Bettina Pucher, BSc

Comparison of Integrative Analysis Methods based on Simulated and Biological Data Sets

MASTER'S THESIS

to achieve the university degree of

Master of Science

Master's degree programme: Biomedical Engineering

submitted to

Graz University of Technology

Supervisors

Dr. Gerhard Thallinger

Dr.techn. Oana Alina Zeleznik, MSc

Computational Biotechnology and Bioinformatics
Institute of Molecular Biotechnology

Graz, October 2015

AFFIDAVIT[1]

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Graz, ____________________________ ________________________________

Date Signature

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Abstract

Integrative analysis methods have become essential tools for the extraction of a small set of features which is assumed to be the driver of the measured molecular-biological processes.
Objectives: In the presented master's thesis, three integrative analysis methods based on different mathematical concepts are compared: sparse Canonical Correlation Analysis (sCCA), Non-Negative Matrix Factorization (NMF) and Microarray Logic Analyzer (MALA). They are applied on synthetic data as well as on biological breast cancer data derived on three different levels: the DNA level, the transcript level, and the protein level.
Methods: The resulting sets of selected features are compared with each other directly as well as on a more general level, the associated Gene Ontology (GO) terms. Additionally, the sets of selected features in the biological datasets are compared to genes known to be involved in cancer development.
Results: The observed overlap on the feature level is modest in both the synthetic and the biological datasets. Considering the associated GO terms, the overlap increases in at least one GO category for both datasets. The features selected from the biological dataset by each of the three methods cover about 10% of the features involved in pathways in cancer according to the KEGG database.
Conclusion: The results of integrative analysis of biological data can hardly be validated, however, they can be compared to the results of other integrative analysis methods. The feature sets resulting from the methods under comparison are not congruent. A better agreement between the results can be observed on a higher functional level, the GO term level.

Contents

1. Introduction

1.1. Motivation

The central dogma of molecular biology [1] describes the relationship and the flow of information between DNA, RNA and proteins and shows that there are multiple interacting levels of information in the cells of an organism. The status of an organism's cell on gene or DNA level is investigated in the field of genomics; investigations on transcript or RNA level are summarized under the term transcriptomics; examining the amount of proteins, that is the analysis on protein level is termed proteomics and the examination of all metabolites present in a cell is referred to as metabolomics. In recent years, due to technological improvements, large amounts of data have been obtained employing high-throughput technologies. These methods enable highly parallel measurements on different biological levels on the same set of samples. The challenge has been shifted from obtaining data towards extracting useful information from it.

A major goal in bioinformatics is the identification of features associated with complex diseases such as diabetes mellitus, breast cancer or Alzheimer's disease which are caused by multiple genetic, environmental and lifestyle factors [2]. The task of being able to classify a sample as case or control boils down to the identification of a preferably small number of genes of an organism's genome that show strong evidence to be associated with a certain disease. The selection of candidate features which are subsequently subjected to further analysis in the wet lab implicates a tremendous reduction of time and money costs compared to the analysis of the whole feature set. For this purpose data is obtained from different biological levels to provide a comprehensive view on the system under study and it is intuitive that the amount of gained information is greater resulting from joint analysis than from the individual analysis of datasets.

1.2. Integrative Analysis

In the past two decades the focus of methods applicable to large biological datasets has been on analysis of data from one single biological level such as the analysis of all transcripts or all proteins in a sample at a time. In recent years the integration of two or more omics-datasets measured on different levels on the same set of samples or measured at different time points or conditions in two different organisms has become more and more important. Their simultaneous mutually dependent analysis is summarized under the term integrative analysis in contrast to the mutually independent analysis of datasets and the combination of individual results termed meta-analysis.

Considering integrative analysis methods one has to distinguish between those which reveal specific or common structures within datasets respectively and those which incorporate a feature selection step and result in a short list of candidate genes to be subjected to further experimental analysis. The reduction of the feature set size is accomplished by various approaches and combinations of them. One important characteristic of large, genome-scale datasets is that the number of features comprised by the dataset usually far exceeds the number of observations and the number of features is further increased by the simultaneous analysis of two or more data sets measured on the same small set of samples. Various approaches have been described in the literature so far that aim to overcome the issue that an under-determined system of equations due to the small number of samples does not have a unique solution and that the resulting set of candidate features contributing to a disease is desired to be rather small. They are based on widespread well-known mathematical concepts adapted for example by inducing sparseness or incorporating heuristic or machine learning approaches into the feature selection step. Some examples of the basic concepts of integrative analysis of datasets are summarized in the following section.

1.3. Literature Review

As a preparatory work for the selection of methods to compare, a literature search on integrative analysis methods for biological datasets with focus on genomic, transcriptomic and proteomic data was conducted. The goal was to review the mathematical concepts of integrative analysis and to provide an overview of methods currently in use. The methods under review are grouped according to their basic concepts into eigenvalue decomposition based methods, regression based methods, clustering based methods and machine learning based methods and are summarized in the following subsections.

1.3.1. Decomposition Based Methods

Several methods described in the literature are based on eigenvalue decomposition but there are also other factorization approaches.
Canonical Correlation Analysis (CCA), Co-Inertia Analysis (CIA), Pseudoinverse Projection (PIP) and General Singular Value Decomposition (GSVD) have in common that some product of the data matrices to be analyzed must be decomposed into its eigenvectors and eigenvalues. Alter and Golub [3] showed that PIP [4] of an arbitrary number of datasets represents a linear transformation into a space spanned by a *basis* set of samples. Each sample can be approximated by a linear combination of the basis samples. Unknown regulatory dependencies may manifest as correlations between samples of the datasets and the basis samples.

Berger *et al.* [5] presented an iterative algorithm for dimension reduction based on GSVD [6] and applied it to gene expression and copy number variation data. In an iterative *steerable gene shaving* process the genes with the highest variance in the datasets are identified. In each iteration, a matrix X containing the generalized singular vectors of the dataset pair on the columns is calculated and the angular distances between the samples of the datasets and the columns of X are determined. The datasets are projected onto the column of X corresponding to the largest angular distance and the genes with the least parallel contribution are shaved off. The steps are repeated until the number of features remaining in the datasets falls below a desired number. Ponnapalli *et al.* [7] developed a higher-order GSVD (HO GSVD), an extended version of conventional GSVD applicable to more than two datasets. They analyzed the genome-scale expression datasets of three organisms in order to reveal structural or functional motifs common to all datasets. The data matrices are decomposed into three factors whereat one of them is identical in all decompositions. The common factor matrix contains the right basis vectors obtained from an eigensystem involving the arithmetic mean of the pairwise combination of all data matrices. The significance of each right basis vector, in other words the amount of information contributed to each of the datasets is indicated by the *higher-order generalized singular value set* associated with each basis vector. Information that is common to all datasets is represented by basis vectors with equal significance. The right basis vectors corresponding to the eigenvalues equal to one determine a common subspace of the HO GSVD.

Many integrative approaches employ in some form the CCA introduced by Hotelling [8]. Conventional CCA maximizes the correlation of the projections of two datasets and is not suitable for integration of more than two datasets. Many groups have made efforts to extend CCA to the application on more than two datasets. Lê Cao *et al.* [9] recalled a regularized variant of CCA described in detail in [10] which used *Elastic Net* [11] penalization, a combination of *lasso* and *ridge* penalties. To obtain unique canonical factors in case the number of features in the datasets exceeds the number of samples, additional information has to be introduced. This could be of the form that the vector containing the weights of the decomposition is subjected to a penalty. Different penalties have been employed such as *ridge*, where the L_2-norm of the vector is bound or *lasso*, that limits the sum of absolute values of the elements of the vector to a given constant, resulting in a sparse vector [11].

Witten *et al.* [12] presented a penalized matrix decomposition (PMA) method which can be used to obtain sparse principle components as well as sparse canonical vectors when applied to the product of two matrices. It is basically a regularized singular value decomposition (SVD) where a given matrix is decomposed into sparse vectors. The vectors are subjected to either *lasso* or *fused lasso* penalties, depending on the appearance of the data.

The non-zero weights of the sparse vectors are associated to features with large influence on the correlation. The sparse version of CCA (sCCA) was further extended by Lin *et al.* [13] who took into account the structure or *group effect* within genomic data for example genes within the same pathway. They developed a method based on the block cyclic coordinate descent algorithm [14] in order to solve the optimization problem which incorporates sparse group lasso penalty.

Another dimension reduction method, the co-inertia analysis (CIA) [15] was applied by Fagan *et al.* [16] for integrative analysis of two datasets, however the authors pointed out that the method is suitable for the analysis of any number of datasets. CIA aims at finding major directions or *axes* of the datasets having maximum covariance. The axis can be obtained by various standard multivariate analysis techniques such as principle components analysis (PCA) or correspondence analysis (CA). The axis pairs with the largest covariance are supposed to represent common themes within the two datasets. GO information was used as a supplement to facilitate biological interpretation. Actual feature selection is not part of the CIA, though the elements of the weight vector of the dimension reduction procedure can be ordered and the features corresponding to the top weights are selected [9].

A method for the extraction of *relevant biological correlations* based on non-negative matrix factorization (NMF) in the form described by Lee and Seung [17] is presented by Brunet *et al.* [18]. They approximate the expression profiles of all genes in a datasets as decomposition into a small number of metagenes and a weight-matrix. The samples can be clustered based on the expression patterns of the metagenes. They also propose a criterion for model selection to determine the number of metagenes used for the decomposition. Zhang *et al.* [19] employed NMF for the factorization of more than one dataset at a time into a matrix containing the shared *building blocks* and a weight-matrix for each of the datasets. The factor-matrices are determined in an iterative update process minimizing the approximation error given a predefined number of building blocks. The method aims at the identification of multi-dimensional modules which are represented by features of all datasets that show similar profiles across all or a subset of samples.

1.3.2. Regression Based Methods

The term regression analysis refers to the process of determining a model describing a relationship between a given set of data points [20]. A simple example for a regression model is a line and the fitting process is called linear regression. The correlation between two sets of OMICS data can be assessed with linear regression analysis as shown in [21]. The global correlation of two genome-scale datasets is usually close to zero, which means there is hardly any correlation. In advance to the actual analysis the data has to be transformed since the data usually is not normally distributed, otherwise the significance of the correlation might not be estimated correctly [21].

Partial least squares regression (PLS) allows to retrieve major driving factors in the datasets referred to as latent variables by maximizing the covariance between lower dimensional projections of the datasets. Lê Cao [22] present a sparse version of PLS, an iterative algorithm which is based on sparse SVD [23] introducing a soft-thresholding penalization on the PLS loading vectors of each dataset. The method is demonstrated by applying it on two datasets measured on the same samples. According to the author results obtained with the presented approach are more promising compared to classical PLS.

1.3.3. Clustering Based Methods

Cluster analysis aims at grouping objects according to some similarity measure [24]. Shen and colleagues [25] present an integrative clustering approach used for tumor subtype discovery. The method is applicable to an arbitrary number of datasets of different types and aims at determining latent variables representing disease driving factors responsible for disease-subtypes. They use an integrative model named iCluster which was introduced earlier [26]. In the so called loading matrix the coefficients of features are subjected to some penalty term and those which do not contribute any information converge to zero. The original datasets can be approximated using the identified variables which are common to all data types.

Another method which results in a list of candidate genes is presented by Cao [27] and is called sparse representation based clustering (SRC). The feature vectors within the dataset are assigned to a predefined number of clusters represented by sparse vectors. Membership of a feature vector to a cluster is determined by the smallest distance to the group vector employing the angle and the difference in length (L2-norm of the vectors) between the feature and the group vector. After the clustering of all feature vectors, a significance measure is used to select candidate features from the groups.

Gusenleitner *et al.* [28] introduce iterative Binary Bi-clustering of Gene sets (iBBiG) where they apply bi-clustering to the results of gene set analysis (GSA) of multiple genome-scale datasets. The method identifies clusters or *modules* by grouping samples with gene expression profiles overrepresented in the same gene sets. The gene sets within a cluster as well as the clusters themselves are ranked according to their homogeneity or their information score respectively. Applied on breast cancer datasets, the majority of clusters found could be associated to molecular subtypes. A mentionable advantage of iBBiG is that the number of clusters is not required to be specified in advance.

1.3.4. Machine Learning Based methods

Machine learning algorithms are used to infer a model from parts of a given dataset (trainings set) that is capable to predict/describe the pattern of the remaining data (test set) as well [29]. The Random Forests approach (RF) [30] can be used to classify samples of a dataset by the aid of classification trees. For appropriate classification of samples the importance of features of either data type is estimated and can thus be employed for feature selection. Reif *et.al.* [31] applied RF to combined genetic and proteomic data and asserted that the combinatorial approach yields more promising results in selecting relevant features for complex disease models than the individual analysis of large-scale datasets.

Weitschek *et al.* [32] presented a tool able to classify microarray experiments (samples). The method comprises three major steps: discretization, feature selection and formula extraction resulting in a set of logic formulas connecting features in conjunctive and disjunctive normal form respectively. Originally developed for microarray data analysis the extension to other genome-scale datasets is straight forward.

1.4. Objectives

One of the major goals of integrative analysis methods in bioinformatics is the identification of a small number of genes or other features evident to contribute to the development of diseases. The presented thesis focuses on methods which actually analyze two or more data sets at the same time, rather than methods where the result of the analysis of one dataset serves as additional information to the second dataset. The focus is explicitly not on the integration of meta-information available for samples. The methods under comparison already involve a feature selection step and result in a flat list of candidate genes.
For the comparison, three integrative methods have been chosen which are based on complementary mathematical concepts compared to the methods reviewed in the recent work of Tomescu *et al.* [33] who compared co-inertia analysis, general singular value decomposition and integrative biclustering. The methods were selected due to three criteria: i) they are based on different mathematical concepts, ii) they are suitable for the extension to an arbitrary number of genome-scale datasets and iii) access to the software implementation is provided.

The specific goals of this thesis are:

- comprehensive understanding of the methods under comparison:
    - sparse Canonical Correlation Analysis
    - Non-Negative Matrix Factorization
    - Microarray Logic Analyzer
- set-up of a software environment as an interface to the pre-implemented methods
- extraction of co-expression networks serving as basis for synthetic data
- synthetic data generation with the tool SynTReN
- application of methods on synthetic and biological data
- comparison of flat lists of candidate genes resulting from each method
- analysis of gene ontology terms associated with candidate genes

The set of candidate features resulting from each method might not be very congruent at the most specific level, the gene level. In order to discover redundancies in the results of the three methods, the lists of candidates are compared on a more general level, the associated Gene Ontology (GO) [34] terms. The comparison is expected to allow inferring an answer to questions like: How big is the overlap of gene lists resulting from each method? How big is the overlap of GO terms?

We hypothesize that the overlap produced by the gene lists resulting from sCCA and the NMF will be grater than the overlap of either of these lists with the genes in the logic formulas resulting from MALA. This is expected because sCCA and NMF both are applied on three datasets of tumor samples and aim to find the similarities, while MALA is applied on sets of tumor and normal samples and aims to discover the differences between them.

2. Methods

In this chapter a comprehensive description of the materials, methods and tools used to accomplish the comparison of three integrative analysis methods is provided. The first part of the chapter deals with the mathematical concepts of the methods; the second part focuses on the description of the structure and origin of the datasets the methods were applied on.

2.1. Sparse Canonical Correlation Analysis

Sparse canonical correlation analysis (sCCA) represents a sparse version of the standard Canonical Correlation Analysis (CCA) [8] which maximizes the correlation of the projections of two datasets in a common space of reduced dimension. CCA has been applied in various contexts to retrieve associations between two datasets by finding projections of them that retain as much information as possible and at the same time maximize the linear association between the projections. The determination of vectors containing the weights for the linear combination of the original variables (canonical weights) involves finding the eigenvalues and the corresponding eigenvectors of the product of the covariance matrices of the datasets. There is an exact solution of CCA for two datasets in case the number of observations (samples) is greater than the number of variables (features) of either dataset. In case the number of variables exceeds the number of observations, the vectors containing the canonical weights used in the projection are not unique.

The sCCA approach employed for the method comparison was presented by Witten *et al.* who showed that sCCA can be reformulated as a penalized matrix decomposition (PMD) problem [12]. Moreover, with PMD a sCCA of multiple datasets can be accomplished [35]. The steps of sCCA via PMD are summarized below.

Datasets are given as matrices with the samples in the rows and the features in the columns. The columns are standardized to have mean equal to zero and standard deviation (SD) equal to one. A matrix X can be represented as the product of its eigenvalues d_k and left- and right-eigenvectors u_k and v_k respectively. This is known as the singular value decomposition (SVD) of a matrix. The best rank-r approximation $\hat{X}$ of X in the sense of the squared Frobenius norm

$$\|X - \hat{X}\|_F^2 = \sum_i \sum_j |x_{ij} - \hat{x}_{ij}|^2 \tag{2.1}$$

involves the r largest eigenvalues and their corresponding eigenvectors:

$$\hat{X} = \sum_{k=1}^{r} d_k u_k v_k^T \tag{2.2}$$

Correspondingly, the approximation of the product of two matrices X and Y that maximizes the correlation involves the largest eigenvalues and the corresponding eigenvectors of the matrix-product (see equation 2.3). It was shown that in the rank-1 approximation the left- and right-eigenvector corresponding to the largest eigenvalue used in the approximation are equal to the canonical weight-vectors u and v of the one-dimensional projection of the two data matrices resulting from conventional CCA [12]. However, these vectors are not unique if the number of features exceeds the number of samples.

$$\begin{gathered} \max_{u,v} cor(Xu, Yv) \text{ is equal to} \\ \max_{u,v} u^T X^T Y v \text{ subject to } u^T X^T X u \leq 1, v^T Y^T Y v \leq 1 \end{gathered} \tag{2.3}$$

Introducing Sparseness

Witten *et al.* subject the vectors u and v in the decomposition of the matrix-product to constraints (PMD) which results in a unique and sparse solution [12]. Additionally, they substitute X^TX and Y^TY with the identity matrix I. This results in the sCCA criterion for two datasets X and Y:

$$\max_{u,v} u^T X^T Y v \text{ subject to } \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1, P_1(u) \leq c_1, P_2(v) \leq c_2 \tag{2.4}$$

with penalty functions P_i and tuning parameters c_i chosen appropriately. For the data used within the presented thesis P is always the L_1 penalty also referred to as *lasso* penalty.

Assuming we have K datasets containing measurements of different sets of features p_k on a shared set of samples n, a generalized form of PMD is applied and can be formulated as

$$\sum_{i<j} w_i^T X_i^T X_j w_j \text{ subject to } w_k^T X_k^T X_k w_k = 1 \quad \forall k \text{ and } w_k \in \mathbb{R}^{p_k}. \tag{2.5}$$

The K canonical weight-vectors w are obtained by solving the *sparse multiple CCA* criterion:

$$\max_{w_1,\ldots w_K} \sum_{i<j} w_i^T X_i^T X_j w_j \text{ subject to } \|w_i\|_2^2 \leq 1, P_i(w_i) \leq c_i \forall i. \tag{2.6}$$

The canonical weights are determined in an iterative approach where w_i is updated in each iteration until convergence:

$$w_i \leftarrow \operatorname{argmax}_{w_i} w_i^T X_i^T \left(\sum_{j \neq i} X_j w_j \right) \text{ subject to } \|w_i\|^2 \leq 1, P_i(w_i) \leq c_i. \tag{2.7}$$

The L_1-penalty on a real vector w of length p is defined as:

$$P(w) = \|w\|_1 = \sum_{i=1}^{p} |w_i|. \tag{2.8}$$

and w will be sparse if $1 \leq c \leq \sqrt{p}$.

The resulting canonical weight-vectors are unique and sparse for appropriate penalty-functions P_i and tuning parameters c_i.

Parameter Selection

In a permutation framework, sets of tuning parameters are tested to assess the significance of the canonical weight-vectors and to determine the best set of tuning parameters $c_{1\ldots K}$. For a given K-dimensional set of tuning parameters the canonical weight-vectors w_i and the corresponding projections are calculated for the original datasets as well as for a number of datasets with randomly permuted samples. As test statistic the sum of pairwise correlations between the projections of the datasets is used. The z-score, which is the standardized test statistic and the p-value are determined. The p-value is given by the fraction of projections of permuted datasets that results in a larger value of the test-statistic than the projections of the original datasets. If there is a significant correlation between features across the original datasets the p-value will be small. The set of tuning parameters corresponding to the highest z-score and the lowest p-value is selected as set of best penalties on the canonical weight-vectors. Alternatively, c_i can be chosen arbitrarily to achieve a certain amount of sparsity. *Sparse* means that many elements in the weight-vectors are equal to zero. The non-zero weights indicate correlated features across datasets. These features can be considered as candidates to be associated with certain attributes shared by the samples in the datasets. The zero-weighted features in the projection of the dataset are thereby assumed to be not as important as to contribute to the inherent structure of the data set.

The number of permutations is set to 100 (default 25). The number of permutations was increased, because the SD of the test statistic is estimated from the permutations. The parameter *type*, is set to *standard* because the features in the datasets are not ordered. As a result, a *lasso* penalty is applied on the canonical weight-vectors. Other parameter settings are left to the default values. The sets of tuning parameters tested and the corresponding statistics for the synthetic datasets are listed in Table 2.1. Additionally, the calculated correlations and z-scores of the tested sets are visualized in Figure 2.1. Employing the set of tuning parameters corresponding to the highest z-score and the lowest p-value, which is highlighted in Table 2.1, the canonical weight-vectors resulting from sCCA comprise 32, 134 and 3 non-zero elements respectively. Since the number of selected features in each datasets is desired not to notedly exceed 5% of the total number of features in that datasets, the tuning parameters are decreased to the values in Table 2.2 iteratively.

Table 2.1.: Sets of tuning parameters for synthetic datasets tested in a permutation framework; c1, c2 and c3 represent the penalties on the canonical weight-vectors for the gene expression, the DNA-methylation and the protein expression datasets respectively.

index	c1	c2	c3	p-value	z-score
1	1.97	4.97	1.10	0.15	0.778
2	3.51	8.84	1.47	0.13	0.867
3	5.05	12.70	2.11	0.68	-0.335
4	6.58	16.57	2.75	0.94	-1.348
5	8.12	20.44	3.39	0.91	-1.134
6	9.65	24.30	4.03	0.88	-1.078
7	11.19	28.17	4.67	0.91	-1.043
8	12.73	32.03	5.31	0.92	-0.971
9	14.26	35.90	5.96	0.90	-0.947
10	15.80	39.77	6.60	0.89	-0.949

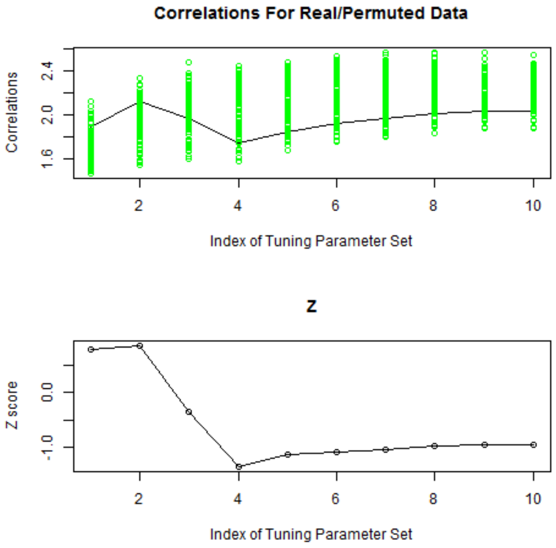

Figure 2.1.: Statistics of tuning parameter sets tested for synthetic datasets.

Table 2.2.: Set of tuning parameters for synthetic datasets adjusted in an adaptive process; c1, c2 and c3 represent the penalties on the canonical weight-vectors for the gene expression, the DNA-methylation and the protein expression datasets respectively.

c1	c2	c3
2.73	8.18	1.47

For the biological datasets, the calculated correlations and the z-scores of tuning parameter sets tested in the permutation test are depicted in Figure 2.2. The set of best penalties which is highlighted in Table 2.3 results in canonical weight-vectors with 8 242, 5 361 and 45 non-zero weights for the biological gene expression, DNA-methylation and protein expression datasets respectively. The penalties on the canonical weight-vectors are hence iteratively decreased to the values in Table 2.4 to reduce the number of selected features to about 5% of the total number of features in the datasets.

Table 2.3.: Sets of tuning parameters for biological datasets tested in a permutation framework; c1, c2 and c3 represent the penalties on the canonical weight-vectors for the gene expression, the DNA-methylation and the protein expression datasets respectively.

index	c1	c2	c3	p-value	z-score
1	14.06	11.67	1.10	0.14	0.944
2	25.00	20.75	1.93	0.07	1.421
3	35.93	29.83	2.78	0.02	1.941
4	46.87	38.91	3.62	0.01	2.641
5	57.80	47.99	4.47	0.00	3.099
6	68.74	57.07	5.31	0.00	3.186
7	79.67	66.15	6.16	0.00	3.156
8	90.61	75.23	7.00	0.00	3.114
9	101.55	84.31	7.85	0.00	3.006
10	112.48	93.38	8.69	0.00	2.830

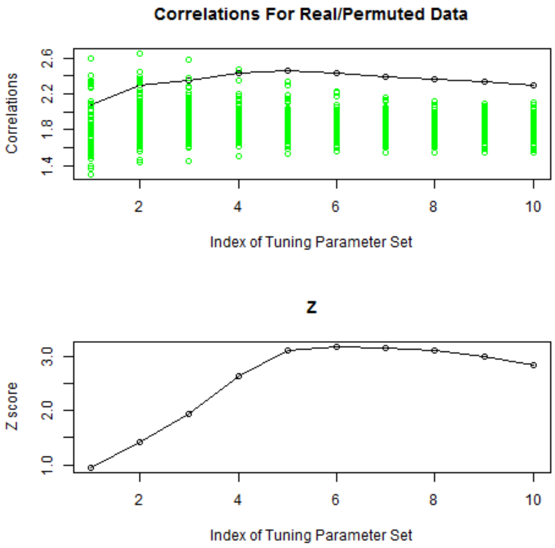

Figure 2.2.: Statistics of tuning parameter sets tested for biological datasets.

Table 2.4.: Set of tuning parameters for biological datasets adjusted in an adaptive process; c1, c2 and c3 represent the penalties on the canonical weight-vectors for the gene expression, the DNA-methylation and the protein expression datasets respectively.

c1	c2	c3
25.78	20.61	1.77

Software

Sparse CCA for multiple datasets using PMD and the permutation framework for tuning parameter selection is available as part of the R-package *PMA* (Penalized Multivariate Analysis) [36].

2.2. Non-Negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) techniques have been described in the literature several times and in various contexts. For example Lee and Seung [17] used NMF to learn the characteristic parts of faces applying NMF on a data set of facial images. Each face in the data set can be approximated by the positively weighted sum of the learned

characteristic parts.
The method employed for the comparison within the scope of this thesis was presented by Zhang and colleagues [19] and aims to find *correlative modules* in multiple genome-scale datasets. These so called multi-dimensional modules (md-modules) are subsets of features within the analyzed datasets that show similar profiles in all or a subset of samples. The large datasets are decomposed into *building blocks* of samples with shared attributes that may reveal the inherent structure of the data. The method is suitable for the simultaneous analysis of an arbitrary number of datasets. Here, a description of the algorithm for the analysis of three datasets is provided.

Given three matrices containing the measurements of a shared set of samples (rows) on a - in general - different number of features (columns). The columns of the matrices are standardized to have mean equal to zero and SD equal to one and the elements of the matrices are scaled so that all matrices have equal Frobenius norm. The method expects input matrices to contain only non-negative elements, hence, according to Kim and Tidor [37] the columns of the matrices were doubled. The first column contains all originally positive elements while the second column contains the absolute value of all originally negative elements. The remaining elements are set to zero. The concept of NMF is based on the fact that a non-negative matrix X of dimension $M \times N$ can be decomposed in two non-negative factor matrices W and H, with $W(M \times K)$ containing the K basis vectors and $H(K \times N)$ containing the K coefficient vectors comprising the weights of the building blocks in W. The columns of X are then approximated by the positively weighted linear combination of the K basis vectors. The weights of the linear combination contained in the matrix H encode for strong or weak presence of the building blocks in the columns of X (features). The matrices W and H are chosen so that they minimize the reconstruction error of the data matrix measured in terms of the squared Frobenius norm:

$$F(W,H) = \|X - WH\|_F^2. \tag{2.9}$$

The joint NMF criterion to determine the best factor matrices W and H_1, H_2 and H_3 in the case of three datasets X_1, X_2 and X_3 measured on the same set of M samples with dimensions $M \times N_1, M \times N_2$ and $M \times N_3$ respectively, can be formulated as

$$\min \sum_{I=1}^{3} \|X_I - WH_I\|_F^2. \tag{2.10}$$

The factorization results in the shared matrix W containing the building blocks common to all datasets and the three different matrices H_1, H_2, H_3 where each row represents a coefficient vector containing the weights of the building blocks.
The matrices W and H are randomly initialized and to minimize the joint reconstruction error in equation 2.10 they are iteratively computed using multiplicative update rules. By this procedure only a local minimum of the objective function is found and thus, the calculation of the factor matrices has to be repeated starting from different random initializations and choosing those which result in the smallest reconstruction error.

Discovery of Multi-Dimensional Modules

To determine membership of a feature in a md-module the coefficient matrices H_1, H_2, H_3 can be used. For this purpose the z-score for each element in the rows of H is calculated as:

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{2.11}$$

where μ_i is the median and σ_i is the median absolute deviation (MAD) of the elements in the i-th row of H. Similarly, the z-score for the elements in the columns of W can be calculated using the median and the MAD of the elements in the columns of W. A feature is assigned to a md-module if the z-score is greater than a given threshold.

Parameter Selection

The number of building blocks, which is at the same time the number of resulting md-modules is problem-dependent and is usually chosen to be $K < min(M, N_i)$. However, since the method aims to reduce the complexity of the data, the number of building blocks K is in general desired to be rather small. Additionally, the choice of the number of building blocks K is suggested to be based on three empirical factors: the trend of the reconstruction error changing with the number of building blocks, the rate of significant vertical correlations within md-modules and the significance of an enrichment analysis of modules. Due to reasons of time only the trend of the reconstruction error was used. In the sense of Kim and Tidor [37], the reconstruction error resulting from NMF is compared with the reconstruction error resulting from the SVD of random datasets with elements stemming from the same distribution as the original ones. The reconstruction error of a dataset is defined as the sum of squared differences between the elements of the original and the reconstructed dataset. The percentage of reconstruction error is specified as the reconstruction error related to the sum of squared elements in the original dataset. As suggested by Kim and Tidor [37], the parameter K is selected as the number of building blocks where the absolute value of the slope of the reconstruction error of the NMF of the original dataset turns lower or equal to the slope of the reconstruction error of the SVD of the random datasets.

The plots of reconstruction errors for each of the synthetic datasets are displayed in Figure 2.3.

The number of building blocks K used in the NMF of the synthetic datasets was chosen to be 5, which is the average number of building blocks derived from the slope of reconstruction errors in Figure 2.3. The plots of reconstruction errors for the biological datasets are shown in Figure 2.4.

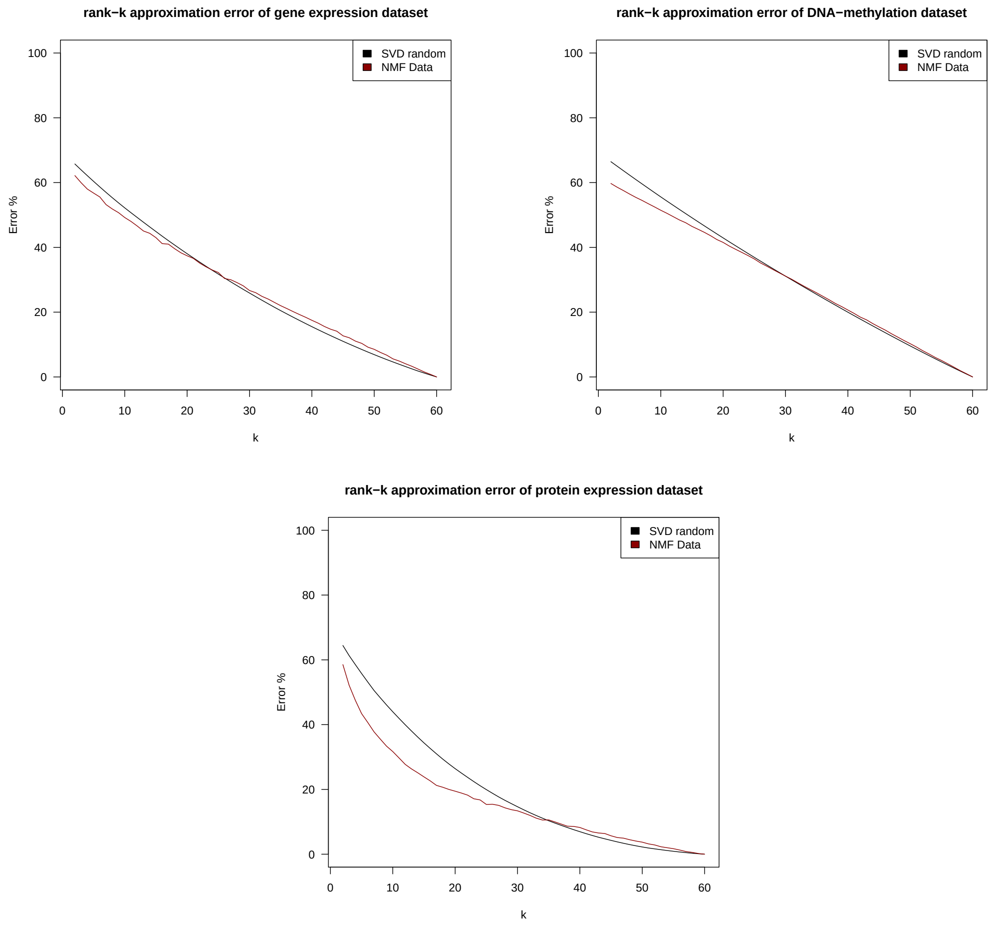

Figure 2.3.: Comparison of reconstruction error of NMF of original datasets and SVD of random datasets for synthetic data.

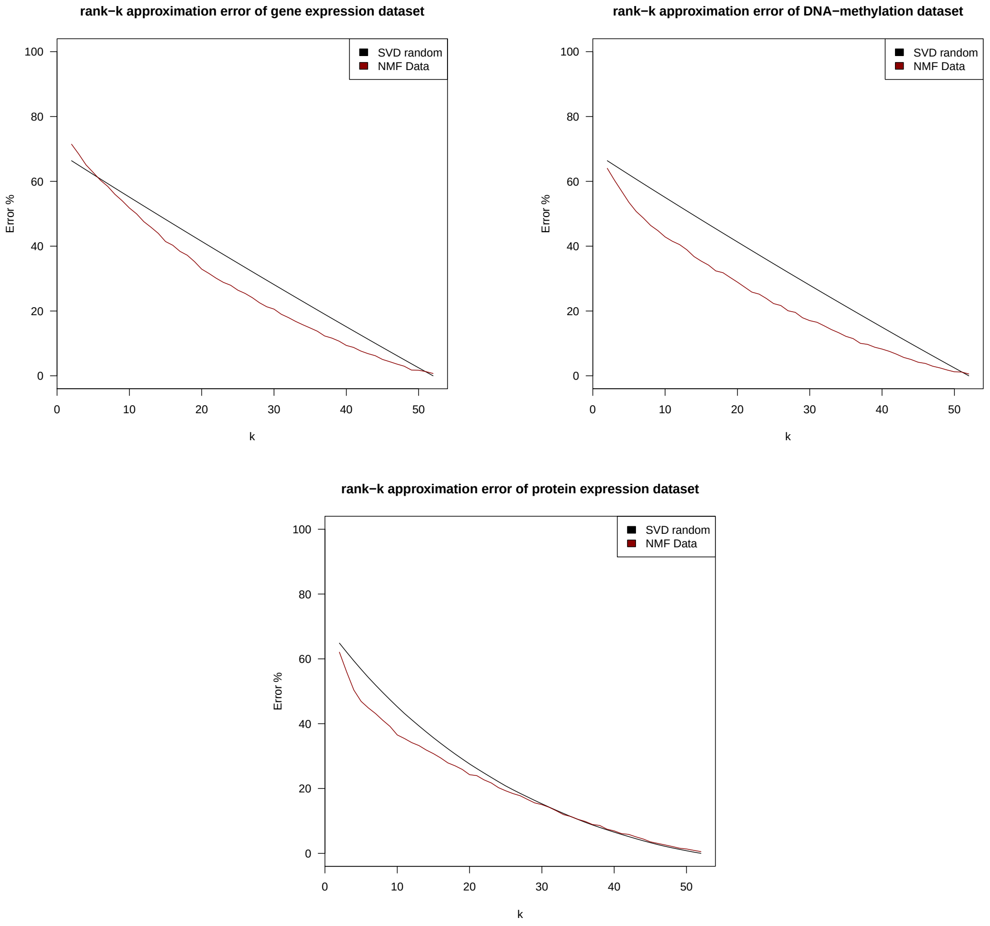

Figure 2.4.: Comparison of reconstruction errors of NMF of original datasets and SVD of random datasets for biological data.

The average number of building blocks derived from Figure 2.4 is 15, which was used as K in the NMF of the biological datasets.

For the choice of a suitable threshold T used to a assign the features/samples to a md-module, the consideration of the fold-change of the enrichment ratio of the gene module within a md-module and randomly constructed modules for a range of thresholds is suggested. The enrichment of gene modules is determined as the number of significantly over-represented (p-value lower than 0.05) GO biological processes associated with the genes comprised by the modules. The number of genes assigned to a module depends on the selected threshold. The enrichment ratio of a module is obtained as the enrichment at a certain threshold in relation to the maximum enrichment. The enrichment ratio is calculated for the gene modules derived by NMF as well as for 100 random modules of the same size. The fold-change of enrichment ratios is obtained as the quotient of enrichment ratios of the original modules and the averaged enrichment ratios of the random modules. High functional homogeneity of a module is indicated by a large fold-change of enrichment ratios.

The choice of parameter T for the NMF of synthetic datasets was based on the analysis of enrichment ratios of derived md-module for T in the range of 2 to 5. At a threshold of 5 or higher, hardly any features are assigned to the md-modules. The enrichment ratios and fold changes of enrichment ratios of gene module 1, as an example for considerable functional homogeneity and module 5, as an example for poor functional homogeneity are illustrated in Figures 2.5 and 2.6. The plots of enrichment ratios and corresponding fold changes of all 5 modules are given in appendix B. According to the enrichment ratios of the gene modules the NMF of synthetic datasets was conducted with parameter T set to 3.

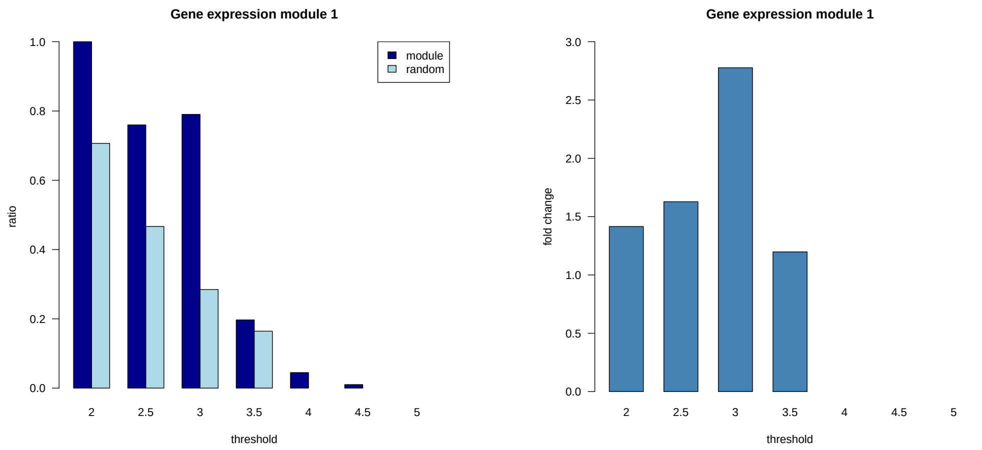

Figure 2.5.: Enrichment ratios and fold change of enrichment ratios of gene module 1 in synthetic data at different thresholds.

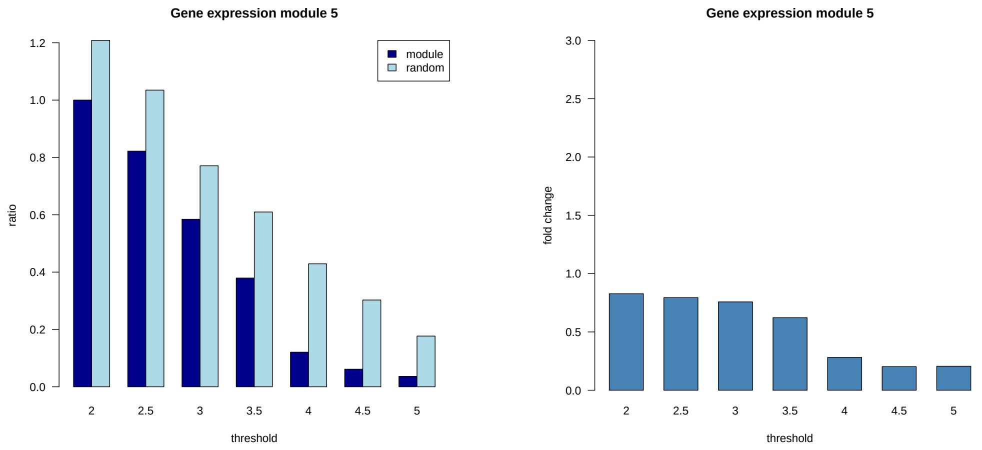

Figure 2.6.: Enrichment ratios and fold change of enrichment ratios of gene module 5 in synthetic data at different thresholds.

The parameter T for the NMF of biological datasets was chosen according to the enrichment ratios and fold change of enrichment ratios of the derived gene modules at values of T in the range of 2 to 6. In the work of Zhang *et al.* [19] the enrichment ratio was assessed for T in the range of 2 to 7 and the best threshold was derived to be at 5. Due to the large effort of time required by the enrichment analysis, the upper limit of T was set to 6. The enrichment ratios and fold changes of enrichment ratios of modules 3, as an example for considerable functional homogeneity and 12, as an example for poor functional homogeneity are illustrated in Figures 2.7 and 2.8. The plots of enrichment ratios and corresponding fold changes of all 15 modules are given in appendix C.

According to the enrichment ratios of the gene modules the NMF of biological datasets was conducted with parameter T set to 4.5.

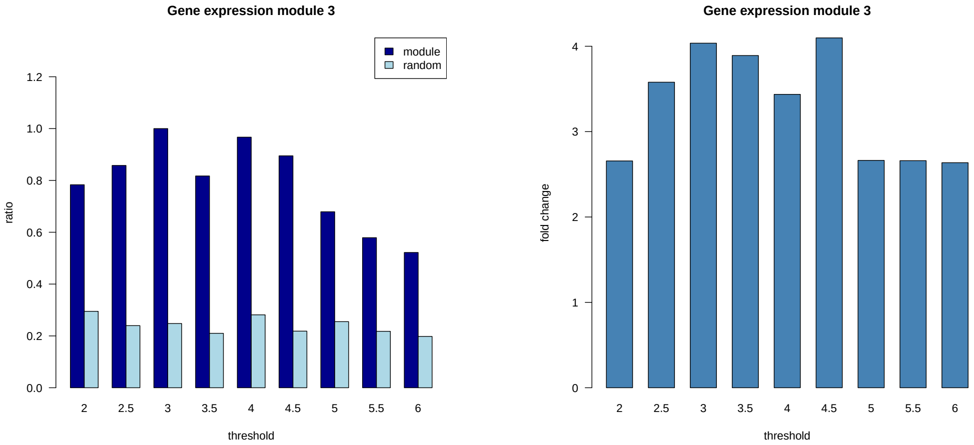

Figure 2.7.: Enrichment ratios and fold change of enrichment ratios of gene module 3 in biological data at different thresholds.

Feature Selection

The identified md-modules are analyzed in regard to their functional homogeneity. The functional homogeneity is assessed by enrichment analysis of GO biological processes. The features comprised by the md-modules with the best functional homogeneity represent the set of selected features. As good functional homogeneity is indicated by a high fold-change of enrichment ratios, module 1 is selected in the synthetic dataset and module 3 is selected in the biological dataset. The features contained in these modules are considered as the resulting set of selected features by the NMF in the synthetic and the biological datasets respectively.

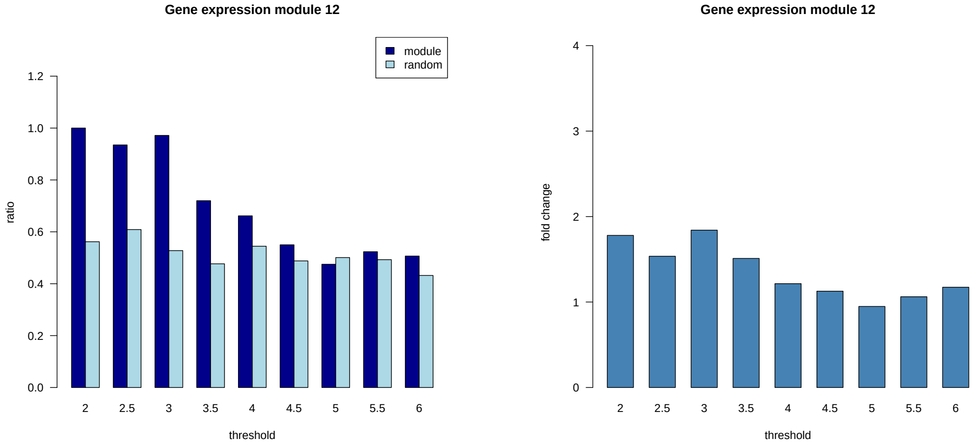

Figure 2.8.: Enrichment ratios and fold change of enrichment ratios of gene module 12 in biological data at different thresholds.

Software

The method has been implemented for the application on three datasets as a Matlab® (Mathworks Inc., Natick, USA) software package and is available as a supplement of the work by Zhang *et. al* [19]. The selection of parameters K and T and the feature selection are not part of the implementation. These were implemented in R as part of this master's thesis.

Parameter Settings

In order to determine a good number of building blocks K, the trend of reconstruction error of NMF was obtained with the following settings: For each K in the range of $2 \leq K \leq min(M, N_i)$, 10 random initializations of W, H_1, H_2 and H_3 are iteratively updated 100 times and the lowest reconstruction error among the 10 runs is reported. The actual calculation of the NMF is conducted with 50 random initializations of the factor matrices and 1000 iterations in each case.

2.3. Microarray Logic Analyzer (MALA)

The application of a logic data mining approach comprising the steps of MALA was described among others by Arisi *et al.* [38]. The MALA software has been developed for the analysis of large microarray gene expression datasets by Weitschek *et al.* [32], however it is suitable for the application on data of any format. To accomplish the integrative analysis of large-scale datasets with MALA, the features of the different data types are concatenated. The two major goals of the method are the gene clustering to reduce the number of features and the classification of samples, that is the differentiation of tissue samples from healthy and ill patients. MALA is based on a machine learning approach and results in a number of *logic formulas* which can be used to classify samples. MALA comprises three major steps: i) the discretization of features and an optional discrete clustering analysis, ii) the selection of the most relevant (clusters) of genes (feature selection), and iii) the assembly of the logic formulas (formula extraction). MALA accepts as input a comma separated value file (*csv*) with the expression profiles of the features in the rows and the individual class-labeled samples in the columns of the dataset. The three analysis steps mentioned above are executed on a subset of samples (training set) and the performance of the classification model is assessed on the remaining samples (test set). The output of MALA comprises a number of files reporting the gene clusters and their sizes in case clustering was done; the classification model as logic formulas; statistical parameters of the model evaluation. For the method comparison, the logic formulas are of special interest, since they consist of the desired candidate features. The steps of the procedure to obtain the logic formulas are summarized in the current section. They are described in detail in [39] and related publications [40, 41].

Discretization and Clustering

The classification algorithm employed by MALA expects the data to be available in a binarized form. Thus, as a first step the features of the datasets have to be discretized [39, 42]. The discretization can be accomplished in two ways, by a supervised or an unsupervised initialization. Owing to the structure of the datasets, the unsupervised discretization is employed: For each feature, a set of equally sized intervals, symmetric around the mean expression level and with size depending on the standard deviation, are defined. The term *unsupervised* in this context indicates that the class-labels of the samples do not have any influence on the choice of the limits of the intervals. For each sample the value of the original feature is mapped on the intervals. The initial number of intervals has to be specified by the user. The number of intervals for each original feature can be reduced according to the following criteria: i) empty intervals are eliminated, ii) two adjacent intervals can be merged if both contain predominantly samples of the same class, iii) an interval can be joint with an adjacent interval, if very few samples of any class are mapped to it. The resulting intervals for each original feature can be represented by a set of binary features. The value of a binary feature is set to one if the expression level of a sample falls within the limits of the corresponding interval, or is set to zero otherwise. Features with the same binary map may be clustered.

Feature Selection

The feature selection (FS) step aims to identify a subset of features suitable to differentiate between the - in our case - two classes of samples. In the binary domain such a feature set can be found by solving a combinatorial problem termed as Set Covering Problem [39]. We consider a dataset of m samples of classes A and B and n features. The binary features can take on two possible values: $\{0, 1\}$. Denoting feature i of a sample h as f_{ih} then a feature f_i is able to discriminate (cover) a pair of samples k, h if $f_{ik} \neq f_{ih}$. In this case, feature i is added to the set of selected features. The problem of finding a feature set of minimal size where all pairs k, h with sample k belonging to class A and h belonging to class B are covered by at least one feature, can be mathematically formulated as:

$$\begin{aligned} &\min \sum_{i=1}^{n} x_i \\ &\sum_{i=1}^{n} a_{ij} x_i \geq 1 \\ &x_i \in \{0, 1\}, \quad i = 1 \ldots n, j = 1 \ldots M, \end{aligned} \tag{2.12}$$

with $x_i = 1$ if f_i is selected and 0 otherwise; M the number of sample pairs k, h; and a_{ij} is equal to one if feature i covers pair j. In order to improve the expected classification performance a certain amount of redundancy α can be introduced by selecting more than one feature to cover each pair of samples. MALA implements a modified version of the optimization problem in equation 2.12 where the number (β) of features to be selected is specified in advance and the redundancy α is maximized:

$$\begin{aligned}
&\sum_{i=1}^{n} x_i \leq \beta \\
&\max \alpha \\
&\sum_{i=1}^{n} a_{ij} x_i - \alpha \geq 0 \\
&x_i \in \{0,1\} \qquad i = 1 \ldots n, j = 1 \ldots M
\end{aligned} \tag{2.13}$$

MALA is also able to find an approximate solution of the set covering problem. The corresponding algorithm is based on the probability that a feature is present $\{1\}$ or absent $\{0\}$ in the samples of a class. If a feature is present in a sample and it is more likely that it is present in the samples of class A then the sample under consideration is classified as member of class A.

Due to the large number of features in the dataset, it is not possible to find an optimal solution for equation 2.13. In order to obtain an approximation of the optimal solution a heuristic approach, the efficient Greedy Randomized Adaptive Search Procedure (GRASP) [39, 43] is used instead. A GRASP iteration comprises two phases, the construction phase and a local search phase. In the first phase, a feasible solution is constructed adding one feature at a time. The features to be added are randomly picked from a restricted candidate list (RCL). The RCL is obtained by ordering all features according to a greedy function and considering a best ranked proportion of features. Depending on the portion of features in the restricted candidate list, the solution renders more greedy (shorter RCL) or more random (larger RCL). The greedy function takes into account the benefit of adding a feature to the solution, that is the number of sample pairs to be additionally covered by adding that feature. The RCL is updated each time a feature has been added. The maximum number of features to be selected is limited to the parameter β. In the second phase of the GRASP iteration, the local neighborhood structure is searched for a better solution compared to the one constructed in phase one and - if a better solution was found - is replaced by the best solution in the neighborhood. Finally, the best solution across all GRASP iterations is kept as the solution to the proposed FS problem. The pseudo-code of GRASP is depicted in listing 2.1.

Listing 2.1: Pseudo-code of GRASP heuristic from Bertolazzi *et al.* [39]

```
procedure GRASP(MaxIterations)
        for i = 1,...,MaxIterations do
                Build a greedy randomized solution x;
                x ← LocalSearch(x);
                if i = 1 then $x^*$ ← x;
                else if w(x) < w($x^*$) then $x^*$ ← x;
        end;
        return($x^*$);
end GRASP;
```

Formula Extraction

In this step, a number of classification rules is inferred from the list of candidate features resulting from the FS step. The features are assembled in a number of logic formulas in Disjunctive Normal Form (DNF) of type: *if feature x is in the value range* R_1 AND *feature y is in the value range* R_2 OR *feature z is in the value range* R_3 *then the sample under consideration is classified as member of class A*. To do so, MALA employs the learning system *Lsquare* described by Felici and Truemper [41]. The problem of finding the classification rules is formulated as a minimum cost satisfiability problem (MINSAT). The solution of the MINSAT problem is described in [44]. The features comprised by a conjunctive clause are of interest because they may account for the main molecular-biological differences between the classes.

Software

MALA is available as a software package written in ANSI C. A compiled command line version was used under Linux and Windows operating systems. The parameter settings of MALA can be changed by editing a text file (./MALA/parameters.dat).

Parameter Settings

The clustering of features is deactivated. As sampling type, random percentage split is selected. This means that a specified percentage of samples in the dataset is randomly selected and assigned to the trainings set. The percentage of samples to be selected for training is set to 80. The number of subsets (how many times should the dataset be split in training and test set) is set to the maximum of 100 for the biological dataset and to 10 for the synthetic dataset respectively. These values have been chosen to obtain a number of selected features comparable with the other methods. The set of selected features resulting from MALA is the accumulation of the features selected in each subset. The number of initial intervals for the feature discretization is left to the default of 7. In order to find not just an approximate solution of the feature selection problem, the type of the set covering problem is chosen to be *quadratic*. The maximum number of features to be chosen during the feature selection step β is set to the maximum of 50. The number of seconds and the number of GRASP iterations to be dedicated to the resolution of the feature selection problem are limited to 960 seconds (maximum value) and to 10 000/1 000 iterations for the biological/synthetic dataset respectively. The values have been set to the maximum for the biological datasets because it was expected that with a larger effort dedicated to the feature selection problem, the performance of the resulting classification model could be improved. The classification model for the synthetic datasets performed quite satisfyingly even at a lower effort than the maximum. The proportion of top ranked features to be included in the RCL is set to the default value of 60%. The costs of the inclusion of literals into the logic formulas are set to a minimum and the extent of the result in terms of numbers of literals and clauses comprised by the logic formulas is set to be maximized.

The performance of the classification model derived by MALA is characterized by the averaged number of correctly, wrong and not classified elements within training and test sets. For the synthetic datasets the average performance of the extracted logic formulas of 10 subsets are given in Table 2.5. The performance statistics of the extracted logic formulas of 100 subsets in the biological datasets are summarized in Table 2.6.

Table 2.5.: Average number of correctly, wrong, and not classified elements in a training and test sets of synthetic data

	Training			Test		
%	correct	wrong	not	correct	wrong	not
mean	99.17	0.21	0.62	59.58	17.08	23.33
sd	1.02	0.42	0.95	25.76	13.10	13.33

Table 2.6.: Average number of correctly, wrong, and not classified elements in training and test sets of biological data

	Training			Test		
%	correct	wrong	not	correct	wrong	not
mean	77.46	0.41	22.12	38.59	3.09	58.32
sd	29.22	3.18	29.02	16.56	4.93	17.70

2.4. Comparison of Methods

The software implementations of the three integrative analysis methods are made available in an environment implemented in R Project for Statistical Computing [45] language. The resulting sets of features and GO terms and their overlaps are visualized with the R package *VennDiagram* [46]. The flat lists of candidate genes resulting from the three integrative analysis methods are compared to the genes in the *Pathways in Cancer* pathway from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [47] PATHWAY Database accessible via the Bioconductor [48, 49] package *graphite* [50]. Additionally, an over-representation analysis of annotated GO terms in three different categories is conducted. The over-representation analysis is accomplished with the Bioconductor packages *GOstats* [51] and *org.Hs.eg.db* [52]. The lists of candidate gene symbols are mapped to Entrez Gene identifiers (Entrez

IDs). Gene symbols which can not be mapped to an Entrez ID are omitted and duplicated Entrez IDs are removed. As gene universe, the whole set of Entrez IDs from the genome wide annotation in the human organism is used. The significance of over-representation is assessed with a hypergeometric test employing the GO terms associated to the genes in the universe and the GO terms associated to the candidate genes. The p-value cut-off for significant over-representation is set to 0.01. For the analysis, GO terms with a minimum category size of 5 are considered.

2.5. Data Sets

2.5.1. Synthetic Data

The synthetic data generation was accomplished with the tool *SynTReN* [53] which was designed for the simulation of large gene expression datasets based on transcriptional regulatory networks. The topology of a network is characterized by its structure, that is the nodes included in the network and the connecting edges between the nodes representing the mode of interaction. *SynTReN* derives a model for a network topology based on a list of pair-wise interacting nodes and by quantitative modeling of interactions between the nodes. Based on the network model the data simulation with *SynTReN* results in synthetic microarray datasets.

The structure of the basis networks for the simulation is inferred from the biological datasets. The network extraction process is described in the following. One network is derived based on the gene expression in tumor samples, the gene expression in normal samples, the DNA-methylation in tumor samples, the DNA-methylation in normal samples and the protein expression in tumor samples respectively. There is no protein expression data available for normal tissue samples in the biological datasets. The five resulting networks serve as basis for the generation of the synthetic microarray datasets referred to as gene expression, DNA-methylation and protein expression datasets. Since the biological datasets comprise 52 samples the number of samples to produce in the data simulation process is limited to 60.

The number of features in each dataset corresponds to the number of nodes provided in the basis network.

Preliminaries for Data Simulation: Gene Regulatory Network Inference

The structure of the networks to serve as basis for the synthetic data generation with *SynTReN* is inferred from the biological datasets described in subsection 2.5.2. The five datasets obtained after preprocessing represent expression profiles of features. They are used to identify features with similar expression profiles under certain conditions. These co-expressed features are supposed to be connected or to be co-regulated in the underlying transcriptional regulatory network. To identify co-expressed genes in each of the five subsets, the pair-wise correlation between the expression profiles of features was calculated in terms of Spearman's correlation coefficient and the significance of the correlation was assessed by the application of a correlation test. A multiple testing correction of p-values was conducted according to the method by Benjamini & Hochberg [54] which is available as part of the build-in R-function *cor.test()*. From each of the five subsets a co-expression network was derived for different cut-off values of the Spearman's correlation coefficient in the range of 0.5 to 0.9 in steps of 0.1. An overview of the resulting network sizes is provided in Tables 2.7, 2.8 and 2.9. Results of the analysis of centrality measures of the inferred networks are given in appendix A. Due to limited computational memory resources and the large number of edges, the cut-off for the Spearman's correlation coefficient was set to 0.9 for all networks except the network derived from the protein expression dataset, where the cut-off was set to 0.5. Based on these networks, 5 datasets are simulated comprising 60 samples and 390/2 748 features in the datasets based on the co-expression networks derived from the gene expression datasets of tumor/normal tissue; 2 471/2 809 features in the datasets based on the co-expression network derived from the DNA-methylation datasets of tumor/normal tissue; and 68 features in the dataset based on the co-expression network derived from the protein expression dataset of tumor tissue.

Table 2.7.: Size of co-expression networks inferred from gene expression data based on Spearman's correlation coefficient.

	tumor		normal	
cut-off	vertices	edges	vertices	edges
0.5	14 352	1 259 030	17 019	13 609 703
0.6	10 395	320 218	16 166	9 082 193
0.7	5 749	81 168	13 762	2 867 084
0.8	2 042	16 611	9 469	455 458
0.9	390	1 171	2 748	15 894

Table 2.8.: Size of co-expression networks inferred from DNA-methylation data based on Spearman's correlation coefficient.

	tumor		normal	
cut-off	vertices	edges	vertices	edges
0.5	9 671	2 343 699	12 996	5 949 001
0.6	7 005	627 017	11 539	2 341 900
0.7	4 852	71 545	8 701	542 535
0.8	3 052	4 418	5 163	71 746
0.9	2 471	1 594	2 809	3 004

Table 2.9.: Size of co-expression network inferred from protein expression data based on Spearman's correlation coefficient.

	tumor	
cut-off	vertices	edges
0.5	68	166
0.6	45	52
0.7	23	17
0.8	13	8
0.9	9	6

SynTReN - Microarray Simulation Tool

The major steps in the simulation process are depicted in Figure 2.9.

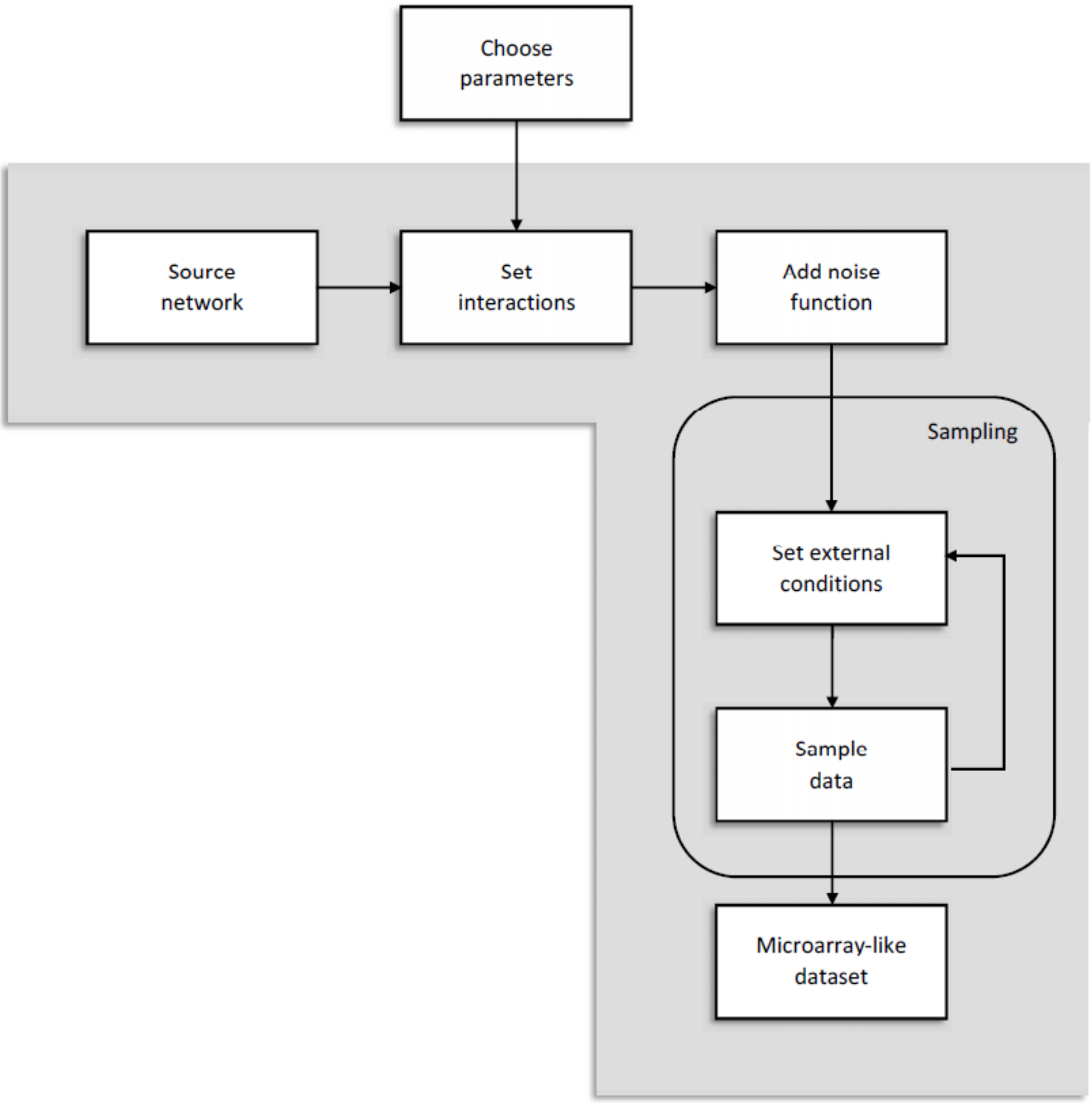

Figure 2.9.: *SynTReN* data generation process. Figure adapted from Van den Bulcke *et al.* [53].

The network structure must be provided as a *sif*-file [55] with the pairs of nodes interacting in the network contained in the lines. All nodes specified in the *sif*-file should be included in the simulated network (parameters createGeneNetwork and selectSubnetwork). The interaction on transcriptional regulatory level between the nodes in each pair is modeled by the assignment of a transition function to the corresponding edge. The transition function defines the dependency of the mRNA level of a gene from the mRNA level of its input nodes. It can optionally be superposed by biological noise (parameter bioNoise), that mimics the stochastic variations in gene expression. The regulatory interaction type between the nodes can directly be specified in the *sif*-file (parameter useEdgeTypesFromSIF) or can be chosen randomly as either activating or inhibiting with weighted probabilities (parameter percentActivators). A value of 50% activating interactions is chosen according to the findings for human gene networks in [56]. Two additional user-definable parameters (interactionCategory and higherOrderProbability) describe the complexity of interactions of nodes in the network. They define the steepness of interactions and the probability to chose a complex interaction to be assigned to an edge respectively. In order to generate different samples, an arbitrary expression level is assigned to nodes defined as externals (parameters fixedExternals and externalInputValues). If the parameter nrExternals is set to -1, all top nodes (nodes which lack of input nodes) are assumed to be externals. None of them will show correlated behavior (parameter nrCorrelatedExternals). Since they are randomly sampled from a uniform distribution, there is no noise added to the input signals (parameter inputNoise). Given a certain constellation of values for the externals, subsequently the mRNA expression level for each gene in the network is derived. Finally, the experimental noise in the microarray data is simulated by adding a user-defined amount of noise from a lognormal distribution. The parameters used for the synthetic data generation are listed in Table 2.10. As emphasized by the authors, *SynTReN* outperforms other network simulators regarding the similarity with real biological networks, which is measured in terms of statistical properties and the computational performance, which is a linear function of the number of nodes in the network. *SynTReN* was implemented in Java and it is available for download. It is embedded in the R environment for the method comparison with a wrapper implemented by de Matos Simoes [57].

Table 2.10.: User-definable parameters and their values used for the synthetic data generation.

parameter	value
createGeneNetwork	true
selectSubnetwork	false
fixedExternals	false
useEdgeTypesFromSIF	false
percentActivators	0.5
interactionCategory	SIGMOIDAL
higherOrderProbability	0.3
nrExternals	-1
nrCorrelatedExternals	0
externalInputValues	RANDOMIZED
bioNoise	0.05
inputNoise	0.00
expNoise	0.01
nrExperiments	60
nrSamplesPerExp	1

2.5.2. Biological Data

The biological datasets employed for the analysis were downloaded from The Cancer Genome Atlas (TCGA) using an open source software for retrieving and processing TCGA data [58] which makes use of the R package *httr* [59]. The data sets comprise gene expression data for 20 531 genes measured on 1 160 samples; DNA-methylation data from 29 988 loci measured in 1 204 samples and protein expression data from 152 proteins measured in 410 samples. The samples originate from patients suffering from breast invasive carcionoma and were obtained either from tissue of the primary tumor or from adjacent normal tissue. A negligible small number of samples originate from other tissue types such as metastatic tissue or from cell line control. These samples are not considered in the analysis. The features in the datasets are associated with a gene symbol by default, which serves as common reference for the features between the datasets.

The data provided at TCGA was obtained by analysis of tumor and normal tissue on three different biological levels. The gene expression dataset reflects the signal due to the mRNA level in the tissue under study. The DNA-methylation dataset represents the percentage of reads where the cytosine base at a position on the DNA is methylated. A methylated position on the DNA may influence the expression level of nearby genes [60] and can hence be flagged with a gene symbol. Since the addition of a methyl group is supposed to down-regulate the expression of the associated gene, the values in the dataset have been transformed in such a way that $x' = 1 - x$. With this transformation it is guaranteed that hight values in all datasets have a similar molecular-biological meaning. The protein expression dataset contains the normalized protein expression level of each gene per sample. The datasets are subjected to comprehensive preprocessing including the following steps: i) remove features not associated to a gene symbol; ii) split the datasets into tumor and normal samples; iii) remove incomplete cases, that is remove features containing NAs and those containing only zero-values; iv) replace sets of features associated with the same gene symbol within a dataset by one single *merged* feature representing the mean of all redundant features with non-zero variances. The preprocessing results in five subsets which are i) gene expression in tumor samples; ii) gene expression in normal samples; iii) DNA-methylation in tumor samples; iv) DNA-methylation in normal samples and v) protein expression in tumor samples. There is no protein expression data available for normal tissue samples. In the next step, the subsets were reduced to contain only samples (patients) where data is available in all five subsets. This results in five subsets comprising 52 samples and 19 769/19 716 features in the gene expression dataset of tumor/normal tissue; 13 627/14 300 features in the DNA-methylation dataset of tumor/normal tissue; and 118 features in the protein expression dataset of tumor tissue.

3. Results

The results of the application of three integrative analysis methods on synthetic and biological datasets are presented in this chapter. The results are divided in two categories: results on synthetic data and results on biological data. On the most specific level, the gene level, each method yields a list of candidate genes. For the biological data, the candidate genes selected by the methods are compared to a set of genes known to be involved in the *Pathways in cancer* pathway from the Kyoto Encyclopedia of Genes and Genomes (KEGG) PATHWAY Database [47]. For each category, over-represented GO terms in three GO categories: biological process (BP), molecular function (MF) and cellular component (CC) associated with the selected candidate genes are identified. The resulting sets of genes and GO terms are compared to each other and their overlaps are illustrated as Venn diagrams.

3.1. Synthetic Data

3.1.1. Sparse Canonical Correlation Analysis

The sCCA results in canonical weight-vectors with 20, 119 and 3 non-zero weights for the gene expression (GE), the DNA-methylation (MET) and the protein expression (PE) datasets respectively. The number of non-zero elements in the canonical weight-vectors corresponds to the number of selected features in the associated dataset. The total number of candidate features selected by sCCA is the unified sum (the *merged* set) of the selected features in each dataset and is 142. The number of over-represented GO terms in each category which are associated with the selected genes from each data set as well as with the merged set of selected genes are listed in Table 3.1.

Table 3.1.: Number of over-represented GO terms (p-value < 0.01) in each category derived from genes selected by sCCA.

GO category	merged	GE	MET	PE
BP	79	103	34	103
MF	23	4	23	16
CC	6	8	11	5

As an example, GO BP terms with category size between 5 and 100 are listed in Table 3.2, GO MF terms with category size between 5 and 1 000 are shown in Table 3.3 and GO CC terms with category size between 5 and 1 000 are presented in Table 3.4. These terms represent the most specific terms derived from the sCCA results.

Table 3.2.: GO BP (p-value < 0.01) associated with genes selected by sCCA of category size between 5 and 100

GO Slim Term	Size	GO Slim Term Description
GO:0002863	9	positive regulation of inflammatory response to antigenic stimulus
GO:0010388	9	cullin deneddylation
GO:0070208	11	protein heterotrimerization
GO:0000338	12	protein deneddylation
GO:0002922	13	positive regulation of humoral immune response
GO:0051383	14	kinetochore organization
GO:0033151	15	V(D)J recombination
GO:0031579	16	membrane raft organization
GO:0006907	17	pinocytosis
GO:1902187	17	negative regulation of viral release from host cell
GO:0032878	19	regulation of establishment or maintenance of cell polarity
GO:0002861	20	regulation of inflammatory response to antigenic stimulus
GO:0007289	20	spermatid nucleus differentiation
GO:0042119	22	neutrophil activation
GO:1901890	22	positive regulation of cell junction assembly
GO:0014009	23	glial cell proliferation
GO:0043267	23	negative regulation of potassium ion transport
GO:2000403	23	positive regulation of lymphocyte migration
GO:0007339	37	binding of sperm to zona pellucida
GO:0043551	37	regulation of phosphatidylinositol 3-kinase activity
GO:1903727	38	positive regulation of phospholipid metabolic process
GO:0043550	44	regulation of lipid kinase activity
GO:0030433	47	ER-associated ubiquitin-dependent protein catabolic process
GO:0042116	47	macrophage activation
GO:0045428	48	regulation of nitric oxide biosynthetic process
GO:0035036	49	sperm-egg recognition
GO:1903725	55	regulation of phospholipid metabolic process
GO:1903426	55	regulation of reactive oxygen species biosynthetic process
GO:0006809	59	nitric oxide biosynthetic process
GO:0009988	61	cell-cell recognition

Table 3.3.: GO MF (p-value < 0.01) associated with genes selected by sCCA of category size between 5 and 1 000

GO Slim Term	Size	GO Slim Term Description
GO:0031683	21	G-protein beta/gamma-subunit complex binding
GO:0004386	145	helicase activity
GO:0003924	246	GTPase activity
GO:0017111	753	nucleoside-triphosphatase activity
GO:0016462	792	pyrophosphatase activity
GO:0016818	794	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
GO:0016817	795	hydrolase activity, acting on acid anhydrides

Table 3.4.: GO CC (p-value < 0.01) associated with genes selected by sCCA of category size between 5 and 1 000

GO Slim Term	Size	GO Slim Term Description
GO:0042599	12	lamellar body
GO:0002080	21	acrosomal membrane

3.1.2. Non-Negative Matrix Factorization

The number of selected genes is 17 in the gene expression (GE) and 120 in the DNA-methylation (MET) dataset. The selected md-module comprises no features in the protein expression dataset. In total the NMF results in a merged set of 137 candidate genes. The number of over-represented GO terms in each category associated with the genes selected by NMF are listed in Table 3.5.

Table 3.5.: Number of over-represented GO terms (p-value < 0.01) in each category derived from genes selected by NMF.

GO category	merged	GE	MET
BP	141	312	37
MF	15	20	14
CC	9	32	5

As an example, GO BP terms with category size between 5 and 100 are listed in Table 3.6, GO MF terms with category size between 5 and 1 000 are shown in Table 3.7 and GO CC terms with category size between 5 and 1 000 are presented in Table 3.8. These terms represent the most specific terms derived from the NMF results.

Table 3.6.: GO BP (p-value < 0.01) associated with genes selected by NMF of category size between 5 and 100

GO Slim Term	Size	GO Slim Term Description
GO:0002291	5	T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell
GO:0002767	5	immune response-inhibiting cell surface receptor signaling pathway
GO:0002309	6	T cell proliferation involved in immune response
GO:0002765	6	immune response-inhibiting signal transduction
GO:0030300	7	regulation of intestinal cholesterol absorption
GO:2001198	7	regulation of dendritic cell differentiation
GO:0060457	8	negative regulation of digestive system process
GO:2001030	8	negative regulation of cellular glucuronidation
GO:0052697	9	xenobiotic glucuronidation
GO:2001029	9	regulation of cellular glucuronidation
GO:0070493	10	thrombin receptor signaling pathway
GO:0001820	11	serotonin secretion
GO:0002664	11	regulation of T cell tolerance induction
GO:1903010	11	regulation of bone development
GO:0002517	12	T cell tolerance induction
GO:0032372	12	negative regulation of sterol transport
GO:0032375	12	negative regulation of cholesterol transport
GO:0002643	13	regulation of tolerance induction
GO:0030299	13	intestinal cholesterol absorption
GO:0045086	13	positive regulation of interleukin-2 biosynthetic process
GO:0030852	15	regulation of granulocyte differentiation
GO:0043931	15	ossification involved in bone maturation
GO:0001711	16	endodermal cell fate commitment
GO:0006837	16	serotonin transport
GO:0060236	17	regulation of mitotic spindle organization
GO:0070977	17	bone maturation
GO:0006882	18	cellular zinc ion homeostasis
GO:0044241	18	lipid digestion
GO:0002507	19	tolerance induction
GO:0045076	19	regulation of interleukin-2 biosynthetic process
GO:0055069	20	zinc ion homeostasis
GO:0009813	21	flavonoid biosynthetic process
GO:0052696	21	flavonoid glucuronidation
GO:0045671	21	negative regulation of osteoclast differentiation
GO:0048799	21	organ maturation
GO:0090224	21	regulation of spindle organization
GO:0042094	22	interleukin-2 biosynthetic process
GO:0043586	22	tongue development
GO:0045922	24	negative regulation of fatty acid metabolic process
GO:0050892	24	intestinal absorption
GO:0052695	25	cellular glucuronidation
GO:0006063	26	uronic acid metabolic process
GO:0019585	26	glucuronate metabolic process
GO:0009812	27	flavonoid metabolic process
GO:0097028	39	dendritic cell differentiation
GO:0002762	40	negative regulation of myeloid leukocyte differentiation
GO:0010677	42	negative regulation of cellular carbohydrate metabolic process
GO:2000107	43	negative regulation of leukocyte apoptotic process
GO:0045912	49	negative regulation of carbohydrate metabolic process
GO:0045670	57	regulation of osteoclast differentiation

GO Slim Term	Size	GO Slim Term Description
GO:0007052	61	mitotic spindle organization
GO:0046634	62	regulation of alpha-beta T cell activation
GO:0045582	64	positive regulation of T cell differentiation
GO:0042440	66	pigment metabolic process
GO:0007588	70	excretion
GO:0031295	75	T cell costimulation
GO:0031294	76	lymphocyte costimulation
GO:0045638	83	negative regulation of myeloid cell differentiation
GO:0042102	85	positive regulation of T cell proliferation
GO:0019886	93	antigen processing and presentation of exogenous peptide antigen via MHC class II
GO:0002495	97	antigen processing and presentation of peptide antigen via MHC class II
GO:0002504	98	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II

Table 3.7.: GO MF (p-value < 0.01) associated with genes selected by NMF of category size between 5 and 1 000

GO Slim Term	Size	GO Slim Term Description
GO:0032393	9	MHC class I receptor activity
GO:0008157	12	protein phosphatase 1 binding
GO:0042288	13	MHC class I protein binding
GO:0017127	18	cholesterol transporter activity
GO:0000993	19	RNA polymerase II core binding
GO:0015248	19	sterol transporter activity
GO:0001098	22	basal transcription machinery binding
GO:0001099	22	basal RNA polymerase II transcription machinery binding
GO:0043175	22	RNA polymerase core enzyme binding
GO:0042287	24	MHC protein binding
GO:0015020	34	glucuronosyltransferase activity
GO:0003823	107	antigen binding
GO:0016758	196	transferase activity, transferring hexosyl groups
GO:0003924	246	GTPase activity

Table 3.8.: GO CC (p-value < 0.01) associated with genes selected by NMF of category size between 5 and 1 000

GO Slim Term	Size	GO Slim Term Description
GO:0043190	8	ATP-binding cassette (ABC) transporter complex
GO:0042613	16	MHC class II protein complex
GO:0042611	27	MHC protein complex
GO:0008023	42	transcription elongation factor complex
GO:0012507	42	ER to Golgi transport vesicle membrane
GO:0030134	52	ER to Golgi transport vesicle
GO:0030133	171	transport vesicle
GO:0009986	688	cell surface

3.1.3. Microarray Logic Analyzer

The logic formulas resulting from MALA comprise 13 genes from gene expression (GE) and 189 genes from the DNA-methylation (MET) dataset respectively. Due to the lack of protein expression data from normal tissue, the protein expression dataset was not analyzed with MALA. In total the merged selected feature set comprises 201 gene symbols. The numbers of over-represented GO terms associated with genes in each dataset selected by MALA are summarized in Table 3.9.

Table 3.9.: Number of over-represented GO terms (p-value < 0.01) in each category derived from genes selected by MALA.

GO category	merged	GE	MET
BP	24	48	19
MF	4	18	4
CC	21	12	16

As an example, GO BP terms with category size between 5 and 100 are listed in Table 3.10, GO MF terms with category size between 5 and 1 000 are shown in Table 3.11 and GO CC terms with category size between 5 and 1 000 are presented in Table 3.12. These terms represent the most specific terms derived from the MALA results.

Table 3.10.: GO BP (p-value < 0.01) associated with genes selected by MALA of category size between 5 and 100

GO Slim Term	Size	GO Slim Term Description
GO:0031573	11	intra-S DNA damage checkpoint
GO:0009219	15	pyrimidine deoxyribonucleotide metabolic process
GO:1902230	27	negative regulation of intrinsic apoptotic signaling pathway in response to DNA damage
GO:0006298	30	mismatch repair
GO:1902229	34	regulation of intrinsic apoptotic signaling pathway in response to DNA damage
GO:0030490	36	maturation of SSU-rRNA
GO:0031572	37	G2 DNA damage checkpoint
GO:0034080	42	CENP-A containing nucleosome assembly
GO:0061641	42	CENP-A containing chromatin organization
GO:0031055	44	chromatin remodeling at centromere
GO:2001021	44	negative regulation of response to DNA damage stimulus
GO:0042274	47	ribosomal small subunit biogenesis

Table 3.11.: GO MF (p-value < 0.01) associated with genes selected by MALA of category size between 5 and 1 000

GO Slim Term	Size	GO Slim Term Description
GO:0016627	55	oxidoreductase activity, acting on the CH-CH group of donors
GO:0003697	78	single-stranded DNA binding

Table 3.12.: GO CC (p-value < 0.01) associated with genes selected by MALA of category size between 5 and 1 000

GO Slim Term	Size	GO Slim Term Description
GO:0032300	11	mismatch repair complex
GO:0030686	24	90S preribosome
GO:0005771	31	multivesicular body
GO:0032040	31	small-subunit processome
GO:0030684	44	preribosome
GO:0044452	55	nucleolar part
GO:0000776	116	kinetochore
GO:0000775	167	chromosome, centromeric region
GO:0000793	186	condensed chromosome
GO:0098687	226	chromosomal region
GO:0005774	290	vacuolar membrane
GO:0005694	785	chromosome

3.1.4. Comparison of Methods

The sets of genes selected by the three integrative analysis methods and the sets of over-represented GO terms are compared for all merged data types in Figure 3.1. The genes selected by two of three methods are listed in Table 3.13. There are no genes in the synthetic datasets which were selected by all methods. The overlap of GO terms associated with the genes selected by each method are presented in Tables 3.14, 3.15 and 3.16.

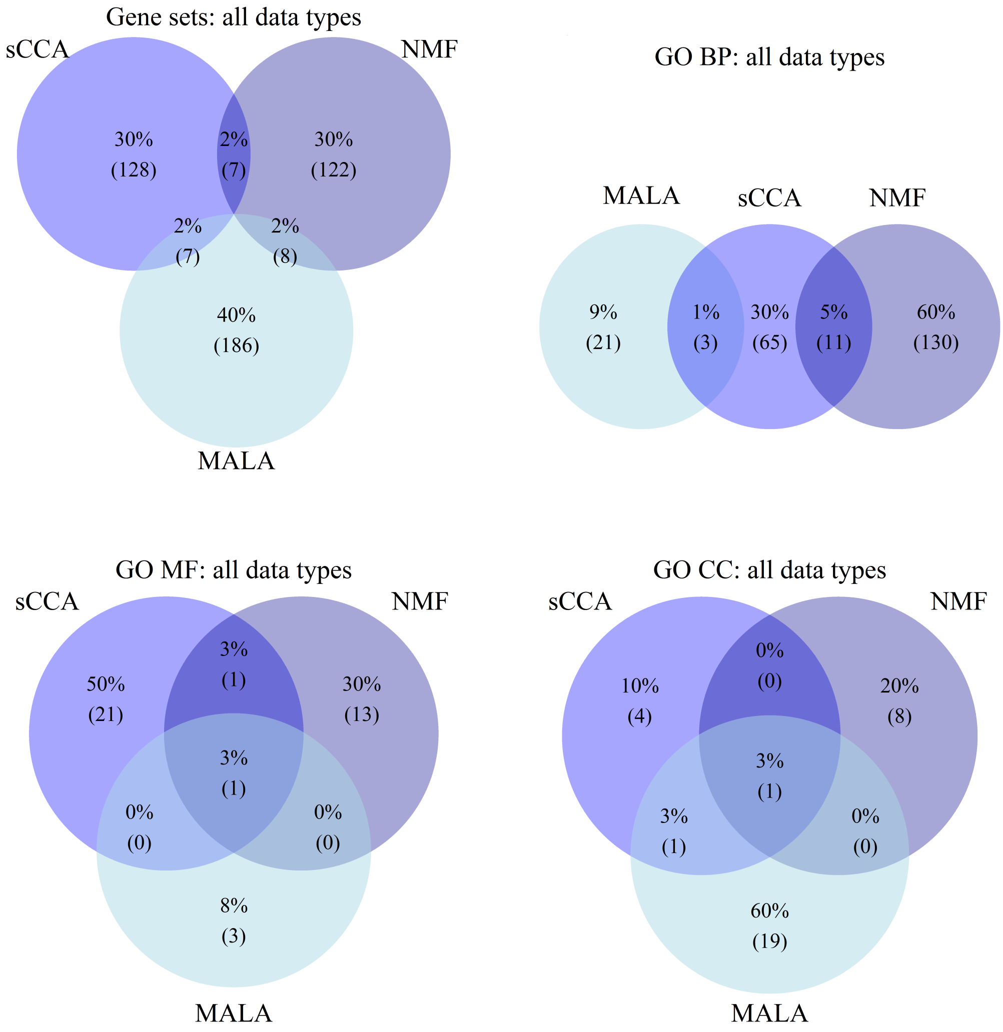

Figure 3.1.: Venn diagrams of gene sets merged from all data types and over-represented GO terms associated with gene sets.

The sets of selected genes from each data type and sets of over-represented GO terms of

Table 3.13.: Genes selected by two of three methods

Symbol	Gene name	sCCA	NMF	MALA
DERL2	derlin 2	✓	✓	
MIS12	MIS12 kinetochore complex component	✓	✓	
MSTO1	misato 1, mitochondrial distribution and morphology regulator	✓	✓	
SBDS	Shwachman-Bodian-Diamond syndrome	✓	✓	
TYW1	tRNA-yW synthesizing protein 1 homolog (S. cerevisiae)	✓	✓	
UGT1A6	UDP glucuronosyltransferase 1 family, polypeptide A6	✓	✓	
ZSCAN29	zinc finger and SCAN domain containing 29	✓	✓	
GALK2	galactokinase 2	✓		✓
LXN	latexin	✓		✓
LYPD4	LY6/PLAUR domain containing 4	✓		✓
MBD4	methyl-CpG binding domain protein 4	✓		✓
MRPS18C	mitochondrial ribosomal protein S18C	✓		✓
RBMXL3	RNA binding motif protein, X-linked-like 3	✓		✓
TPT1	tumor protein, translationally-controlled 1	✓		✓
TXNDC9	thioredoxin domain containing 9		✓	✓
ABCG8	ATP-binding cassette, sub-family G (WHITE), member 8		✓	✓
CNTD1	cyclin N-terminal domain containing 1		✓	✓
DNAJC25-GNG10	DNAJC25-GNG10 readthrough		✓	✓
LRRC57	leucine rich repeat containing 57		✓	✓
TRIM23	tripartite motif containing 23		✓	✓
UBFD1	ubiquitin family domain containing 1		✓	✓
USMG5	up-regulated during skeletal muscle growth 5 homolog (mouse)		✓	✓

categories BP, MF and CC associated with the selected genes are displayed in Figures 3.2, 3.3 and 3.4.

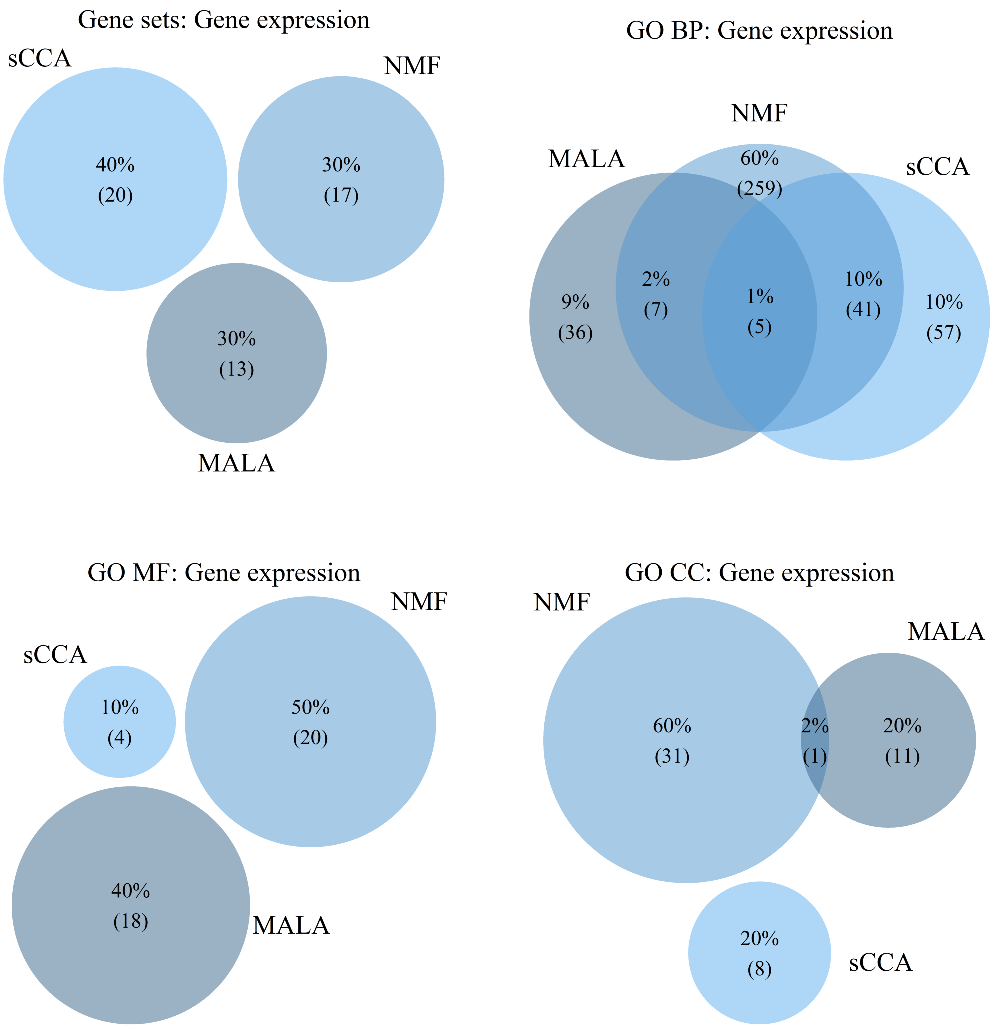

Figure 3.2.: Venn diagrams of gene sets and over-represented GO terms extracted from the gene expression dataset.

Table 3.14.: Overlap of GO BP

GO Slim Term ID	GO Slim Term Description	sCCA	NMF	MALA
GO:0001775	cell activation	✓	✓	
GO:0002376	immune system process	✓	✓	
GO:0002682	regulation of immune system process	✓	✓	
GO:0002684	positive regulation of immune system process	✓	✓	
GO:0007159	leukocyte cell-cell adhesion	✓	✓	
GO:0009605	response to external stimulus	✓	✓	
GO:0042110	T cell activation	✓	✓	
GO:0048534	hematopoietic or lymphoid organ development	✓	✓	
GO:0070486	leukocyte aggregation	✓	✓	
GO:0070489	T cell aggregation	✓	✓	
GO:0071593	lymphocyte aggregation	✓	✓	
GO:0000278	mitotic cell cycle	✓		✓
GO:0006974	cellular response to DNA damage stimulus	✓		✓
GO:0007049	cell cycle	✓		✓

Table 3.15.: Overlap of GO MF

GO Slim Term ID	GO Slim Term Description	sCCA	NMF	MALA
GO:0003723	RNA binding	✓	✓	✓
GO:0003924	GTPase activity	✓	✓	

Table 3.16.: Overlap of GO CC

GO Slim Term ID	GO Slim Term Description	sCCA	NMF	MALA
GO:0044422	organelle part	✓	✓	✓
GO:0044446	intracellular organelle part	✓		✓

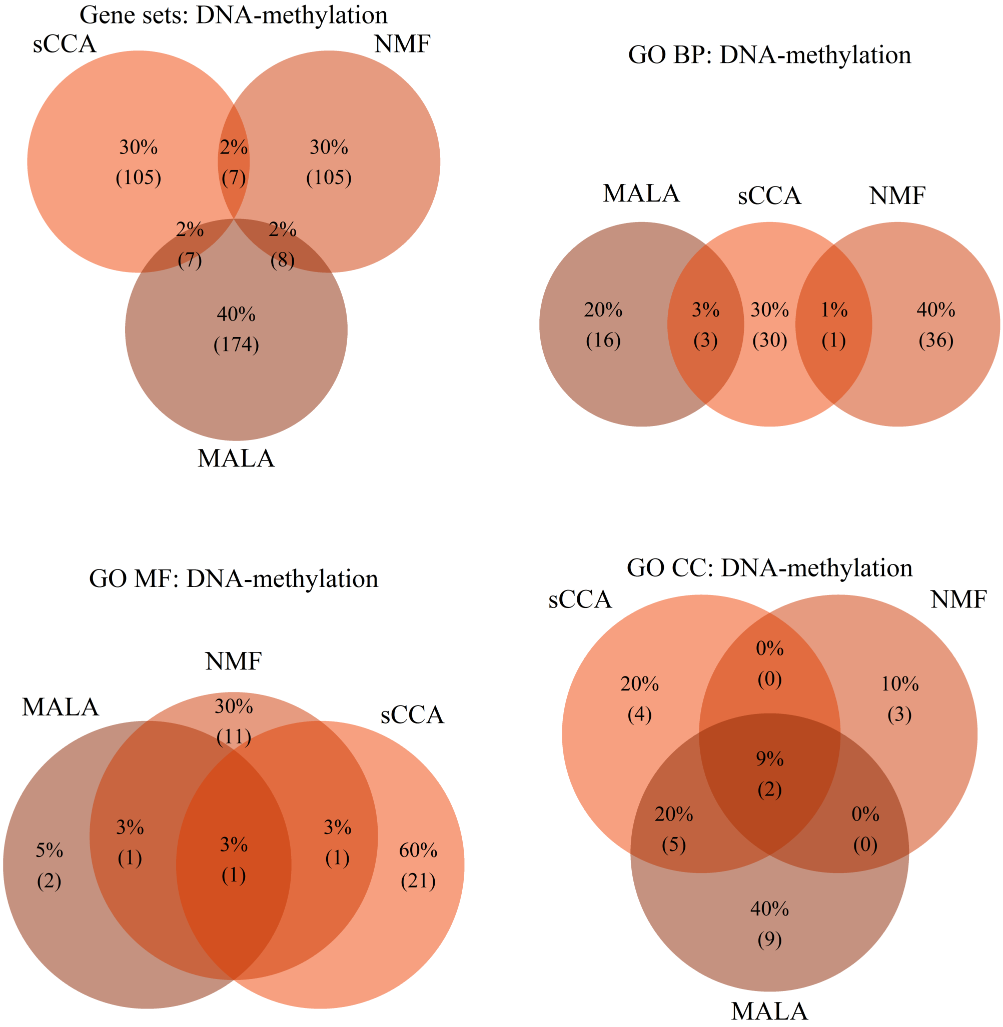

Figure 3.3.: Venn diagrams of gene sets and over-represented GO terms extracted from the DNA-methylation dataset.

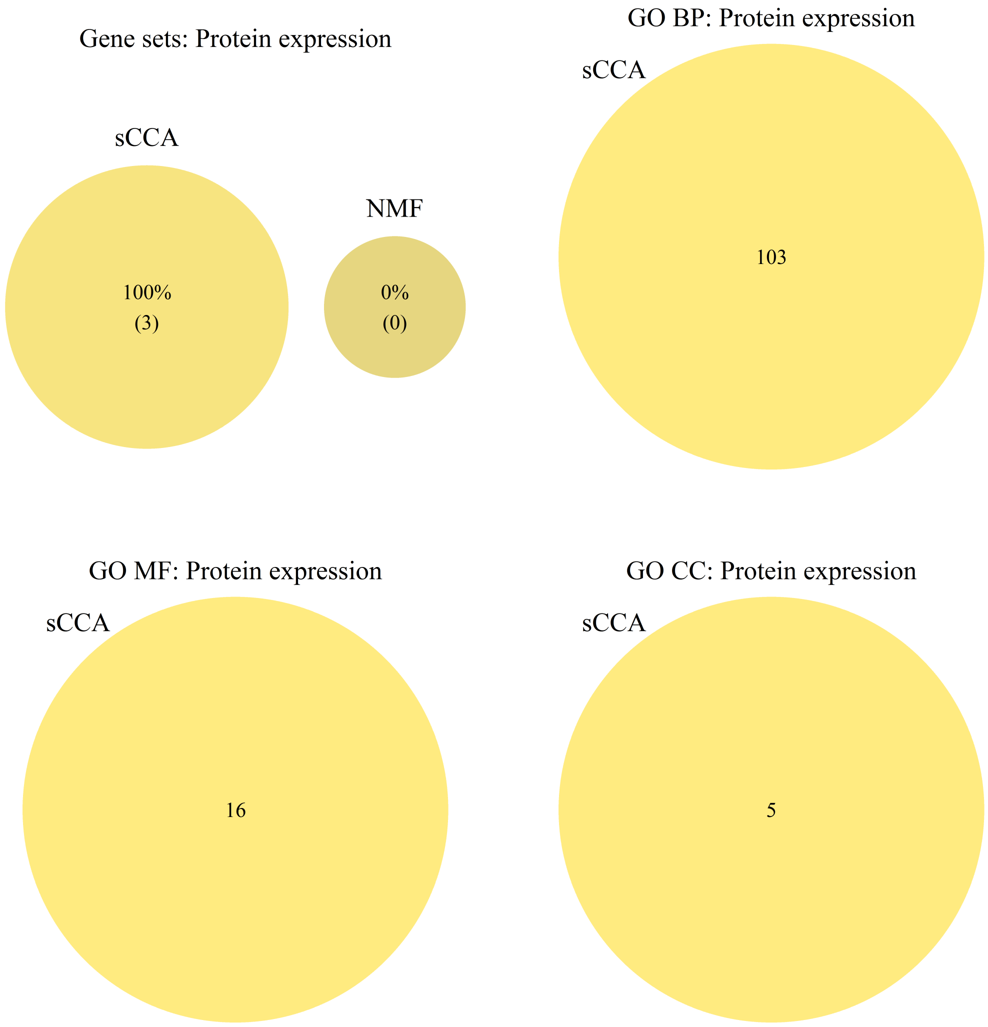

Figure 3.4.: Venn diagrams of gene sets and over-represented GO terms extracted from the protein expression dataset.

3.2. Biological Data

3.2.1. Sparse Canonical Correlation Analysis

The number of selected features (non-zero elements in the canonical weight-vectors) are 1 014 for the gene expression (GE), 690 for the DNA-methylation (MET) and 7 for the protein expression (PE) dataset respectively.

Table 3.17.: Genes selected by sCCA known to be involved in cancer pathways.

Symbol	Gene name
BCR	breakpoint cluster region
CDKN2A	cyclin-dependent kinase inhibitor 2A
DAPK2	death-associated protein kinase 2
E2F3	E2F transcription factor 3
FZD7	frizzled class receptor 7
LAMA1	laminin, alpha 1
LAMB1	laminin, beta 1
PLCG1	phospholipase C, gamma 1
PTCH1	patched 1
RALGDS	ral guanine nucleotide dissociation stimulator
SKP2	S-phase kinase-associated protein 2, E3 ubiquitin protein ligase
SMO	smoothened, frizzled class receptor
SOS2	son of sevenless homolog 2 (Drosophila)
TCF7L1	transcription factor 7-like 1 (T-cell specific, HMG-box)
TCF7L2	transcription factor 7-like 2 (T-cell specific, HMG-box)
WNT8B	wingless-type MMTV integration site family, member 8B
EPAS1	endothelial PAS domain protein 1
RHOA	ras homolog family member A
CSF1R	colony stimulating factor 1 receptor
CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
FZD1	frizzled class receptor 1
IKBKB	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta
MYC	v-myc avian myelocytomatosis viral oncogene homolog
CCDC6	coiled-coil domain containing 6
FGFR2	fibroblast growth factor receptor 2
FOXO1	forkhead box O1
MAX	MYC associated factor X
AKT1	v-akt murine thymoma viral oncogene homolog 1
DVL2	dishevelled segment polarity protein 2
MAP2K2	mitogen-activated protein kinase kinase 2
PIK3R2	phosphoinositide-3-kinase, regulatory subunit 2 (beta)
PDGFB	platelet-derived growth factor beta polypeptide
AR	androgen receptor
BRAF	B-Raf proto-oncogene, serine/threonine kinase
CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa
CCND1	cyclin D1
CCNE1	cyclin E1

The merged and unified set of selected features comprises 1 663 genes. The intersection of the merged gene set with known cancer genes from the KEGG database results in 37 genes. They are listed in Table 3.17. The number of over-represented GO terms in each category associated with the selected genes in each dataset and to the merged set of all selected genes are listed in Table 3.18.

Table 3.18.: Number of over-represented GO terms (p-value < 0.01) in each category derived from genes selected by sCCA.

GO category	merged	GE	MET	PE
BP	205	121	242	459
MF	38	35	28	40
CC	32	5	46	23

As an example, GO BP terms with category size between 5 and 10 are listed in Table 3.19, GO MF terms with category size between 5 and 100 are shown in Table 3.20 and GO CC terms with category size between 5 and 100 are presented in Table 3.21. These terms represent the most specific terms derived from the sCCA results.

Table 3.19.: GO BP (p-value < 0.01) associated with genes selected by sCCA of category size between 5 and 10

GO Slim Term	Size	GO Slim Term Description
GO:0006824	5	cobalt ion transport
GO:0010757	5	negative regulation of plasminogen activation
GO:0060916	5	mesenchymal cell proliferation involved in lung development
GO:0071281	6	cellular response to iron ion
GO:0003149	6	membranous septum morphogenesis
GO:0009744	6	response to sucrose
GO:0034285	6	response to disaccharide
GO:0070141	6	response to UV-A
GO:0097210	6	response to gonadotropin-releasing hormone
GO:0097211	6	cellular response to gonadotropin-releasing hormone
GO:0036297	8	interstrand cross-link repair
GO:0048318	8	axial mesoderm development
GO:0002676	9	regulation of chronic inflammatory response
GO:0031507	9	heterochromatin assembly
GO:0010755	10	regulation of plasminogen activation
GO:0032988	10	ribonucleoprotein complex disassembly
GO:0051918	10	negative regulation of fibrinolysis

Table 3.20.: GO MF (p-value < 0.01) associated with genes selected by sCCA of category size between 5 and 100

GO Slim Term	Size	GO Slim Term Description
GO:0004630	6	phospholipase D activity
GO:0005247	10	voltage-gated chloride channel activity
GO:0005542	10	folic acid binding
GO:0008409	11	5′-3′ exonuclease activity
GO:0015168	11	glycerol transmembrane transporter activity
GO:0015250	11	water channel activity
GO:0015254	11	glycerol channel activity
GO:0008199	12	ferric iron binding
GO:0016538	20	cyclin-dependent protein serine/threonine kinase regulator activity
GO:0030507	25	spectrin binding
GO:0004004	37	ATP-dependent RNA helicase activity
GO:0008186	38	RNA-dependent ATPase activity
GO:0005089	77	Rho guanyl-nucleotide exchange factor activity

Table 3.21.: GO CC (p-value < 0.01) associated with genes selected by sCCA of category size between 5 and 100

GO Slim Term	Size	GO Slim Term Description
GO:0042382	6	paraspeckles
GO:0005861	8	troponin complex
GO:0070688	8	MLL5-L complex
GO:0019908	9	nuclear cyclin-dependent protein kinase holoenzyme complex
GO:0032300	11	mismatch repair complex
GO:0097381	15	photoreceptor disc membrane
GO:0000307	20	cyclin-dependent protein kinase holoenzyme complex
GO:0005865	22	striated muscle thin filament
GO:0036379	25	myofilament
GO:0000791	29	euchromatin
GO:0009925	29	basal plasma membrane
GO:0045178	42	basal part of cell
GO:0001750	59	photoreceptor outer segment
GO:0000792	73	heterochromatin
GO:1902911	86	protein kinase complex

3.2.2. Non-Negative Matrix Factorization

The selected feature sets in the gene expression (GE), the DNA-methylation (MET) and the protein expression (PE) dataset comprise 664, 478 and 5 genes respectively. The total number of genes selected by NMF is the unified sum of the three sets (merged) and is 1 127. The 25 genes selected by NMF which are involved in pathways in cancer are displayed in Table 3.22.

Table 3.22.: Genes selected by NMF which are known to be involved in cancer pathways.

Symbol	Gene name
CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)
FGF19	fibroblast growth factor 19
FIGF	c-fos induced growth factor (vascular endothelial growth factor D)
HHIP	hedgehog interacting protein
IGF1	insulin-like growth factor 1 (somatomedin C)
ITGA2B	integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)
MAP2K2	mitogen-activated protein kinase kinase 2
PPARG	peroxisome proliferator-activated receptor gamma
RELA	v-rel avian reticuloendotheliosis viral oncogene homolog A
RUNX1T1	runt-related transcription factor 1; translocated to, 1 (cyclin D-related)
RXRB	retinoid X receptor, beta
CYCS	cytochrome c, somatic
WNT5B	wingless-type MMTV integration site family, member 5B
BCL2	B-cell CLL/lymphoma 2
RXRG	retinoid X receptor, gamma
FGF2	fibroblast growth factor 2 (basic)
PDGFA	platelet-derived growth factor alpha polypeptide
RAC1	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)
SHH	sonic hedgehog
PTCH1	patched 1
RET	ret proto-oncogene
FADD	Fas (TNFRSF6)-associated via death domain
WNT11	wingless-type MMTV integration site family, member 11
TCEB2	transcription elongation factor B (SIII), polypeptide 2 (18kDa, elongin B)
CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa

The number of over-represented GO terms in each category associated with the selected genes in each dataset and to the merged set of selected genes are listed in Table 3.23.

Table 3.23.: Number of over-represented GO terms (p-value < 0.01) in each category derived from genes selected by NMF.

GO category	merged	GE	MET	PE
BP	723	416	300	261
MF	56	45	27	22
CC	31	22	46	26

As an example, GO BP terms with category size between 5 and 10 are listed in Table 3.24, GO MF terms with category size between 5 and 100 are shown in Table 3.25 and GO CC terms with category size between 5 and 100 are presented in Table 3.26. These terms represent the most specific terms derived from the NMF results.

Table 3.24.: GO BP (p-value < 0.01) associated with genes selected by NMF of category size between 5 and 10

GO Slim Term	Size	GO Slim Term Description
GO:0072300	5	positive regulation of metanephric glomerulus development
GO:0010193	5	response to ozone
GO:0060331	5	negative regulation of response to interferon-gamma
GO:0060336	5	negative regulation of interferon-gamma-mediated signaling pathway
GO:0061526	5	acetylcholine secretion
GO:1903431	5	positive regulation of cell maturation
GO:0051552	6	flavone metabolic process
GO:0072298	6	regulation of metanephric glomerulus development
GO:0010887	6	negative regulation of cholesterol storage
GO:0015870	6	acetylcholine transport
GO:0060024	6	rhythmic synaptic transmission
GO:0060509	6	Type I pneumocyte differentiation
GO:0071763	6	nuclear membrane organization
GO:1902285	6	semaphorin-plexin signaling pathway involved in neuron projection guidance
GO:0014041	7	regulation of neuron maturation
GO:0008300	7	isoprenoid catabolic process
GO:0036295	7	cellular response to increased oxygen levels
GO:0045084	7	positive regulation of interleukin-12 biosynthetic process
GO:0071455	7	cellular response to hyperoxia
GO:1901374	7	acetate ester transport
GO:2001030	8	negative regulation of cellular glucuronidation
GO:0030638	8	polyketide metabolic process
GO:0044597	8	daunorubicin metabolic process
GO:0044598	8	doxorubicin metabolic process
GO:0048548	8	regulation of pinocytosis
GO:0090193	8	positive regulation of glomerulus development
GO:0035630	8	bone mineralization involved in bone maturation
GO:0060426	8	lung vasculature development
GO:0090037	8	positive regulation of protein kinase C signaling
GO:2000316	8	regulation of T-helper 17 type immune response
GO:0052697	9	xenobiotic glucuronidation
GO:2001029	9	regulation of cellular glucuronidation
GO:0021612	9	facial nerve structural organization
GO:0030647	9	aminoglycoside antibiotic metabolic process
GO:0021561	10	facial nerve development
GO:0021610	10	facial nerve morphogenesis
GO:0090520	10	sphingolipid mediated signaling pathway

Table 3.25.: GO MF (p-value < 0.01) associated with genes selected by NMF of category size between 5 and 100

GO Slim Term	Size	GO Slim Term Description
GO:0004024	5	alcohol dehydrogenase activity, zinc-dependent
GO:0004957	5	prostaglandin E receptor activity
GO:0005381	5	iron ion transmembrane transporter activity
GO:0008131	6	primary amine oxidase activity
GO:0005021	7	vascular endothelial growth factor-activated receptor activity
GO:0001758	7	retinal dehydrogenase activity
GO:0004032	7	alditol:NADP+ 1-oxidoreductase activity
GO:0004022	8	alcohol dehydrogenase (NAD) activity
GO:0005113	8	patched binding
GO:0004955	9	prostaglandin receptor activity
GO:0004954	10	prostanoid receptor activity
GO:0005451	10	monovalent cation:proton antiporter activity
GO:0008106	12	alcohol dehydrogenase (NADP+) activity
GO:0004953	14	icosanoid receptor activity
GO:0004806	15	triglyceride lipase activity
GO:0004033	20	aldo-keto reductase (NADP) activity
GO:0005504	24	fatty acid binding
GO:0005501	31	retinoid binding
GO:0019840	33	isoprenoid binding
GO:0015020	34	glucuronosyltransferase activity
GO:0017046	35	peptide hormone binding
GO:0016709	38	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NAD(P)H as one donor, and incorporation of one atom of oxygen
GO:0004879	49	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity
GO:0098531	49	direct ligand regulated sequence-specific DNA binding transcription factor activity
GO:0033293	51	monocarboxylic acid binding
GO:0016655	53	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor
GO:0003707	55	steroid hormone receptor activity
GO:0004714	64	transmembrane receptor protein tyrosine kinase activity
GO:0042562	65	hormone binding
GO:0005254	71	chloride channel activity
GO:0019199	80	transmembrane receptor protein kinase activity
GO:0005496	85	steroid binding
GO:0004497	94	monooxygenase activity
GO:0016651	95	oxidoreductase activity, acting on NAD(P)H

Table 3.26.: GO CC (p-value < 0.01) associated with genes selected by NMF of category size between 5 and 100

GO Slim Term	Size	GO Slim Term Description
GO:0097208	7	alveolar lamellar body
GO:0005771	31	multivesicular body
GO:0022627	40	cytosolic small ribosomal subunit
GO:0005891	41	voltage-gated calcium channel complex
GO:0022625	52	cytosolic large ribosomal subunit
GO:0005811	54	lipid particle
GO:0034704	63	calcium channel complex
GO:0015935	64	small ribosomal subunit
GO:0030667	74	secretory granule membrane
GO:0022626	100	cytosolic ribosome

3.2.3. Microarray Logic Analyzer

The classification model consisting of logic formulas derived by MALA comprises 329 features from the gene expression (GE) dataset and 950 features from the DNA-methylation (MET) dataset. Due to the lack of protein expression data from normal tissue, the protein expression dataset was not analyzed with MALA. In total MALA results in a merged set of 1 266 gene symbols selected from the datasets. The 30 genes selected by MALA which are part of the *Pathways in cancer* pathway from the KEGG database are given in Table 3.27.

Table 3.27.: Genes selected by MALA known to be part of pathways in cancer.

Symbol	Gene name
EGFR	epidermal growth factor receptor
FGF2	fibroblast growth factor 2 (basic)
FGF10	fibroblast growth factor 10
WNT11	wingless-type MMTV integration site family, member 11
PTGS2	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)
SMAD3	SMAD family member 3
WNT7B	wingless-type MMTV integration site family, member 7B
FZD1	frizzled class receptor 1
TRAF2	TNF receptor-associated factor 2
ZBTB16	zinc finger and BTB domain containing 16
EGLN3	egl-9 family hypoxia-inducible factor 3
LEF1	lymphoid enhancer-binding factor 1
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog
RHOA	ras homolog family member A
TRAF5	TNF receptor-associated factor 5
GLI1	GLI family zinc finger 1
CBL	Cbl proto-oncogene, E3 ubiquitin protein ligase
TRAF3	TNF receptor-associated factor 3
CCDC6	coiled-coil domain containing 6
RXRG	retinoid X receptor, gamma
PDGFA	platelet-derived growth factor alpha polypeptide
FOXO1	forkhead box O1
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
MAX	MYC associated factor X
BAX	BCL2-associated X protein
FOS	FBJ murine osteosarcoma viral oncogene homolog
PTCH1	patched 1
IGF1R	insulin-like growth factor 1 receptor
RALA	v-ral simian leukemia viral oncogene homolog A (ras related)
HIF1A	hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)

The number of over-represented GO terms in each category associated with the selected genes in each dataset as well as with the merged set of selected genes are listed in Table 3.28.

Table 3.28.: Number of over-represented GO terms (p-value < 0.01) in each category derived from genes selected by MALA.

GO category	merged	GE	MET
BP	223	151	260
MF	42	53	35
CC	79	35	97

As an example, GO BP terms with category size between 5 and 10 are listed in Table 3.29, GO MF terms with category size between 5 and 100 are shown in Table 3.30 and GO CC terms with category size between 5 and 100 are presented in Table 3.31. These terms represent the most specific terms derived from the MALA results.

Table 3.29.: GO BP (p-value < 0.01) associated with genes selected by MALA of category size between 5 and 10

GO Slim Term	Size	GO Slim Term Description
GO:0008627	5	intrinsic apoptotic signaling pathway in response to osmotic stress
GO:0019896	5	axon transport of mitochondrion
GO:0046069	6	cGMP catabolic process
GO:0060534	6	trachea cartilage development
GO:0071321	6	cellular response to cGMP
GO:1902262	6	apoptotic process involved in patterning of blood vessels
GO:1902913	6	positive regulation of neuroepithelial cell differentiation
GO:0006930	7	substrate-dependent cell migration, cell extension
GO:0015810	7	aspartate transport
GO:0045634	7	regulation of melanocyte differentiation
GO:0051657	7	maintenance of organelle location
GO:0070305	7	response to cGMP
GO:0015740	8	C4-dicarboxylate transport
GO:0047484	8	regulation of response to osmotic stress
GO:0006621	9	protein retention in ER lumen
GO:0046886	9	positive regulation of hormone biosynthetic process
GO:0048340	9	paraxial mesoderm morphogenesis
GO:0035437	10	maintenance of protein localization in endoplasmic reticulum

Table 3.30.: GO MF (p-value < 0.01) associated with genes selected by MALA of category size between 5 and 100

GO Slim Term	Size	GO Slim Term Description
GO:0045545	5	syndecan binding
GO:0015556	6	C4-dicarboxylate transmembrane transporter activity
GO:0045295	12	gamma-catenin binding
GO:0031996	14	thioesterase binding
GO:0043274	16	phospholipase binding
GO:0005234	18	extracellular-glutamate-gated ion channel activity
GO:0008301	18	DNA binding, bending
GO:0004970	19	ionotropic glutamate receptor activity
GO:0030552	24	cAMP binding
GO:0008066	27	glutamate receptor activity
GO:0043548	27	phosphatidylinositol 3-kinase binding
GO:0046915	28	transition metal ion transmembrane transporter activity
GO:0030551	37	cyclic nucleotide binding
GO:0030295	51	protein kinase activator activity
GO:0019209	57	kinase activator activity
GO:0004702	83	receptor signaling protein serine/threonine kinase activity
GO:0015294	88	solute:cation symporter activity

Table 3.31.: GO CC (p-value < 0.01) associated with genes selected by MALA of category size between 5 and 100

GO Slim Term	Size	GO Slim Term Description
GO:0071204	6	histone pre-mRNA 3'end processing complex
GO:0071541	7	eukaryotic translation initiation factor 3 complex, eIF3m
GO:0031616	10	spindle pole centrosome
GO:0000780	16	condensed nuclear chromosome, centromeric region
GO:0043596	34	nuclear replication fork
GO:0097542	43	ciliary tip
GO:0032154	45	cleavage furrow
GO:0097610	45	cell surface furrow
GO:0032153	48	cell division site
GO:0032155	48	cell division site part
GO:0005876	50	spindle microtubule
GO:0005657	59	replication fork

3.2.4. Comparison of Methods

The comparison of gene sets resulting from each method and of associated over-represented GO terms is visualized in Figure 3.5.

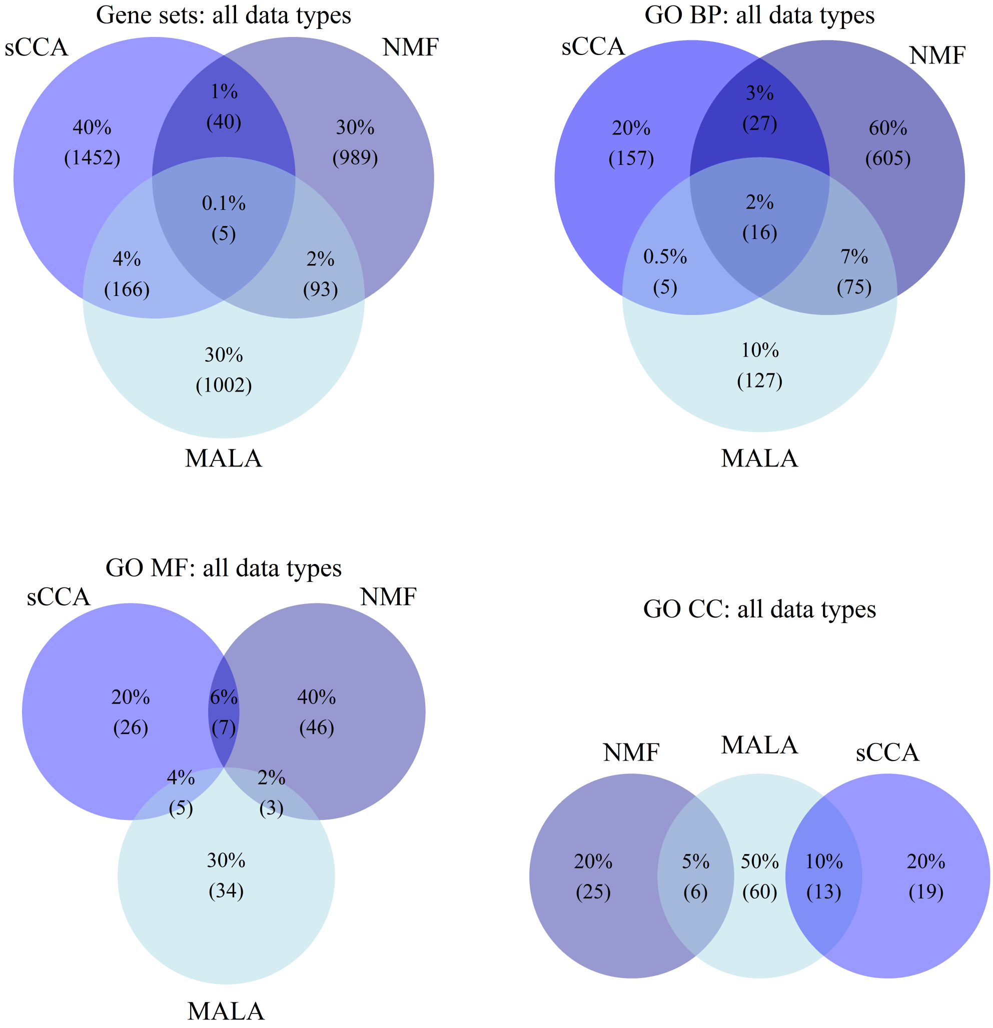

Figure 3.5.: Venn diagrams of gene sets merged from all data types and over-represented GO terms associated with gene sets.

Five genes were selected by all methods; they are presented in Table 3.32. An overview of numbers of genes selected by each method known to be involved in cancer pathways is given in Table 3.33. Genes involved in the *Pahtways in cancer* pathway from the KEGG database that were selected by at least two methods are shown in Table 3.34. The overlap of GO terms of category biological process are listed in Table 3.35. The overlap of GO terms in categories molecular function and cellular component associated with the genes selected by at least two of three methods are listed in Tables 3.36 and 3.37.

Table 3.32.: Genes selected by all methods.

Symbol	Gene name
GLIPR2	GLI pathogenesis-related 2
PTCH1	patched 1
TCEAL2	transcription elongation factor A (SII)-like 2
TTYH1	tweety family member 1
C7orf25	chromosome 7 open reading frame 25

Table 3.33.: Total number of genes involved in *Pahtways in cancer* and number of genes retrieved by each method.

total	sCCA	NMF	MALA
310	37	25	30

Table 3.34.: Genes involved in cancer pathways selected by at least two methods.

Symbol	Gene name	sCCA	NMF	MALA
PTCH1	patched 1	✓	✓	✓
MAP2K2	mitogen-activated protein kinase kinase 2	✓	✓	
CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa	✓	✓	
RHOA	ras homolog family member A	✓		✓
FZD1	frizzled class receptor 1	✓		✓
CCDC6	coiled-coil domain containing 6	✓		✓
FOXO1	forkhead box O1	✓		✓
MAX	MYC associated factor X	✓		✓
RXRG	retinoid X receptor, gamma		✓	✓
FGF2	fibroblast growth factor 2 (basic)		✓	✓
PDGFA	platelet-derived growth factor alpha polypeptide		✓	✓
WNT11	wingless-type MMTV integration site family, member 11		✓	✓

Table 3.35.: Overlap of GO BP

GO Slim Term ID	GO Slim Term Description
GO:0007167	enzyme linked receptor protein signaling pathway
GO:0007275	multicellular organismal development
GO:0007399	nervous system development
GO:0007417	central nervous system development
GO:0007420	brain development
GO:0009653	anatomical structure morphogenesis
GO:0022008	neurogenesis
GO:0030154	cell differentiation
GO:0030182	neuron differentiation
GO:0035239	tube morphogenesis
GO:0048468	cell development
GO:0048546	digestive tract morphogenesis
GO:0048699	generation of neurons
GO:0048729	tissue morphogenesis
GO:0048731	system development
GO:0048856	anatomical structure development

Table 3.36.: Overlap of GO MF

GO Slim Term ID	GO Slim Term Description	sCCA	NMF	MALA
GO:0000975	regulatory region DNA binding	✓	✓	
GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity	✓	✓	
GO:0001012	RNA polymerase II regulatory region DNA binding	✓	✓	
GO:0001067	regulatory region nucleic acid binding	✓	✓	
GO:0001071	nucleic acid binding transcription factor activity	✓	✓	
GO:0003700	sequence-specific DNA binding transcription factor activity	✓	✓	
GO:0044212	transcription regulatory region DNA binding	✓	✓	
GO:0004672	protein kinase activity	✓		✓
GO:0004674	protein serine/threonine kinase activity	✓		✓
GO:0005488	binding	✓		✓
GO:0005515	protein binding	✓		✓
GO:0016773	phosphotransferase activity, alcohol group as acceptor	✓		✓
GO:0005102	receptor binding		✓	✓
GO:0015267	channel activity		✓	✓
GO:0022803	passive transmembrane transporter activity		✓	✓

Table 3.37.: Overlap of GO CC

GO Slim Term ID	GO Slim Term Description	sCCA	NMF	MALA
GO:0005622	intracellular	✓		✓
GO:0031974	membrane-enclosed lumen	✓		✓
GO:0031981	nuclear lumen	✓		✓
GO:0032991	macromolecular complex	✓		✓
GO:0043229	intracellular organelle	✓		✓
GO:0043231	intracellular membrane-bounded organelle	✓		✓
GO:0043233	organelle lumen	✓		✓
GO:0044424	intracellular part	✓		✓
GO:0044428	nuclear part	✓		✓
GO:0070013	intracellular organelle lumen	✓		✓
GO:0072372	primary cilium	✓		✓
GO:1902494	catalytic complex	✓		✓
GO:1990234	transferase complex	✓		✓
GO:0005576	extracellular region		✓	✓
GO:0031982	vesicle		✓	✓
GO:0031988	membrane-bounded vesicle		✓	✓
GO:0043005	neuron projection		✓	✓
GO:0043235	receptor complex		✓	✓
GO:0044421	extracellular region part		✓	✓

The sets of selected genes from each data type and sets of over-represented GO terms of categories BP, MF and CC associated with the selected genes are displayed in Figures 3.6, 3.7 and 3.8.

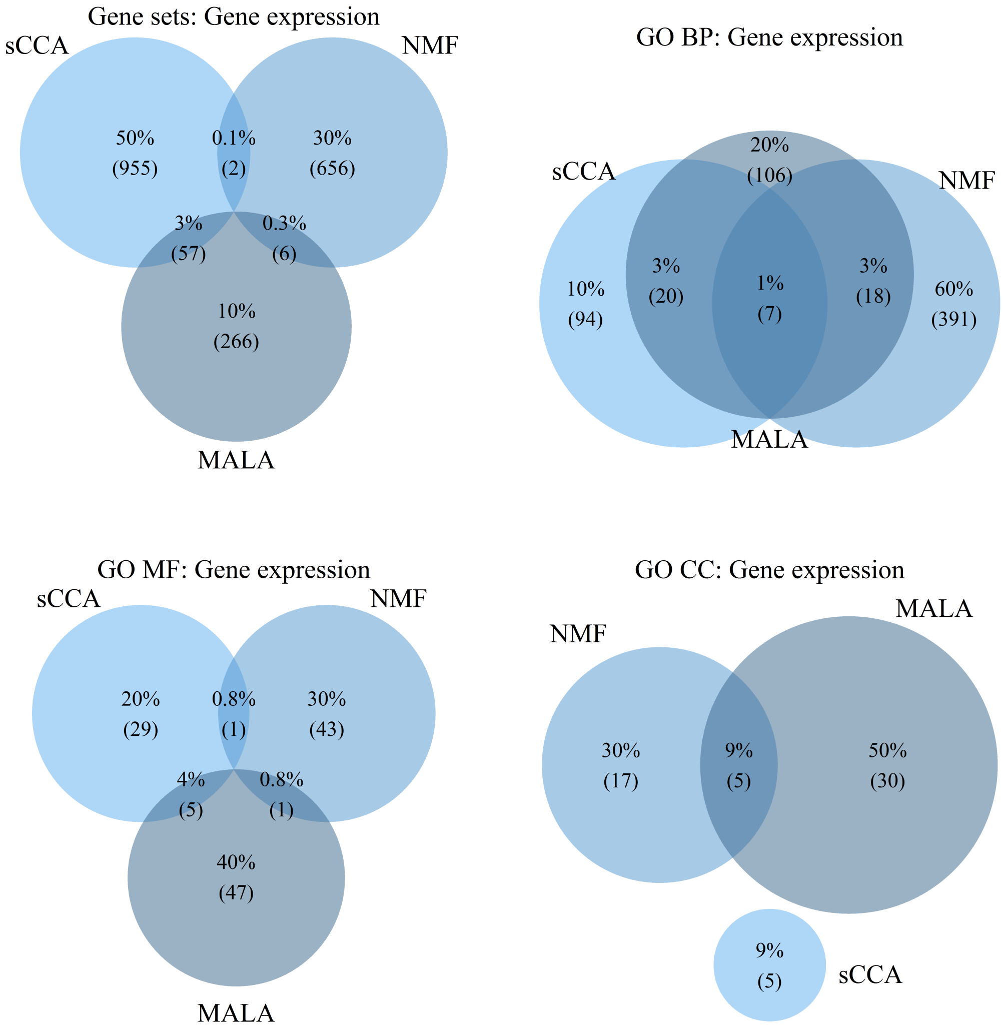

Figure 3.6.: Venn diagrams of gene sets and over-represented GO terms extracted from the gene expression dataset.

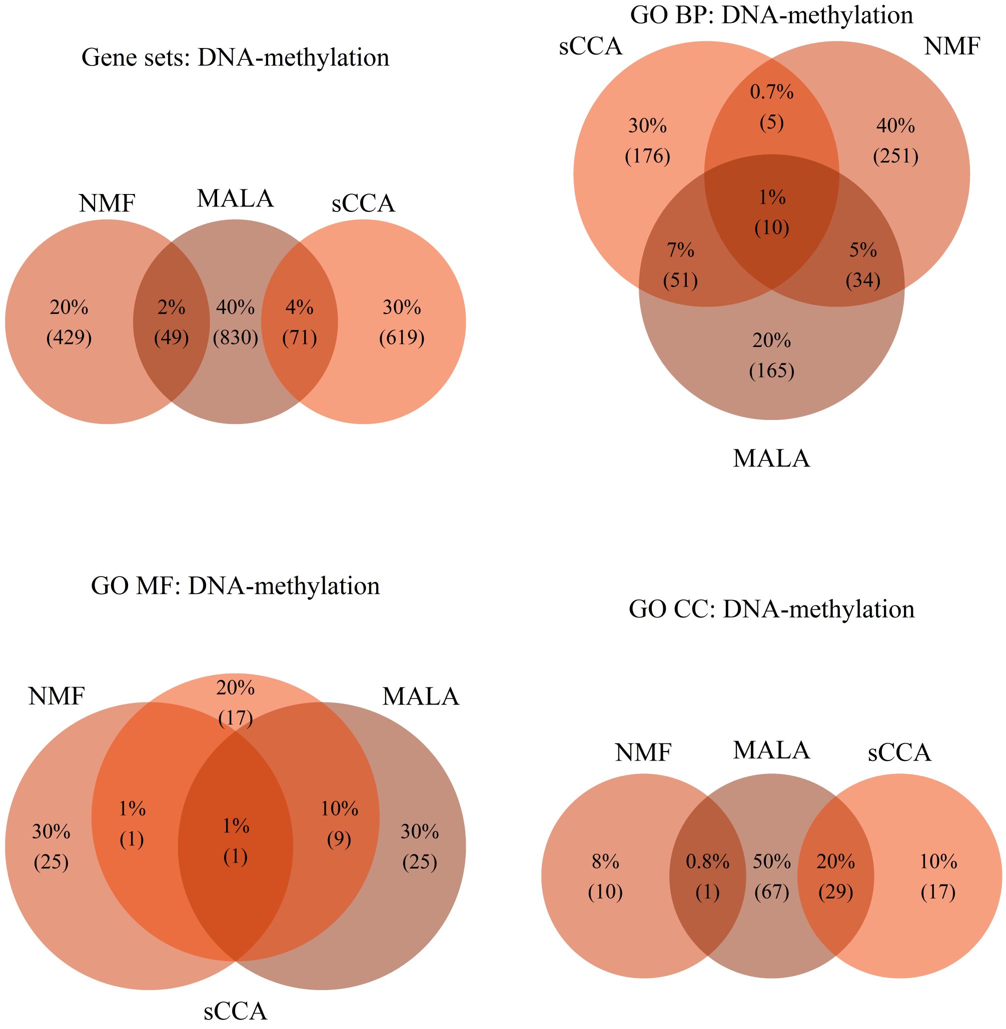

Figure 3.7.: Venn diagrams of gene sets and over-represented GO terms extracted from the DNA-methylation dataset.

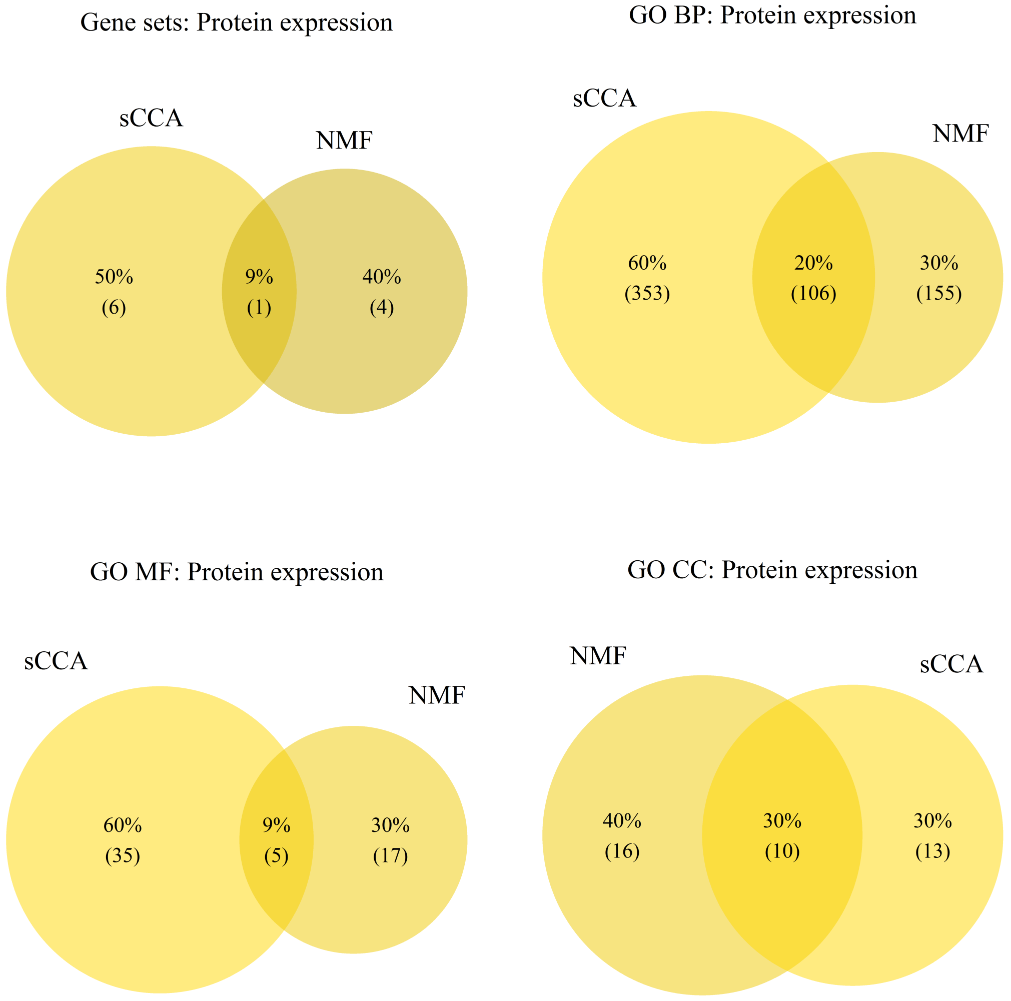

Figure 3.8.: Venn diagrams of gene sets and over-represented GO terms extracted from the protein expression dataset.

4. Discussion

The aim of the presented master's thesis was to compare three integrative analysis methods. These were applied on synthetic and biological datasets and their results were assessed on gene level and on the level of associated GO terms. The biological datasets comprise measurements from three different biological levels: the transcript level, represented by the gene expression dataset, the gene level, represented by the DNA-methylation dataset and the protein level represented by the protein expression dataset. The samples were obtained from patients suffering from breast invasive carcinoma and originate from solid tumor and adjacent normal tissue. The synthetic data comprises gene expression datasets from simulated microarray experiments based on co-expression networks derived from the biological datasets.

Integrative analysis methods have become more and more important recently and are used to derive information from large datasets obtained with high-throughput technologies on different biological levels for the same samples/conditions. They aim to identify a subset of candidate genes to account for the development of complex diseases which should thus be subjected to further analysis. The relevance and accuracy of integrative analysis methods can hardly be assessed because most of the mechanisms underlaying the development of complex diseases are currently unknown and the functional annotation of genes in the laboratory is time and costs demanding. A promising strategy to validate the results of integrative analysis is, hence, to compare them with already validated genes or with the results of other methods. Here the results of the three methods are compared with each other and to already validated genes involved in the KEGG cancer pathway. The KEGG PATHWAY database provides manually annotated pathway maps in a variety of biological domains such as *Human Diseases*. The maps describe and visualize networks of molecular interactions and reactions. The map of *Pathways in Cancer* from the KEGG PATHWAY database is shown in Figure 4.1.

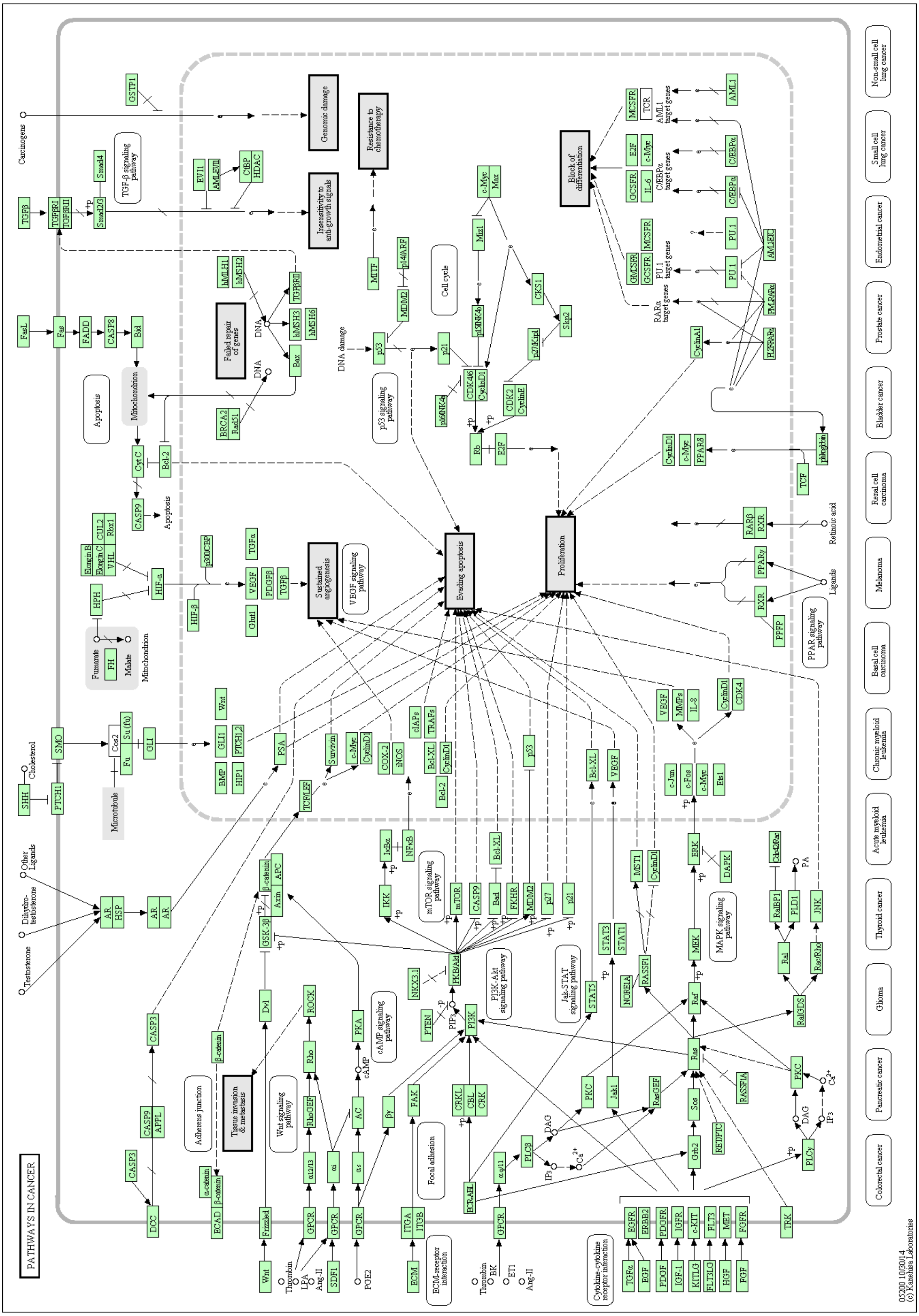

Figure 4.1.: *Pathway in Cancer* map from the KEGG PATHWAY database.

4.1. Mathematical Concept of Investigated Methods

The three integrative analysis methods compared in this master's thesis are based on different mathematical concepts. This fact was one of the motivating aspects for the selection of the methods. The sCCA aims to find sparse canonical weight-vectors that result in the one-dimensional projections of three or more datasets with the highest pair-wise correlation. The non-zero elements in the canonical weight-vectors are of interest due to their association with correlated features accounting for the inherent structure of the datasets.

The goal of NMF is to identify a number of building blocks which are common to all (three or more) datasets and can be used for the reconstruction of the original datasets as their positively weighted sum. The factorization is subjected to the constraint that the joint reconstruction error is minimized. Features and samples in the datasets that are associated with high weights are grouped and form md-modules. The sCCA and the NMF have in common that the datasets are decomposed and similarities in terms of correlation (sCCA) or significant contribution (NMF) are considered to select features that account for the inherent structure of datasets. Both methods integrate datasets originating from three different biological levels (gene expression, DNA-methylation and protein expression datasets) sampled under the same condition (tumor tissue) and aim to find commonalities within datasets. The columns of the datasets (containing the features) are standardized before the application of either method.

In contrast to that, MALA is based on a machine learning approach. It operates on datasets obtained from the same patient under different conditions (tumor and normal tissue) and aims to find the differences in the two classes. Due to the lack of protein expression data originating from normal tissue, only the gene expression and the DNA-methylation datasets could be subjected to the analysis with MALA. MALA derives a classification model based on a training set of samples which is validated on the remaining samples (test set). The main functionality of MALA is the solution of a feature selection problem with a GRASP algorithm. The features selected in this procedure are assembled to a classification model, a set of logic formulas. Since each feature in the datasets is assessed for classification capacity individually, a transformation of values is not necessary.

4.1.1. Advantages and Drawbacks of the Three Integrative Analysis Methods

An advantage of sCCA is that it is available as an R-package and the analysis of datasets is straightforward and intuitive. A drawback of the method, considering the purpose of feature selection, is that the number of non-zero weights in the canonical weight-vectors must be rather large to achieve significantly correlated projections. At the expense of correlation, the number of non-zero elements and thus the number of selected features is reduced to a more convenient number by decreasing the penalty terms on the canonical weight-vectors.

The NMF, as emphasized by Kim and Tidor [37], not only detects correlations across the whole set of features or samples but is also able to unravel local similarities limited to a subset of features or samples in the datasets. An elementary shortcoming of the NMF is that the number of building blocks, which is equal to the rank of the approximation, and the threshold for the assignment of a feature to a md-module must be specified in advance by the user. This is a complex and time consuming task and the method would benefit from an automatic selection of parameters.

MALA provides a comprehensive set of parameters adjustable for the analysis. A clustering step is implemented in the method, however, this could not be used because the inference of the classification model failed when clustering was activated. A similar problem occurred, when the sampling type was set to *cross validation*. The program terminated incorrectly with a *segmentation fault*. Thus, the alternative sampling type, *random percentage split* was chosen. For each subset the feature selection problem is solved and the classification model is derived. Each classification model represented by the logic formulas comprising the selected features is validated on the samples in the test set. The maximum number of features to be selected - regardless of the number of features comprised by the datasets - is 60. As a consequence, the average performance of the classification models derived for the relatively small synthetic datasets is quite satisfying. The performance of the classification models derived form the substantially larger biological dataset, however, was not as good. In order to compensate this effect, the number of subsets was set to 10 for the synthetic datasets and to 100 for the biological datasets and the features in the merged formulas derived from all subsets are considered as the selected feature set. This results in a selected features set of size comparable to those resulting from the other methods. An advantage of MALA is that the complex features selection problem, which grows quadratically with the number of features in the dataset, is solved using a heuristic approach with a remarkably small effort of time. On the other hand, this approach has the disadvantage of only being able to find local solutions.

4.2. Comparison on the Feature and GO Term Levels

4.2.1. Resulting Sets of Genes and GO Terms

In this section, influencing factors on the size of the resulting sets of features and GO terms are discussed.

Generally speaking, the resulting sets of candidate genes and associated GO terms respectively are rather big and the actual size of the gene sets was limited to a maximum of 5% of the features in each dataset for the sCCA and the NMF. The maximum number of features selected in one run by MALA is 60. This is the maximum number of features employed in the classification model derived from one training set. In order to compensate for this difference, the features of 100 runs of MALA were accumulated to achieve a comparable number of selected features by MALA. The feature sets resulting from the two conceptually more similar methods, sCCA and NMF, tend to be larger than the feature set resulting from the third method, MALA. A possible reason for that may be that the sCCA and the NMF are based on a top-down approach, while MALA pursues a bottom-up strategy. The sCCA and the NMF start from the whole feature set and seek to reduce the number of features by the introduction of certain criteria. In contrast, MALA subsequently adds one feature at a time during an iterative search procedure.

The difference in size between the analyzed datasets represents an additional reason for the different sizes of the resulted features sets. The synthetically generated datasets comprise considerably fewer features than the biological datasets. The synthetic datasets based on the co-expression network derived from the gene expression datasets of tumor/normal tissue comprise 390/2 748 nodes, the network derived from the DNA-methylation datasets of tumor/normal tissue comprise 2 471/2 809 and the network derived from protein expression dataset in tumor tissue consists of 68 nodes respectively. Similar differences in dataset size are observed in the biological datasets. These comprise 19 769/19 716 features in the gene expression dataset of tumor/normal tissue; 13 627/14 300 features in DNA-methylation dataset of tumor/normal tissue; and 118 features in the protein expression dataset of tumor tissue.

The number of resulting GO terms was reduced by limiting the category size. The over-represented GO terms in three categories BP, MF and CC comprise general terms which are associated with a large number of genes, as well as very specific terms. This is indicated by the category size of the GO terms. The lower limit of the category size of a GO term considered in the analysis was set to 5 in order to exclude the terms which are associated with very few genes. This was done because GO terms of small category size are prone to random enrichment. GO terms with large category sizes represent general biological processes which are not suitable for the characterization of the obtained results.

4.2.2. Overlap in Synthetic Datasets

In this section, the overlap on feature level and GO term level of results in synthetic datasets is discussed. For the synthetic data, an overview of the resulting overlaps in terms of percentage of the total number of selected feature and associated GO terms respectively, derived from the Venn diagrams in section 3.1.4 is given in Table 4.1. It can be observed that there are no features which are selected by all methods. The pair-wise overlap exclusively results from the features selected in the DNA-methylation dataset. There is no overlap of features selected by the methods originating from the gene expression or the protein expression dataset. This might be associated with the fact that the DNA-methylation dataset represents the largest synthetic dataset.

On the GO term level, several GO terms in category BP were found by two of three methods. The results of sCCA and NMF on the gene expression datasets produce a considerable overlap of 11% (46 terms) in category BP. The results of sCCA and MALA on the DNA-methylation dataset produce a remarkable overlap of 29% (7 terms) in the category CC. For the GO terms associated with the merged gene set, an overlap of 1 GO term (representing 3% in either case) in categories MF and CC respectively, is also observed. In summary, in the synthetic datasets there are no genes that were selected by all methods. However, considering the GO terms associated with the selected features, an overlap can be observed in at least one GO category in all datasets except the protein expression dataset.

4.2.3. Overlap in Biological Datasets

In this section, the overlap on feature level and GO term level of results in biological datasets is discussed. The resulting overlaps between the three methods on the feature and GO term level derived from the Venn diagrams in section 3.2.4 is shown in Table 4.2. A small number of 5 genes representing 0.1% in the merged set of features is selected by all three methods. Considering the three different biological levels separately, there are no features which were selected by all three methods. This means that features selected by a method on one biological level was selected on another biological level by other methods. This emphasizes the relevance of integrative analysis methods.

In general, an overlap of all GO categories can be observed for the merged set of features. A high overlap of the results of sCCA and NMF is especially observed for the protein expression dataset. The highest overlap of the results of sCCA and MALA can be observed for the DNA-methylation dataset in categories MF and CC.

Similar to the results on the synthetic datasets, the overlap is increased on the more general level of associated over-represented GO terms. There is an overlap of associated GO terms of at least one category on each biological level, as well as for the merged set of selected features. Specifically, there is an overlap in category BP of 7 (1%) and 10 (1%) GO terms on gene expression and DNA-methylation level respectively and an overlap of 16 GO terms (2%) associated with the merged feature set. The GO terms associated with the features selected in the DNA-methylation datasets even show an overlap of 1 term in the MF category. An overview of the resulting overlaps in terms of percentage of the total number of selected feature and associated GO terms respectively is given in Table 4.2.

Table 4.1.: Overview of overlaps on different levels for the synthetic datasets.

	Feature level				GO term level											
					BP				MF				CC			
	GE	MET	PE	merged	GE	MET	PE	merged	GE	MET	PE	merged	GE	MET	PE	merged
sCCA ∩ NMF ∩ MALA	-	-	-	-	1%	-	-	-	-	3%	-	3%	-	9%	-	3%
sCCA ∩ NMF	-	2%	-	2%	11%	1%	-	5%	-	6%	-	6%	-	9%	-	3%
sCCA ∩ MALA	-	2%	-	2%	1%	3%	-	1%	-	3%	-	3%	-	29%	-	6%
NMF ∩ MALA	-	2%	-	2%	3%	-	-	-	-	6%	-	3%	2%	9%	-	3%

Table 4.2.: Overview of overlaps on different levels for the biological datasets.

	Feature level				GO term level											
					BP				MF				CC			
	GE	MET	PE	merged	GE	MET	PE	merged	GE	MET	PE	merged	GE	MET	PE	merged
sCCA ∩ NMF ∩ MALA	-	-	-	0.1%	1%	1%	-	2%	-	1%	-	-	-	-	-	-
sCCA ∩ NMF	0.1%	-	9%	1.1%	1%	1.7%	20%	5%	0.8%	2%	9%	6%	-	-	30%	-
sCCA ∩ MALA	3%	4%	-	4.1%	4%	8%	-	2.5%	4%	11%	-	4%	-	20%	-	10%
NMF ∩ MALA	0.3%	2%	-	2.1%	4%	6%	-	9%	0.8%	1%	-	2%	9%	0.8%	-	5%

4.3. Biological Annotation of Results

The genes and GO terms derived from the synthetic datasets are not analyzed regarding their biological annotation or their overlap with genes known to be involved in *Pathways in Cancer* because the regulatory dependencies in the biological datasets could not be reproduced in the synthetic datasets. Instead, the regulatory interactions between the nodes in the co-expression networks, which serve as basis for the synthetic datasets, were set randomly.

The number of genes selected in the biological datasets which are known to be involved in *Pathways in Cancer*, is notably quite the same for each method (37, 25 and 30 of 310 genes in the *Pathways in Cancer* pathway). These gene sets represent 10% of the genes in *Pathways in Cancer* and thus, it is not likely that they were selected by coincidence. One may draw the conclusion that the methods are equally suitable to retrieve genes involved in cancer development.

Genes originating from the biological datasets which were selected by all methods are shown in Table 3.32. Genes which were selected by at least two of three methods and which are, additionally, involved in the *Pathways in Cancer* pathways from the KEGG database are listed in Table 3.34. They are analyzed in regard to their biological meaning and importance. Among the genes selected by all three methods, the *tweety family member 1 (TTYH1)* gene was recently shown to be related to pediatric brain tumors [61]. Alterations of the *patched 1 (PTCH1)* gene such as aberrant frequency of methylation were associated with the development of cervical carcinoma [62]. *Transcription elongation factor A (SII)-like 2 (TCEAL2), chromosome 7 open reading frame 25 (C7orf25)* and *GLI pathogenesis-related 2 (GLIPR2)* could not be directly associated with cancerogenesis. However, the name of TCEAL2 suggests a general influence in transcription regulation.

Among the genes involved in *Pathways in Cancer* which were selected by at least two methods the *Mitogen-activated protein kinase kinase 2 (MAP2K2)* for example may play a role in cell proliferation [63]. Disorders of the expression of *platelet-derived growth factor alpha* *polypeptide (PDGFA)* are associated with neoplasia and, hence, with tumorigenic processes, since it is involved in the regulation of cell proliferation [64]. Another example, the *MYC associated factor X (MAX)* can bind to *Myc* which is known to be an oncoprotein due to its involvement in cell proliferation, differentiation and apoptosis, according to the gene summary page in Entrez Gene [65]. Another prominent example, which is described in Entrez Gene, is the *ras homolog family member A (RHOA)* which takes influence on tumor cell proliferation and metastasis. Two widely-known representatives involved in the development of breast invasive carcinoma are BRCA1 and BRCA2, also known as *breast cancer 1, early onset* and *breast cancer 2, early onset*. Mutations of these genes are known to increase the probability of genetically caused breast cancer. However, the role of these genes is due to mutations which were not part of this analysis. This could be the reason why these two genes have not been selected by any of the three methods.

4.4. Conclusion

The three integrative analysis methods compared in this master's thesis yield rather comprehensive lists of selected features which produce a modest overlap on the gene level. Not even the results of sCCA and NMF, which are based on more similar mathematical concepts, produce a considerable overlap. Significantly over-represented GO terms derived from the selected genes are more congruent. About 10% of the features known to be involved in *Pathways in Cancer* from the KEGG database are retrieved by each method, however, only one of them is selected by all methods.

4.5. Outlook

In order to evaluate and validate the results of the three integrative analysis methods, the role of the genes selected by each method in the development of complex diseases must be revealed. A comprehensive review of the biological annotation of the selected genes would shed light on the biological homogeneity of the selected feature sets. Moreover, the selected feature sets could be compared to the results of further integrative analysis methods. Additionally, a ranking of the selected features would be of interest.

Bibliography

[1] Crick F: **Central dogma of molecular biology**. *Nature* 1970. 227(5258):561–563.

[2] Hunter DJ: **Gene–environment interactions in human diseases**. *Nature Reviews Genetics* 2005. 6(4):287–298.

[3] Alter O and Golub GH: **Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between dna replication and rna transcription**. *Proceedings of the National Academy of Sciences of the United States of America* 2004. 101(47):16577–16582.

[4] Golub GH and Van Loan CF: **Matrix computations**. Johns Hopkins Univ Press, Baltimore, MD, USA, 3 edition, 1996.

[5] Berger JA, Hautaniemi S, Mitra SK and Astola J: **Jointly analyzing gene expression and copy number data in breast cancer using data reduction models**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2006. 3(1):2–16.

[6] Van Loan CF: **Generalizing the singular value decomposition**. *SIAM Journal on Numerical Analysis* 1976. 13(1):76–83.

[7] Ponnapalli SP, Saunders MA, Van Loan CF and Alter O: **A higher-order generalized singular value decomposition for comparison of global mrna expression from multiple organisms**. *PloS One* 2011. 6(12):e28072.

[8] Hotelling H: **Relations between two sets of variates**. *Biometrika* 1936. 28(3/4):321–377.

[9] Lê Cao KA, Martin PG, Robert-Granié C and Besse P: **Sparse canonical methods for biological data integration: application to a cross-platform study**. *BMC Bioinformatics* 2009. 10(34).

[10] Waaijenborg S, Verselewel de Witt Hamer PC and Zwinderman AH: **Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis**. *Statistical Applications in Genetics and Molecular Biology* 2008. 7(1):Article 3.

[11] Zou H and Hastie T: **Regularization and variable selection via the elastic net**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005. 67(2):301–320.

[12] Witten DM, Tibshirani R and Hastie T: **A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis**. *Biostatistics* 2009. 10(3):515–34.

[13] Lin D, Zhang J, Li J, Calhoun VD, Deng HW and Wang YP: **Group sparse canonical correlation analysis for genomic data integration**. *BMC Bioinformatics* 2013. 14(1):article 245.

[14] Simon N, Friedman J, Hastie T and Tibshirani R: **A sparse-group lasso**. *Journal of Computational and Graphical Statistics* 2013. 22(2):231–245.

[15] Dolédec S and Chessel D: **Co-inertia analysis: an alternative method for studying species–environment relationships**. *Freshwater Biology* 1994. 31(3):277–294.

[16] Fagan A, Culhane AC and Higgins DG: **A multivariate analysis approach to the integration of proteomic and gene expression data**. *Proteomics* 2007. 7(13):2162–2171.

[17] Lee DD and Seung HS: **Learning the parts of objects by non-negative matrix factorization**. *Nature* 1999. 401(6755):788–791.

[18] Brunet JP, Tamayo P, Golub TR and Mesirov JP: **Metagenes and molecular pattern discovery using matrix factorization**. *Proceedings of the National Academy of Sciences* 2004. 101(12):4164–4169.

[19] Zhang S, Liu CC, Li W, Shen H, Laird PW and Zhou XJ: **Discovery of multi-dimensional modules by integrative analysis of cancer genomic data**. *Nucleic Acids Research* 2012. 40(19):9379–91.

[20] Draper NR and Smith H: **Applied regression analysis**. John Wiley & Sons, New York City, USA, 3 edition, 2014.

[21] Kohl M, Megger DA, Trippler M, Meckel H, Ahrens M, Bracht T, Weber F, Hoffmann AC, Baba HA, Sitek B *et al.*: **A practical data processing workflow for multi-omics projects**. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 2014. 1844(1):52–62.

[22] Lê Cao KA, Rossouw D, Robert-Granié C and Besse P: **A sparse PLS for variable selection when integrating omics data**. *Statistical Applications in Genetics and Molecular Biology* 2008. 7(1):Article 35.

[23] Shen H and Huang JZ: **Sparse principal component analysis via regularized low rank matrix approximation**. *Journal of Multivariate Analysis* 2008. 99(6):1015–1034.

[24] Härdle W and Simar L: **Applied multivariate statistical analysis**. Springer Science & Business Media, Berlin, Germany, 2007.

[25] Shen R, Wang S and Mo Q: **Sparse integrative clustering of multiple omics data sets**. *The Annals of Applied Statistics* 2013. 7(1):269–294.

[26] Shen R, Olshen AB and Ladanyi M: **Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis**. *Bioinformatics* 2009. 25(22):2906–2912.

[27] Cao H, Duan J, Lin D and Wang YP: **Sparse representation based clustering for integrated analysis of gene copy number variation and gene expression data**. *International Journal of Computers and Applications (IJCA)* 2012. 19(2):131–144.

[28] Gusenleitner D, Howe EA, Bentink S, Quackenbush J and Culhane AC: **iBBiG: iterative binary biclustering of gene sets**. *Bioinformatics* 2012. 28(19):2484–2492.

[29] Kohavi R and Provost F: **Glossary of terms**. *Machine Learning* 1998. 30(2-3):271–274.

[30] Breiman L: **Random forests**. *Machine Learning* 2001. 45(1):5–32.

[31] Reif DM, Motsinger A, McKinney B, Crowe Jr JE, Moore JH *et al.*: **Feature selection using a random forests classifier for the integrated analysis of multiple data types**. In **Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB'06. 2006 IEEE Symposium on**. IEEE, 2006 pages 1–8.

[32] Weitschek E, Felici G and Bertolazzi P: **Mala: A microarray clustering and classification software**. In **Database and Expert Systems Applications (DEXA), 2012 23rd International Workshop on**. IEEE, 2012 pages 201–205.

[33] Tomescu OA, Mattanovich D and Thallinger GG: **Integrative omics analysis. a study based on plasmodium falciparum mrna and protein data**. *BMC Systems Biology* 2014. 8(Suppl 2):S4.

[34] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al.*: **Gene ontology: tool for the unification of biology**. *Nature Genetics* 2000. 25(1):25–29.

[35] Witten DM and Tibshirani RJ: **Extensions of sparse canonical correlation analysis with applications to genomic data**. *Statistical Applications in Genetics and Molecular Biology* 2009. 8(1):1–27.

[36] Witten DM, Tibshirani RJ, Gross S and Narasimhan B: **PMA: Penalized Multivariate Analysis**, 2013. URL `http://CRAN.R-project.org/package=PMA`. R package version 1.0.9.

[37] Kim PM and Tidor B: **Subsystem identification through dimensionality reduction of large-scale gene expression data**. *Genome Research* 2003. 13(7):1706–1718.

[38] Arisi I, D'Onofrio M, Brandi R, Felsani A, Capsoni S, Drovandi G, Felici G, Weitschek E, Bertolazzi P and Cattaneo A: **Gene expression biomarkers in the brain of a mouse model for Alzheimer's disease: mining of microarray data by logic classification and feature selection**. *Journal of Alzheimer's Disease* 2011. 24(4):721.

[39] Bertolazzi P, Felici G, Festa P and Lancia G: **Logic classification and feature selection for biomedical data**. *Computers & Mathematics with Applications* 2008. 55(5):889–899.

[40] Felici G and Truemper K: **A MINSAT approach for learning in logic domains**. *INFORMS Journal on Computing* 2002. 14(1):20–36.

[41] Felici G, Truemper K and Wang J: **The lsquare system for mining logic data**. *Encyclopedia of Data Warehousing and Mining* 2005. 2:693–697.

[42] Kurgan L and Cios KJ: **Caim discretization algorithm**. *IEEE Transactions on Knowledge and Data Engineering* 2004. 16(2):145–153.

[43] Resende MG: **Greedy randomized adaptive search procedures**. In CA Floudas and PM Pardalos, editors, **Encyclopedia of Optimization**, pages 1460–1469. Springer, New York, NY, USA, 2 edition, 2009.

[44] Truemper K: **Design of logic-based intelligent systems**. John Wiley & Sons, Hoboken, NJ, USA, 2004.

[45] R Development Core Team: **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[46] Chen H: **VennDiagram: Generate High-Resolution Venn and Euler Plots**, 2015. URL `http://CRAN.R-project.org/package=VennDiagram`. R package version 1.6.16.

[47] Kanehisa M and Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Research* 2000. 28(1):27–30.

[48] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al.*: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biology* 2004. 5(10):R80.

[49] Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T *et al.*: **Orchestrating high-throughput genomic analysis with bioconductor**. *Nature Methods* 2015. 12(2):115–121.

[50] Sales G, Calura E and Romualdi C: **graphite: GRAPH Interaction from pathway Topological Environment**, 2015. R package version 1.14.1.

[51] Falcon S and Gentleman R: **Using GOstats to test gene lists for GO term association.** *Bioinformatics* 2007. 23(2):257–8.

[52] Carlson M: **org.Hs.eg.db: Genome wide annotation for Human**, 2015. R package version 3.1.2.

[53] Van den Bulcke T, Van Leemput K, Naudts B, van Remortel P, Ma H, Verschoren A, De Moor B and Marchal K: **SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms**. *BMC Bioinformatics* 2006. 7(1):43.

[54] Benjamini Y and Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995. 57(1):289–300.

[55] Hall SR, Allen FH and Brown ID: **The crystallographic information file (CIF): a new standard archive file for crystallography**. *Acta Crystallographica Section A: Foundations of Crystallography* 1991. 47(6):655–685.

[56] Cui Q: **A network of cancer genes with co-occurring and anti-co-occurring mutations**. *PLoS One* 2010. 5(10):e13180.

[57] de Matos Simoes R: **Rsyntren: syntren.jar wrapper functions**, 2014. R package version 1.0.

[58] Zhu Y, Qiu P and Ji Y: **TCGA-assembler: open-source software for retrieving and processing TCGA data**. *Nature Methods* 2014. 11(6):599–600.

[59] Wickham H: **httr: Tools for Working with URLs and HTTP**, 2015. URL `http://CRAN.R-project.org/package=httr`. R package version 1.0.0.

[60] Razin A and Riggs AD: **DNA methylation and gene function**. *Science* 1980. 210(4470):604–610.

[61] Kleinman CL, Gerges N, Papillon-Cavanagh S, Sin-Chan P, Pramatarova A, Quang DAK, Adoue V, Busche S, Caron M, Djambazian H *et al.*: **Fusion of TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B isoform in the embryonal brain tumor ETMR**. *Nature Genetics* 2014. 46(1):39–44.

[62] Chakraborty C, Dutta S, Mukherjee N, Samadder S, Roychowdhury A, Roy A, Mondal RK, Basu P, Roychoudhury S and Panda CK: **Inactivation of PTCH1 is associated with the development of cervical carcinoma: clinical and prognostic implication**. *Tumor Biology* 2015. 36(2):1143–1154.

[63] Lee CS, Dykema KJ, Hawkins DM, Cherba DM, Webb CP, Furge KA and Duesbery NS: **MEK2 is sufficient but not necessary for proliferation and anchorage-independent growth of SK-MEL-28 melanoma cells**. *PloS One* 2011. 6(2):e17165.

[64] Yu JH, Ustach C and ChoiKim HR: **Platelet-derived growth factor signaling and human cancer**. *BMB Reports* 2003. 36(1):49–59.

[65] Maglott D, Ostell J, Pruitt KD and Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Research* 2005. 33(suppl 1):D54–D58.

Appendix A.

Centrality Measures of Co-Expression Networks

A.1. Degree

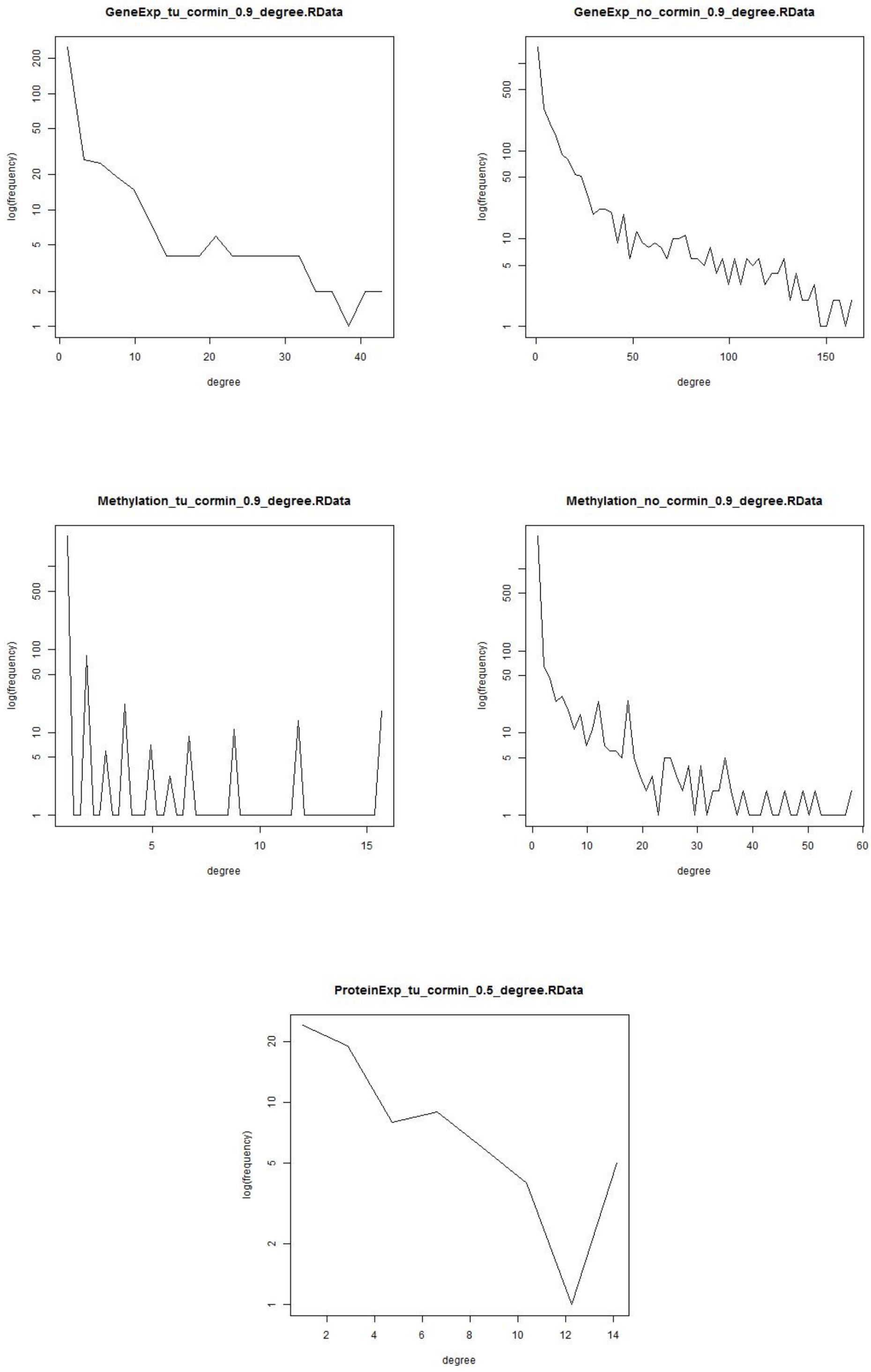

Figure A.1.: Degree of co-expression networks derived from biological datasets at cut-off values for Spearman's correlation coefficient of 0.9 or 0.5 respectively.

A.2. Betweenness

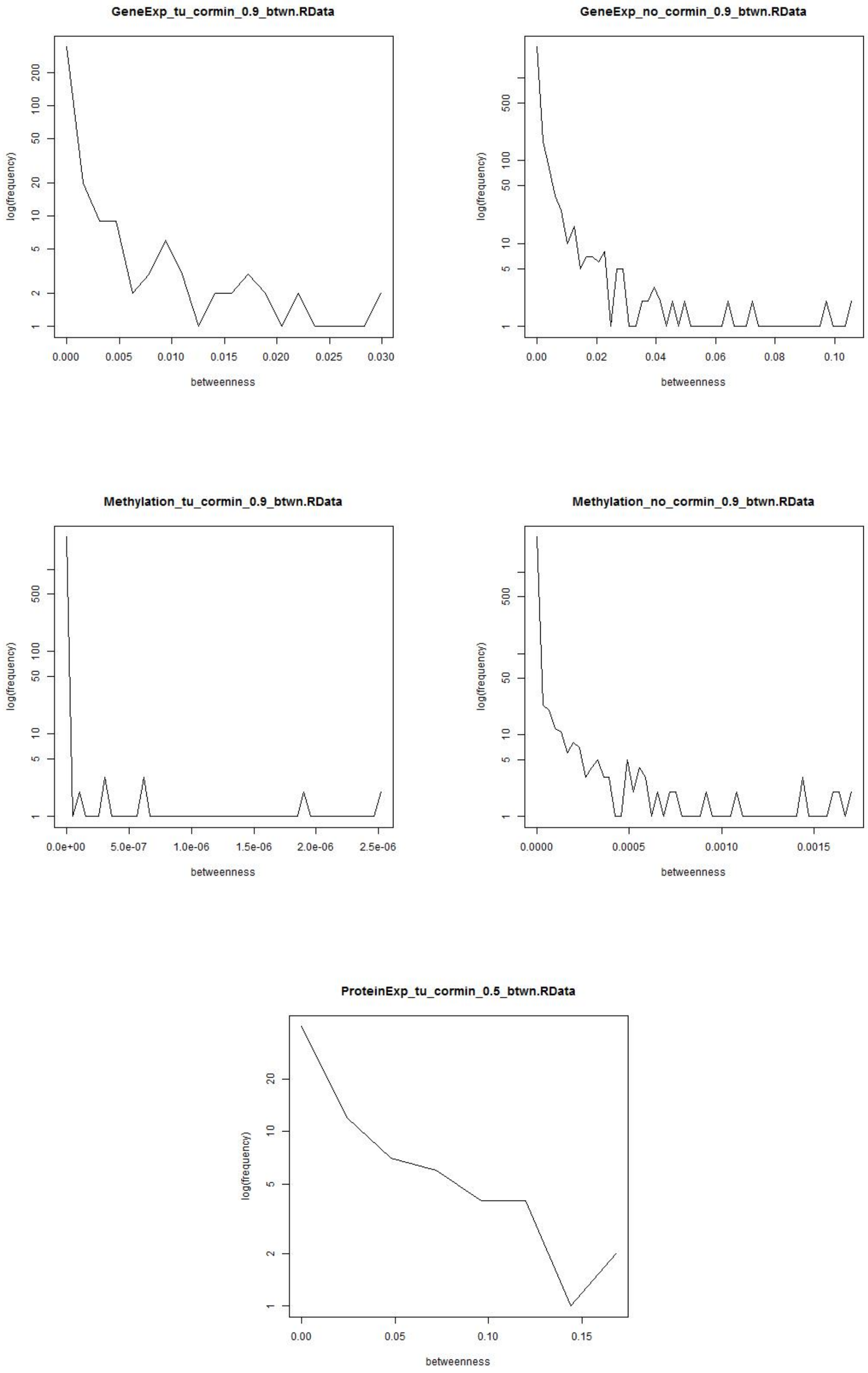

Figure A.2.: Betweenness of co-expression networks derived from biological datasets at cut-off values for Spearman's correlation coefficient of 0.9 or 0.5 respectively.

Appendix B.

Enrichment Analysis of Modules in Synthetic Gene Expression Datasets

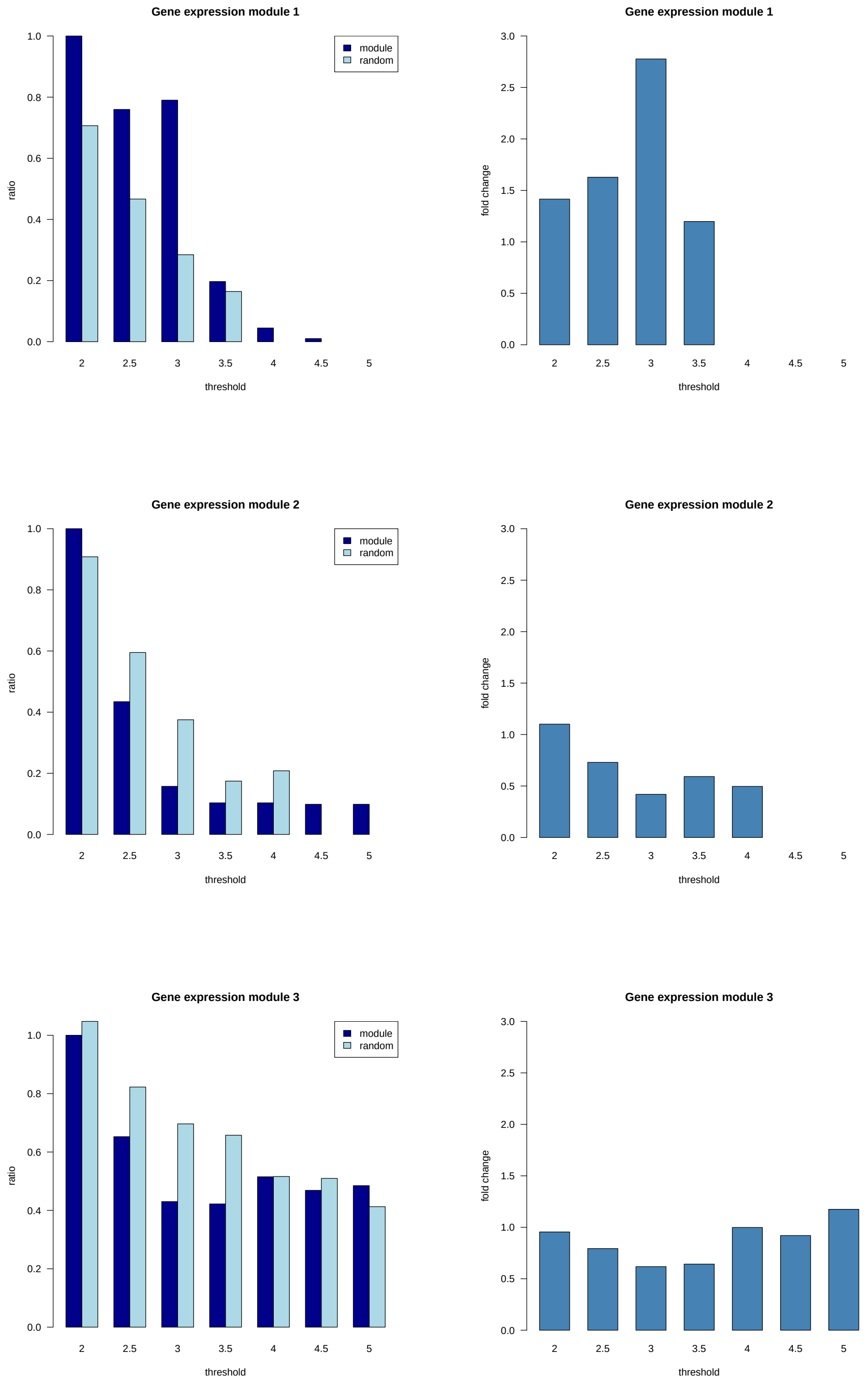

Figure B.1.: Enrichment ratios and fold-change of enrichment ratios of modules 1 to 3

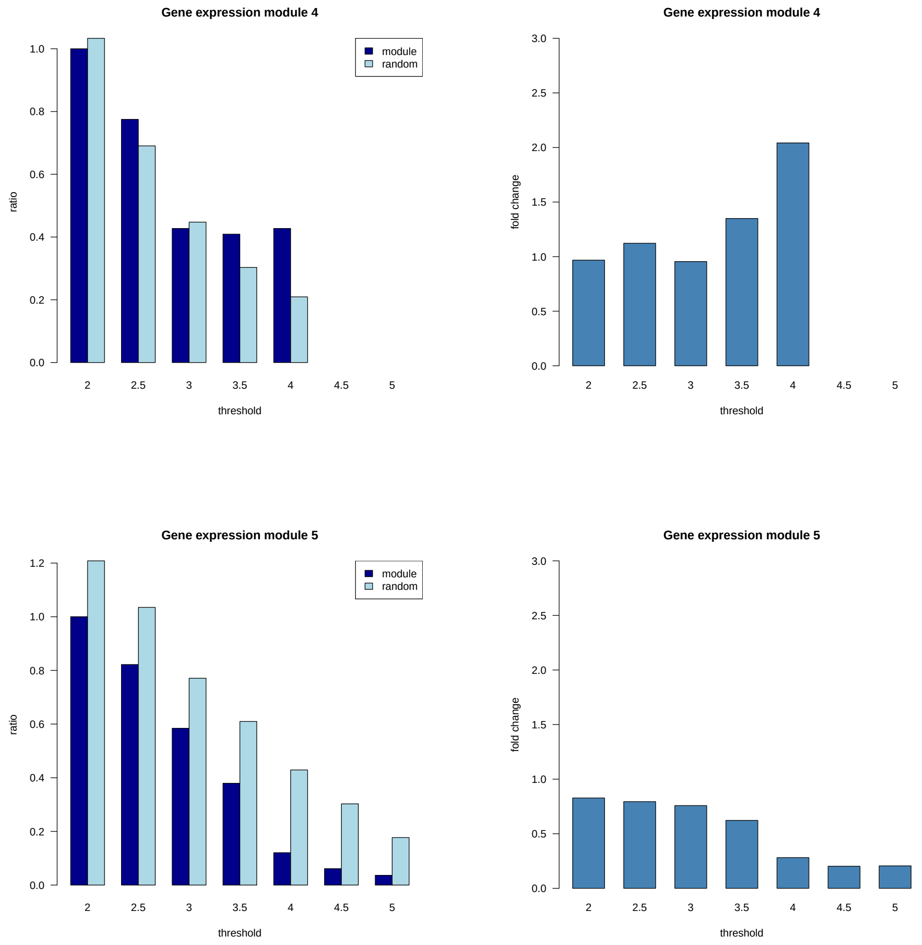

Figure B.2.: Enrichment ratios and fold-change of enrichment ratios of modules 4 and 5

Appendix C.

Enrichment Analysis of Modules in Biological Gene Expression Datasets

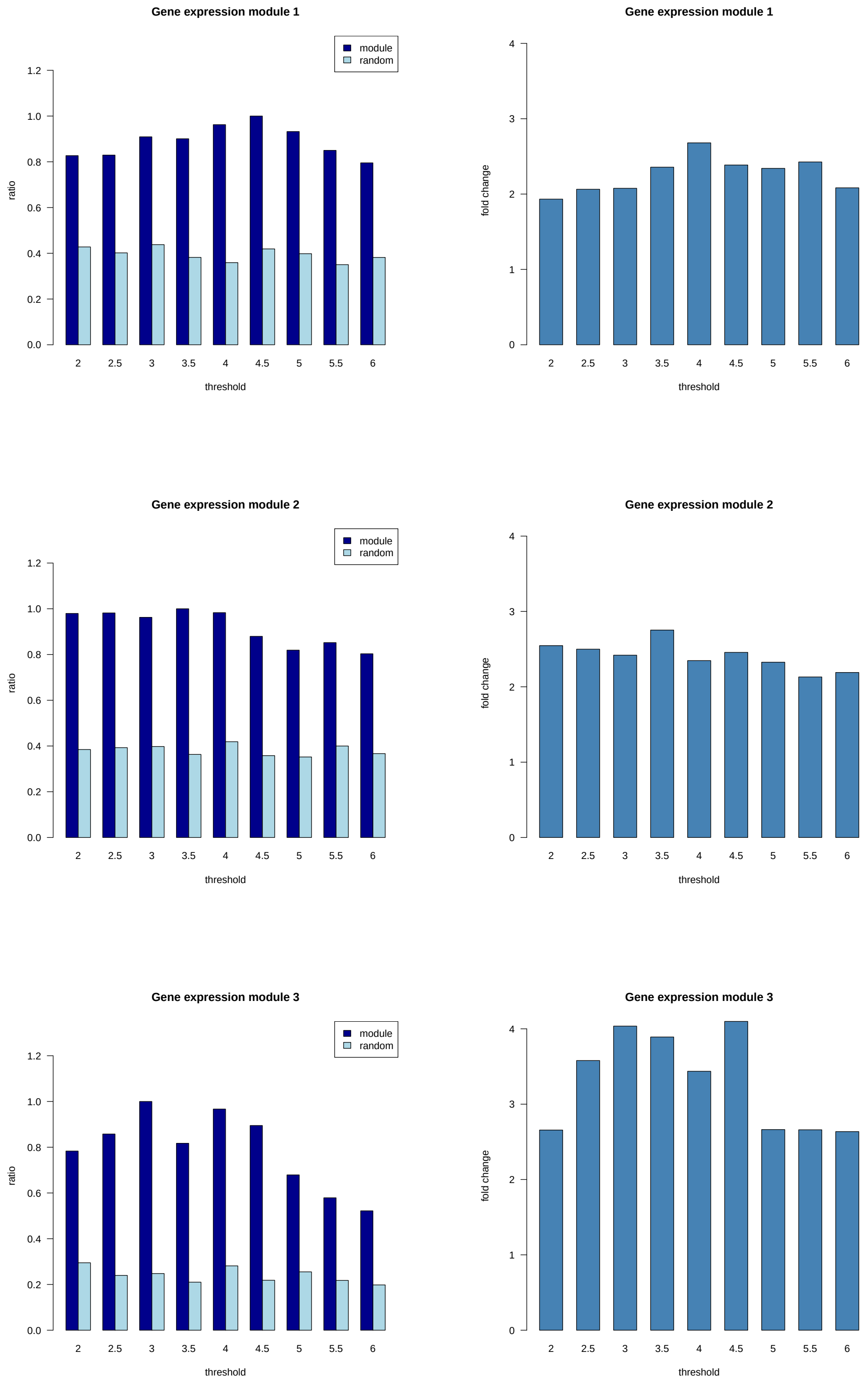

Figure C.1.: Enrichment ratios and fold-change of enrichment ratios of modules 1 to 3

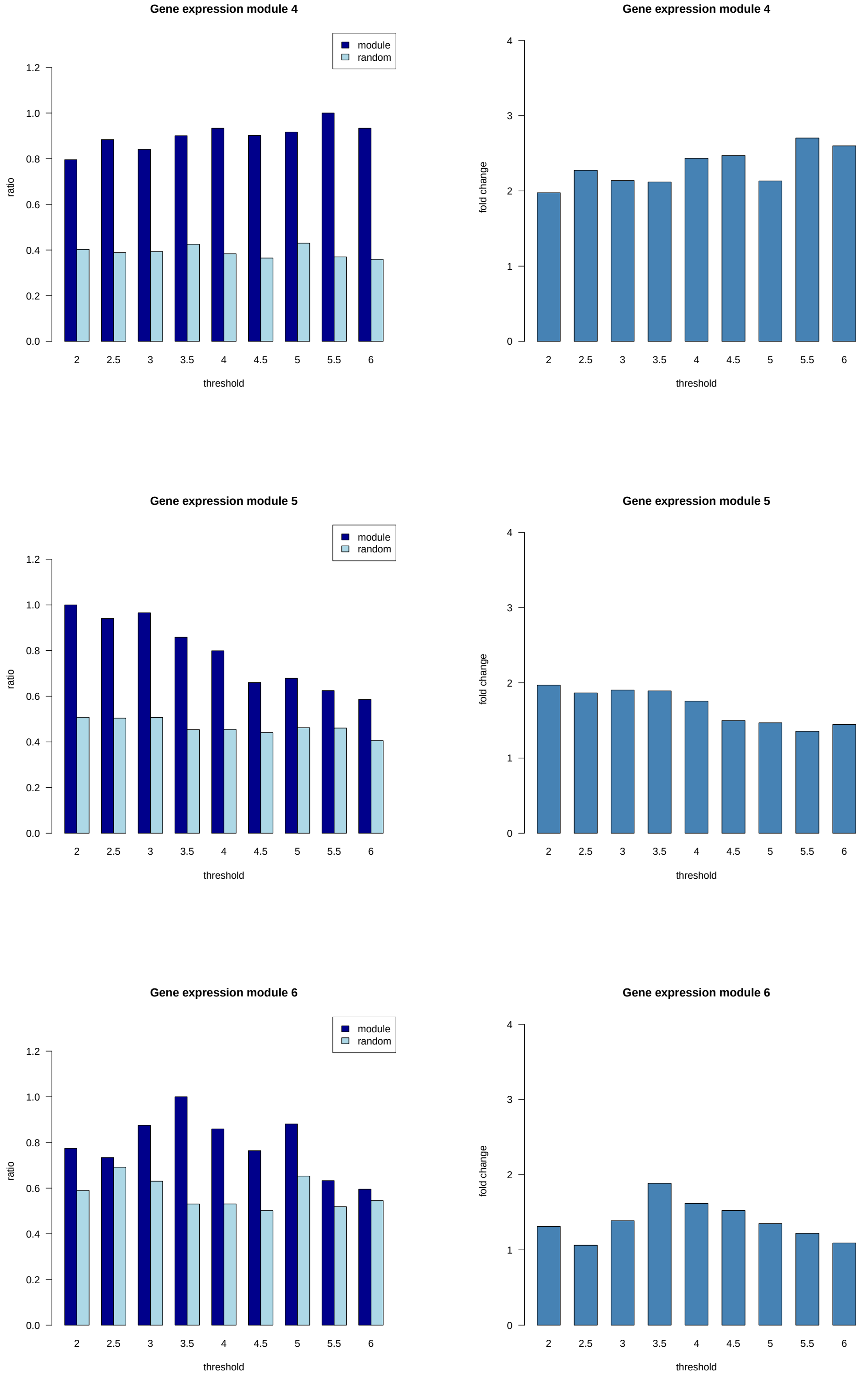

Figure C.2.: Enrichment ratios and fold-change of enrichment ratios of modules 4 to 6

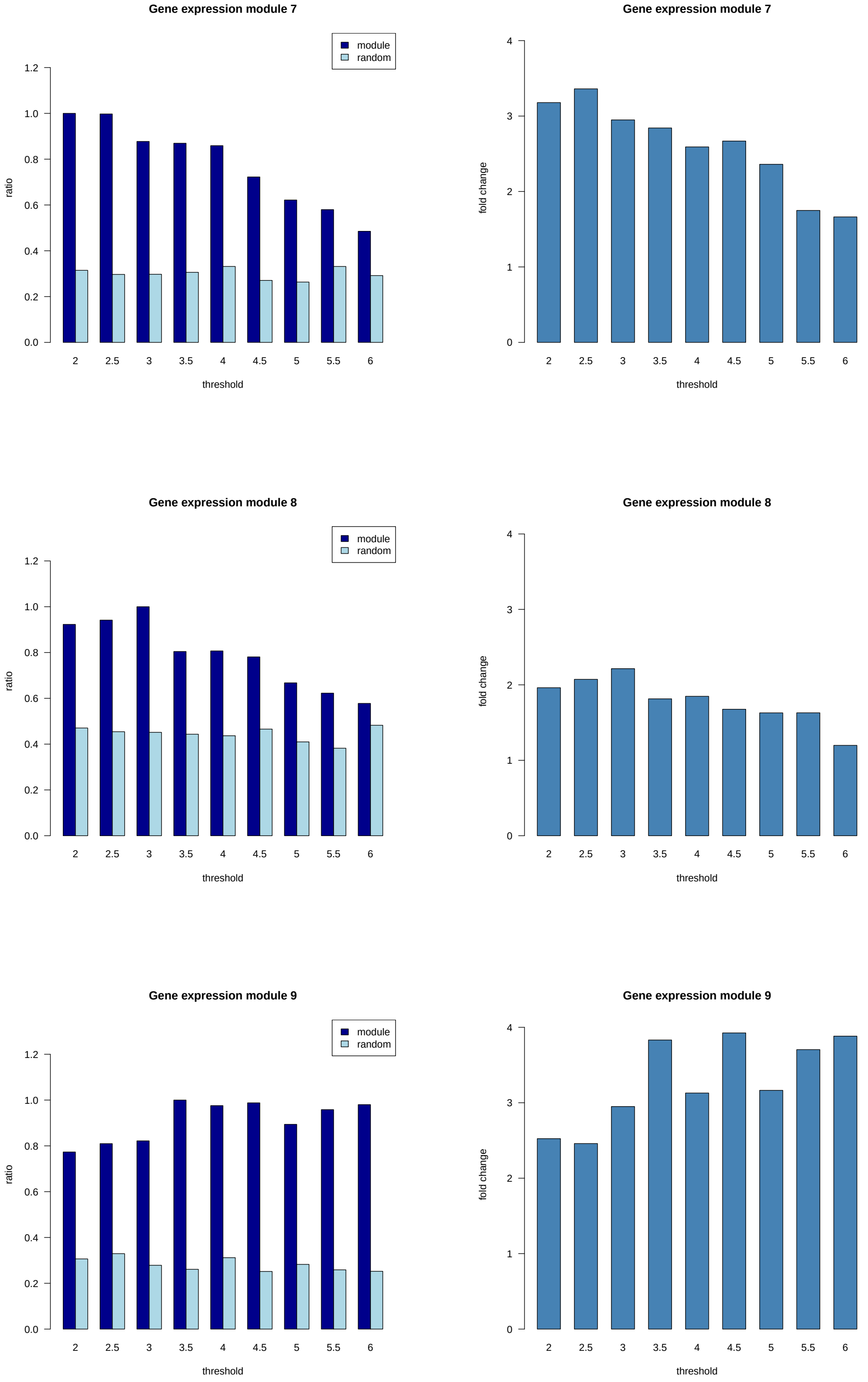

Figure C.3.: Enrichment ratios and fold-change of enrichment ratios of modules 7 to 9

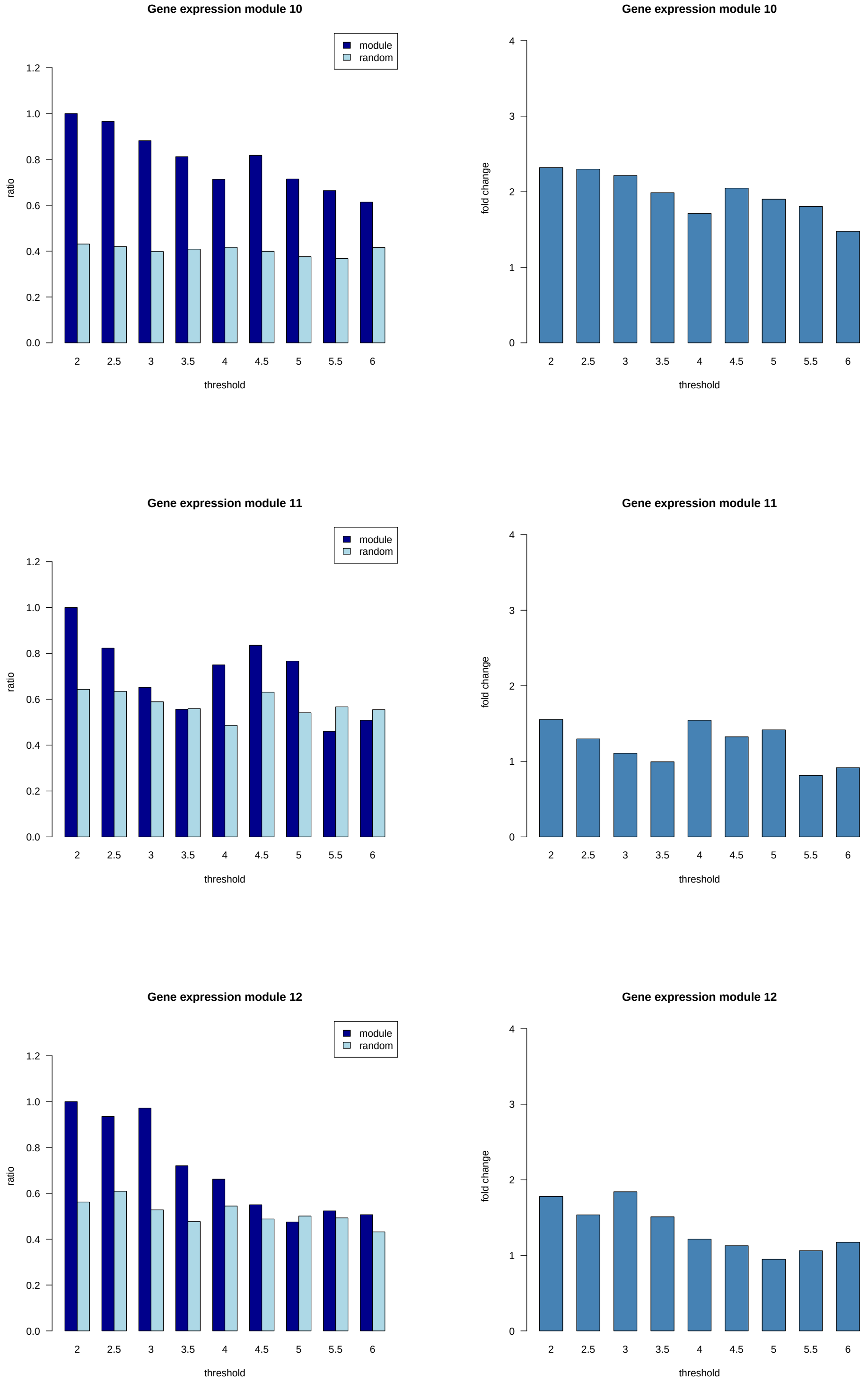

Figure C.4.: Enrichment ratios and fold-change of enrichment ratios of modules 10 to 12

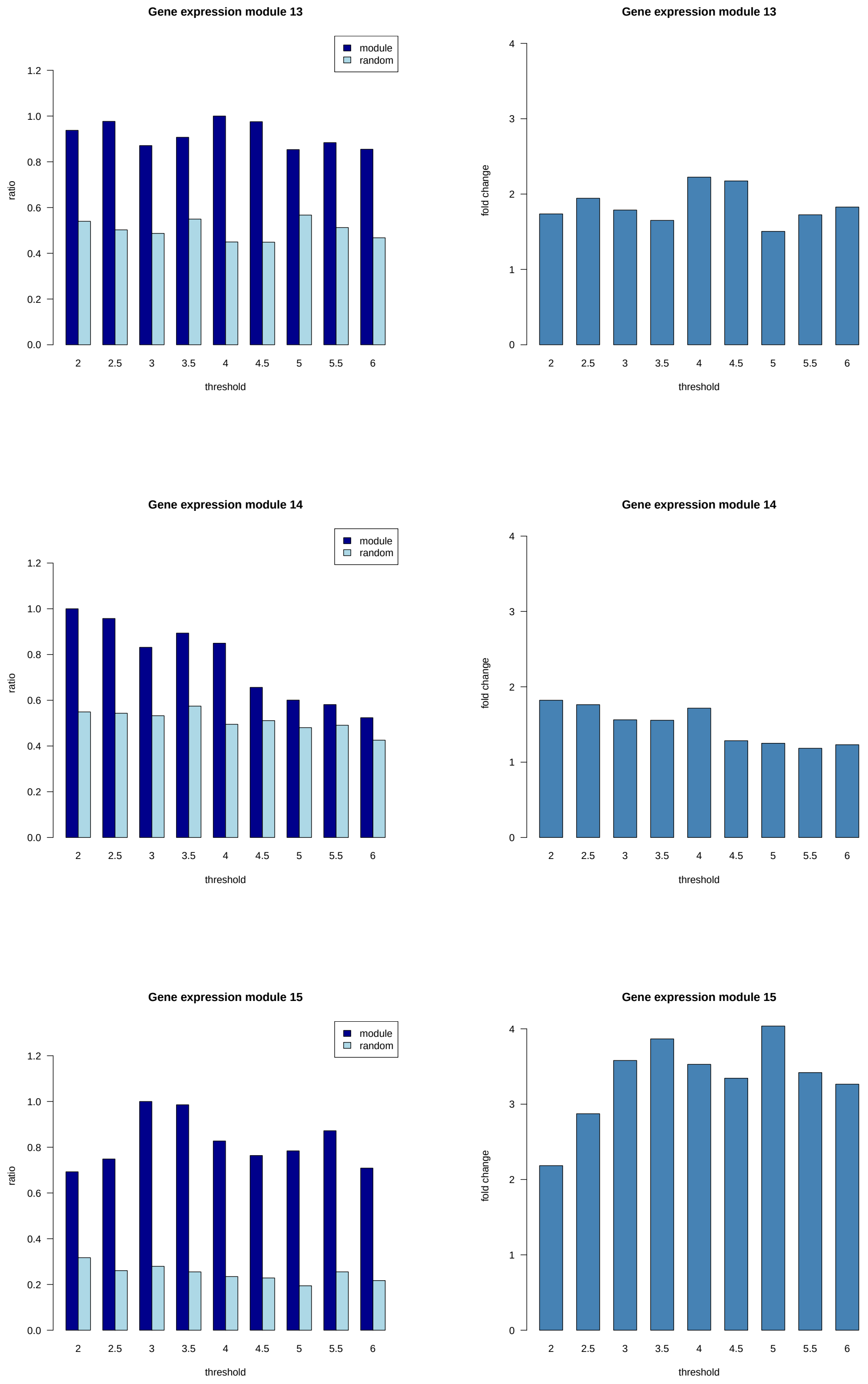

Figure C.5.: Enrichment ratios and fold-change of enrichment ratios of modules 13 to 15

List of Figures

List of Tables