.625,NC=4, NC2=NULL, mds=FALSE)
ClusterPlot(Data1=MABC, nrclusters=4,cols=Colours,main="M_ABC",ylim=c(-0.1,10))

# ADC method

D11H_ADC=ADC(L, distmeasure = "euclidean", normalize = FALSE, method = NULL,
         clust = "agnes", linkage = "ave", alpha = 0.625)
ClusterPlot(Data1=D11H_ADC, nrclusters=4,cols=Colours,main="ABC",ylim=c(-0.1,10))

# Weighted similarity-based clustering

w=WeightedClust(L, type='data',distmeasure=c("euclidean","euclidean"))
ComparePlot(list(w),nrclusters=4,
cols=Colours,fusionsLog=TRUE,margins=c(8.1,10.1,4.1,4.1),names=c('mRNA
only','0.9+0.1','0.8+0.2','0.7+0.3','0.6+0.4','0.5+0.5','0.4+0.6','0.3+0.7','0.2+0.8','0.1+0.9','prote
ins only'), weightclust=F,plottype="new",location=NULL)

plot(w$Clust$Clust, main='weights=0.5')

# CEC method

CEC_2_10_1=CEC(List=L,distmeasure=c("euclidean","euclidean"),normalize=c(F, F)
          ,t=100, r=c(100,100),
nrclusters=list(seq(2,10,1),seq(2,10,1)),clust="agnes",linkage=
            c("ave","ave"),weightclust=0.5)


ClusterPlot(Data1=CEC_2_10_1$Clust, nrclusters=4,cols=Colours,ylim=c(-0.1,10))
ClusterPlot(Data1=CEC_2_10_1$Clust, nrclusters=2,cols=Colours,ylim=c(-0.1,10))
```