Selecting Biomarkers for Pluripotency and Alzheimer's Disease

The Real Strength of the GA/SVM

Vom Fachbereich Mathematik/Informatik der Universität Osnabrück zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften

Dissertation

vorgelegt von

Diplom-Bioinformatikerin
Lena Scheubert
geboren am 2.1.1983 in Gera

Tag der mündlichen Prüfung: 27. Juni 2012

Abstract

Pluripotency and Alzheimer's disease are two very different biological states. Even so, they are similar in the lack of knowledge about their underlying molecular mechanisms. Identifying important genes well suited as biomarkers for these two states improves our understanding. We use different feature selection methods for the identification of important genes usable as potential biomarkers.

Beside the identification of biomarkers for these two specific states we are also interested in general algorithms showing good results in biomarker detection. For this reason we compare three feature selection methods with each other. Particularly good results show a rarely noticed wrapper approach of genetic algorithm and support vector machine (GA/SVM). More detailed investigations of the results show the strength of the small gene sets selected by our GA/SVM.

In our work we identify a number of promising biomarker candidates for pluripotency as well as for Alzheimer's disease. We also show that the GA/SVM is well suited for feature selection even if its potential is not yet exhausted.

Acknowledgements

Without the support of several people writing this thesis would not have been possible. First of all I want to thank my advisor Volker Sperschneider who strengthened me in many tough decisions and always helped me with his experience and a new perspective. Without his support writing this thesis would not have been possible.

I also want to thank Georg Füllen who guided me through multiple challenging issues of this thesis and always provided me with a large number of interesting studies and recent publications.

I am indebted my coauthors Dirk Repsilber, Mitja Luštrek and Rainer Schmidt for the productive collaboration. I like to express special thanks to Dirk for many valuable discussions and for ensuring we always used the right statistical tests.

I want to thank Alexander Fanghänel and Kerstin Scheubert for proofreading this thesis and discussing numerous issues from a point of view outside my specific field of research.

Finally, I want to thank my parents who were always just a phone call away. Without them I would never have reached this point in my life.

Contents

1 Introduction **1**

2 Basic methods **7**
2.1 Statistical methods 7
2.1.1 Fold change 7
2.1.2 T-test 8
2.1.3 χ^2-test 9
2.1.4 False discovery rate correction 11
2.1.5 Shannon entropy and mutual information 12
2.2 Special methods on microarrays 13
2.2.1 Robust multichip average 13
2.2.2 Gene set enrichment analysis 14
2.3 Machine learning and feature selection methods 15
2.3.1 Basic definitions 16
2.3.2 The C4.5 classifier 18
2.3.3 The k-nearest neighbor classifier 19
2.3.4 The naive Bayes classifier 20
2.3.5 Support vector machines for classification 21
2.3.6 Random forest for classification and feature selection . 22
2.3.7 Information gain for feature selection 23
2.3.8 GA/SVM for feature selection 24
2.4 Software packages 29

3 Microarray data and data preprocessing **31**
3.1 DNA microarrays 31
3.1.1 Architecture of DNA microarrays 32
3.1.2 Steps in a microarray experiment 33
3.1.3 Preprocessing of microarray data 34
3.2 Data acquisition and preprocessing 35
3.2.1 Data acquisition 35
3.2.2 From raw data to preprocessed data sets 38

4 Quality of feature selection methods **43**
4.1 Methods for microarray analysis 43
4.1.1 Statistical methods 44
4.1.2 Machine learning methods 45
4.1.3 Feature selection methods – Identification of potential biomarkers 47
4.2 Quality of different classification and feature selection methods. 52
4.2.1 Classification performance of different classifiers 53
4.2.2 Classification performance of selected features 55
4.3 The true potential of our GA/SVM 61
4.3.1 Classification capability of small biomarker sets 62
4.3.2 Analysis of gene pairs in small sets 65
4.4 Comparison of our data sets 74

5 Promising biomarker candidates **77**
5.1 Pluripotency 78
5.1.1 Embryonic and adult stem cells 78
5.1.2 Molecular mechanisms of pluripotency 80
5.1.3 iPS cells - first successes in ESC research 83
5.1.4 Stem cell therapies - A glace to the future of medicine 84
5.2 Identified biomarkers for pluripotency 85
5.2.1 Gene set enrichment analysis 86
5.2.2 Biological relevance of the selected biomarkers 87
5.3 Alzheimer's disease 93
5.3.1 Molecular mechanisms of Alzheimer 93
5.3.2 Diagnosis and treatment 97
5.3.3 Perspective 98
5.4 Identified biomarkers for Alzheimer's disease 99
5.4.1 Gene set enrichment analysis 100
5.4.2 Biological relevance of the selected biomarkers 102
5.5 Focusing on our data sets 106
5.5.1 Problems in gene expression data analysis 106
5.5.2 Comparison of our data sets 109

6 Conclusion **111**

A Appendix **115**
A.1 All GEO data series forming the PLURI data set 115
A.2 Partitioning of the PLURI data set 118
A.3 Partitioning of the AD data set 121

Bibliography **123**

Chapter 1

Introduction

If we look into somebodies face we only need a split of a second to decide in which mood the person is. Drawing on our experiences we make a decision intuitively and we are not always able to specify which characteristic of the face is determining our perception.

During the last decade algorithms also became very strong in solving this problem [1]. Based on a number of images of happy and sad faces machines are able to learn a rule for distinguishing the two classes of people and this way classify new images. Even so, in the field of image analysis they are still outperformed by humans.

When leaving the visual field and pass over to more abstract numerical problems machine learning algorithms outclass humans by far. Instead of a face we now look onto a vector of thousands of real number values and have to decide if this vector belongs to one class or another. Imagine that, we are not able to make an intuitive decision any longer. Usually, those vectors contain values that are less important than other or even have no influence on the classification of the vector. For example, the color of the eyes it is not relevant for the facial expression of a person but we perceive it anyway. Eliminating irrelevant values and reduce the size of the vector to the most relevant characteristic simplifies the problem and supports the intuitive understanding.

Biomarkers

In medicine and biology distinguishing tissues of different biological states is a common task, for example the identification of tumor cells compared to healthy ones [2]. In medical practice a classification of biological tissues allows reliable diagnosis of various diseases. In cell biology we mainly focus on the detection of cells satisfying a specific function or structure. Usually

biological tissues are defined by a huge number of characteristics but only few of the characteristics are needed associating the tissue with a specific biological state. Those characteristics can serve as biological markers, so called biomarkers.

Formal, a biomarker is defined as an objective measurable characteristic used as an indicator for a biological state [3].

What kind of characteristics we use as biomarkers mainly depends on the question we want to answer. In this work, we focus on the underlying molecular mechanisms of the investigated biological states and the question which genes are involved in the regulatory processes. For this reason we use gene expression levels as characteristics for the investigated biological state and identify genes that are well suited for distinguishing the biological state from others as biomarkers.

For the measurement of the gene expression levels in a biological tissue we use DNA microarrays.

Microarrays

DNA microarrays allow the measurement of thousands of gene expression levels simultaneously using only small amount of biological material [4]. Microarrays are very popular for biological and medical use. Besides the fact that microarrays got less expensive during the last years, most scientific journals require a publication of the experimental data. For this reason various databases exist containing large series of freely available microarray data.

Because of the large number of measured gene expression levels microarrays give a widespread insight into cellular processes. Nevertheless, as we are usually interested in a specific biological state and in processes that are important for maintaining this state most of the measured genes are irrelevant. Identifying the few important genes usable as biomarkers is a challenging task in microarray data analysis.

A large number of methods starting with simple statistical techniques [5] up to complex machine learning models are used for the identification of biomarkers from microarray data. Among them feature selection methods [6] show good results identifying genes with a high classification capability as potential biomarkers.

Feature selection

Feature selection methods are not specifically developed for the identification of biomarkers. Even so, they show good results on large-scale microarray data sets.

Usually a microarray data set consists of multiple arrays, so called samples, each containing the gene expression levels of thousands of genes. Each sample is clearly associated with a biological state. For selecting genes that are important for distinguishing between different biological states samples of at least two different states are required in the data set.

Feature selection methods reduce the dimension of the data set by removing irrelevant or redundant genes. Using feature selection techniques we are able to select small subsets of relevant features well suited to build robust learning models. The elimination of redundant features improves the generalization capability and the speed of supervised and unsupervised machine learning algorithms that are usually not designed for dealing with a large number of unimportant features.

There are a large number of widely used feature selection algorithms that allows us to include large amount of information into microarray data analysis. Feature selection methods show good results in detecting genes that serve as indicator for a specific biological state. A further investigation of those genes may give insights into complex regulatory processes. We consider those genes to be potential biomarkers. In our work we compare three different feature selection methods regarding their ability to identify potential biomarkers for two very different biological states.

Pluripotency and Alzheimer's disease

In this work, we focus on the identification of new biomarkers for two different biological states.

First we are interested in pluripotent cells. Pluripotency is the potential of a cell to differentiate into any of the three germ layers endoderm, mesoderm or ectoderm. For researchers this ability of embryonic stem cells is very important because of the potential benefit of stem cell therapies [7, 8]. They assume that the use of embryonic stem cells enables the treatment of multiple at this time incurable diseases.

As a second biological issue we are interested in Alzheimer's disease affected brain cells. Alzheimer causes high costs in the health care system and recent studies show that these costs will increase year after year [9, 10]. The disease leads to an irreversible damage of the neural cells in the brain and a loss of memory and other important brain functions. An early detection of the disease is the best chance to enable a successful treatment.

In this work, we identify potential biomarkers for pluripotency and Alzheimer. We assume that these biomarkers may improve the overall knowledge and give insights into the underlying molecular processes of the two biological states. As the processes in pluripotent and Alzheimer's disease affected tissue

differ a lot, we are able to analyze the quality of our feature selection methods on two highly diversified tasks.

Outline of this thesis

In this thesis we focus on the quality of different feature selection methods for the identification of potential biomarkers form microarray data. Besides the performance of different algorithms the second issue is the actual identification of biomarkers for pluripotency and Alzheimer's disease.

Part of the results presented in this thesis are already published [11], with valuable contributions of Rainer Schmidt, Dirk Repsilber, Mitja Luštrek and Georg Fuellen. As we did not consider the partial dependencies between the samples of the pluripotency data set in this paper, it reports slightly elevated accuracies due to overoptimistic cross-validation. To obtain the results presented in this thesis, we corrected the data preprocessing and redid the cross-validation calculations. At the time of writing, a second paper, with the same set of authors, reporting results that are also presented in this thesis (with a focus on the Alzheimer data set), is under review by BMC Bioinformatics.

In Chapter 2, we start with a detailed introduction to all standard methods used in the work. Then, in Chapter 3 we focus on the technical and methodical background of DNA microarrays as well as on the microarray data sets we use for our analyses. As we do not perform our own microarray experiments we use freely available data from Gene Expression Omnibus. As we are interested in biomarkers for pluripotency and for Alzheimer's disease we compose two data sets for our analyses. The first data set referred to as PLURI data set contains samples of pluripotent and non-pluripotent cells. The second data set referred to as AD data set contains samples of Alzheimer's disease affected brain tissue as well as of brain tissue samples derived from a healthy control group. For both data sets we use raw microarray data and perform various preprocessing steps such as background correction, normalization and summarization. Additionally, we use a number of statistical methods to reduce the number of genes in each data set before applying feature selection methods for the identification of potential biomarkers.

We split our work into two main parts. In Chapter 4 we identify methods well suited for biomarker selection in microarray data. For this we compare the quality of three feature selection algorithms regarding the classification capability of the selected genes. Here, our wrapper of genetic algorithm and support vector machine (GA/SVM) outperforms information gain as well as random forest. We assume that the main reason for the good performance

of our GA/SVM is the dependencies of the genes selected together in a small set. Concluding that the small gene sets are the real strength of the GA/SVM we following investigate the dependencies within the small sets in-depth. To the best of our knowledge this is a new approach in biomarker selection and offers promising opportunities.

In Chapter 5 we focus on the identification of potential biomarkers for pluripotency and Alzheimer's disease. As the analysis of the quality of all three feature selection methods show good results we assume to identify genes well suited as biomarkers. We discuss our results in context to other recent publications finding many similarities. Additional we identify a number of new biomarker candidates not yet associated with pluripotency or Alzheimer.

This work shows the strength of feature selection methods in microarray data analysis focusing on a rarely used wrapper approach of genetic algorithm and support vector machine. Using those algorithms we are able to identify promising candidate genes for further biological examination.

Chapter 2

Basic methods

In this chapter we give an introduction to all standard methods we use in this thesis. As we use statistical and machine learning methods for the analysis of microarray data, our description differs slightly from the definitions given in the most general sense and refer to microarray data analysis. We classify the methods into three groups: Statistical methods, methods specially designed for microarray analysis and machine learning methods. Additionally we give a brief introduction to the software packages we use in our work.

2.1 Statistical methods

The following statistical methods are commonly known and among others widely used in microarray data analysis [12, 13]. They are implemented in most of the software suites available for statistical data analysis such as R [14], MATLAB [15] or Mathematica [16].

2.1.1 Fold change

The fold change is a measure for the quantitative difference of two values. In microarray data analysis, it is widely used for the identification of differentially expressed genes [17–19].

In literature, we find two different definitions for the fold change of a gene that is expressed under two different conditions. Based on the scale of the gene expression data the fold change is either given by the ratio [20] or the difference [18, 19, 21] of the mean expression values in the two classes of samples. As we work on log scaled microarray data we focus on the second opportunity.

Let $X_1 = \{x_1, \ldots, x_k\}$ be a set of microarray samples belonging to a

certain class and $X_2 = \{x_{k+1}, \ldots, x_n\}$ a set of samples belonging to another class. For example, we could interpret this as classifying the microarrays into healthy and diseased samples. Each sample x is a vector containing the log-transformed gene expression values of all examined genes on the array. The fold change FC_i for each gene i can be computed as

$$FC_i = mean_i(X_1) - mean_i(X_2).$$

Here, $mean_i(X)$ denotes the arithmetic mean of the gene expression values of gene i over all samples of set X.

In this thesis we use the fold change as a part of the filtering process applied to the microarray data before feature selection. The filtering of the microarray data is described in Section 3.2.2

2.1.2 T-test

The t-test is a statistical test that specifies how likely a result is due to chance [22–24]. It is used to determine whether there is a statistical significant difference between two groups of samples or not. In microarray analysis, we use the t-test to identify genes differentially expressed under two different conditions.

Let $X_1 = \{x_1, \ldots, x_k\}$ be a set of microarray samples belonging to a certain class and $X_2 = \{x_{k+1}, \ldots, x_n\}$ a set of samples belonging to another class. Each sample x_j is a vector containing the log-transformed gene expression values of all examined genes of a microarray. The test statistic T_i for the t-test for gene i is designed as

$$T_i = \frac{|FC_i|}{\sqrt{\frac{var_i(X_1)}{k} + \frac{var_i(X_2)}{n-k}}},$$

where FC_i is the fold change of gene i, described in Section 2.1.1, and $var_i(X)$ denotes the estimated variance of the expression values of i in set X. The t-test combines the fold change as measurement for the difference between two classes with the variances inside each class.

The Student's t-distribution illustrated in Figure 2.1 is defined as

$$f_{df}(x) = \frac{\Gamma\left(\frac{1+df}{2}\right)}{\Gamma\left(\frac{df}{2}\right) \cdot \sqrt{df \cdot \pi}} \cdot \left(1 + \frac{x^2}{df}\right)^{\frac{-df+1}{2}},$$

where Γ is the Gamma function [25] and df is the number of degrees of freedom. The test statistic T_i follows the t-distribution with $df = n - 2$

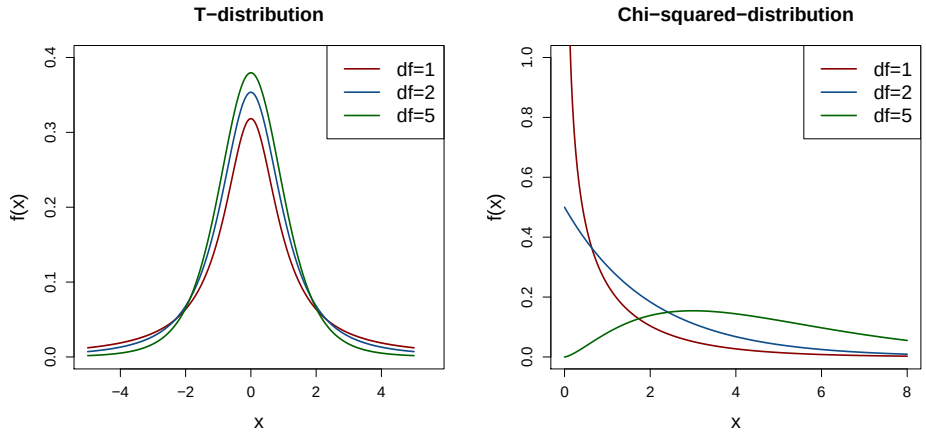


Figure 2.1: T- and χ^2-distribution with 1, 2 and 5 degrees of freedom (df).

degrees of freedom under the assumption that there is no difference between the samples of X_1 and X_2.

The significance level, the so called p-value, of the two-tailed t-test is given by

$$pValue(T_i) \quad = \quad \int_{-T_i}^{-\infty} f_{df}(x)dx + \int_{T_i}^{\infty} f_{df}(x)dx \quad = \quad 2 \cdot \int_{T_i}^{\infty} f_{df}(x)dx.$$

This term determines the sum of two areas below the function $f_{df}(x)$, bounded by our test statistic T_i respectively by the negative test statistic $-T_i$. As the Student's t-distribution is symmetric we can summarize the term. We call a difference between the two classes significant if the p-value is smaller than 0.05.

For the calculation of the p-values we use the t-test with unequal variances [26] of the R package 'stat' [27].

Similar to the fold change, the t-test is used as part of the data filtering described in Section 3.2.2.

2.1.3 χ^2-test

Pearson's χ^2-test [22, 23] is a statistical test for the analysis of distributions. On the one hand we can determine whether an observed distribution differs from a theoretical one. On the other hand it is used to test the independence of two observed variables. In this work we use the χ^2-test for the test of independence.

More specifically, we test the independence of two genes G_1 and G_2 occurring together in small gene sets selected by our wrapper of genetic algorithm and support vector machine (GA/SVM, see Section 2.3.8). For each gene we define two possible observations:

1. The gene occurs in the small set. (obs_1)
2. The gene does not occur in the small set. (obs_2)

The observations for G_1 and G_2 in an experiment with n small sets can be expressed in the following contingency table:

		G_2		
		obs_1	obs_2	
G_1	obs_1	a_{11}	a_{12}	$a_{1\cdot}$
	obs_2	a_{21}	a_{22}	$a_{2\cdot}$
		$a_{\cdot 1}$	$a_{\cdot 2}$	n

The values a_{ij} determine the number of times we observe obs_i for gene G_1 and obs_j for gene G_2, for example a_{11} defines the number of small sets, in which gene G_1 as well as gene G_2 occur. Consequently, the number of times we observe obs_i for gene G_1 is given by $a_{i\cdot} = a_{i1} + a_{i2}$. Analogous, for gene G_2 we calculate the number of times we observe obs_i as $a_{\cdot i} = a_{1i} + a_{2i}$.

For this contingency table the χ^2-test statistic X^2 can be calculated as

$$X^2 = \sum_{i=1,2} \sum_{j=1,2} \frac{(a_{ij} - a'_{ij})^2}{a'_{ij}},$$

where a'_{ij} is the expected value of a_{ij} under the assumption that G_1 and G_2 are independent. The expected value a'_{ij} can be estimated as

$$a'_{ij} = \frac{a_{i\cdot} a_{\cdot j}}{n}.$$

For the usability of the χ^2-test we need a sufficiently large number of samples ($n \geq 60$ [22]) and the expected value for a_{ij} should not come below five [23].

Figure 2.1 shows the χ^2-distribution defined as

$$f_{df}(x) = \frac{1}{2^{\frac{df}{2}} \cdot \Gamma\left(\frac{df}{2}\right)} \cdot x^{\frac{df}{2}-1} \cdot e^{\frac{-x}{2}},$$

where Γ is the Gamma function [25] and df is the number of degrees of freedom. Under the assumption that G_1 and G_2 are independent the distribution of our test statistic X^2 can be approximated by the χ^2-distribution with $df = 1$.

If the p-value of the one-tailed χ^2-test is smaller than 0.05 we assume G_1 and G_2 as independent. The p-value is given as

$$pValue(X^2) = \int_{X^2}^{\infty} f_{df}(x)dx$$

that defines the area below the function $f_{df}(x)$ bounded by the value for the test statistic X^2 on the left side.

For the calculation of the p-values we use the χ^2-test of the R package 'stat' [27].

In this thesis we use the χ^2-test for the identification of significant over- and under-represented gene pairs in the small sets of genes selected by our GA/SVM in Section 4.3.

2.1.4 False discovery rate correction

False discovery rate [28] correction is a statistical method used to adjust the expected proportion of incorrectly rejected null hypotheses, so called type I errors, in multiple hypothesis testing.

The following table shows the possible outcome of m hypotheses tests with m_0 true null hypotheses h_0.

	reject h_0	accept h_0	
h_0 is true	FP	$m_0 - FP$	m_0
$\overline{h_0}$ is true	TP	$m - m_0 - TP$	$m - m_0$
	R	$m - R$	m

The observable random variable R determines the total number of tests for which we reject h_0. The number of true positives TP is the number of true alternative hypotheses $\overline{h_0}$ for which we reject h_0 and FP (false positives) is the number of true h_0 for which we reject h_0, the so called type I error.

As TP and FP are unobservable variables we estimate the false discovery rate FDR as the expectation value

$$FDR = \mathrm{E}\left(\frac{FP}{FP + TP}\right) = \mathrm{E}\left(\frac{FP}{R}\right).$$

For $R = 0$, $\frac{FP}{R}$ is defined as 0.

We use the false discovery rate correction for measuring the overall accuracy of a set of significant genes. The q-value [29] gives a measure for the significance attached to each gene. It is the minimal rate at which a test can still be considered significant. Therefore it is the p-value equivalent for the false discovery rate.

We use the R-package 'qvalue' [30] for our data analysis.

In our thesis we use the t-test as well as the χ^2-test for testing multiple hypotheses. FDR is used to correct the resulting p-values of these tests. A detailed description of the implementation is given in Section 3.2.2 and Section 4.3.

2.1.5 Shannon entropy and mutual information

The Shannon [31] entropy $H(X)$ is a measurement for the uncertainty about a random variable X with the possible outcomes $\{x_1, \ldots, x_n\}$. It quantifies the expected information contained in x_i, a specific realization of X. The entropy is defined as

$$H(X) = -\sum_{i=1}^{n} \Pr(x_i) \log_2 \Pr(x_i),$$

where $\Pr(x_i)$ determines the probability that $X = x_i$. The Shannon entropy is often used within decision tree algorithms.

Additional to the entropy the mutual information of two variables depends on the conditional entropy $H(X|Y)$. The conditional entropy determines the average uncertainty about a random variable X after observing another random variable $Y = \{y_1, \ldots, y_m\}$. It is defined as

$$H(X|Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} \Pr(x_i, y_j) log_2 \frac{\Pr(y_j)}{p(x_i, y_j)},$$

where $\Pr(x_i, y_j)$ denotes the probability that $X = x_i$ and $Y = y_j$.

Using the terms of entropy and conditional entropy we define the mutual information $I(X, Y)$ [32, 33] of two random variables X and Y as

$$I(X, Y) = H(X) - H(X|Y).$$

The mutual information defines the mutual dependency of two random variables. It is a measurement for the reduction of uncertainty about a random variable X after observing Y. Among other, mutual information is used to characterize the redundancy of two variables [34].

In Section 4.2.2 we use the mutual information of two genes as a measurement for their redundancy. For estimating the mutual information we use the R package 'parmigene' [35] based on k-nearest neighbor distances [36].

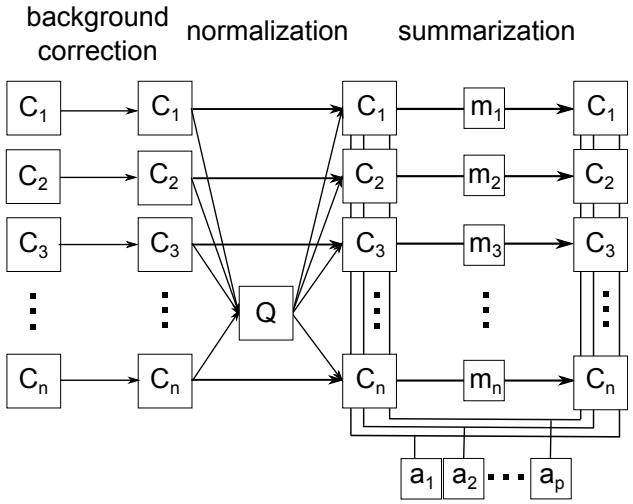


Figure 2.2: Model of the robust multichip average (RMA) algorithm. Starting with the raw intensities of n microarrays we perform background correction, normalization and summarization to obtain the $\log_2$-scaled gene expression values for each array.

2.2 Special methods on microarrays

Besides the standard statistical methods, more complex algorithms have been developed [17] specifically for the analysis of microarray data. These algorithms are often based on simple statistical tests that are adapted to the requirements and questions in microarray data analysis. In the following we introduce two popular methods for the preprocessing and the evaluation of microarray data.

2.2.1 Robust multichip average

Robust multichip average (RMA) [37,38], also called robust multiarray average, is a three-step process for computing the expression measures of multiple microarray chips from the raw fluorescence intensity of the single probes. The algorithm is illustrated in Figure 2.2.

Step 1: Background adjustment. The background correction for each microchip is performed independently. It is based on a convolution model inspired by the probe intensities distribution. The convolution model is defined by

$$O = S + Y,$$

where O is the observed intensity for a probe, S is the true signal and Y is the background noise. The true signal $S \sim \exp(\frac{1}{\alpha})$ is assumed to be exponentially distributed and the background noise $Y \sim \mathcal{N}(\mu, \sigma^2)$ is assumed to be normally distributed. All background corrected values are positive as the normal noise distribution is truncated at zero. After background correction all adjusted probe intensities are in log2-space.

Step 2: Normalization. The normalization is performed using the quantile normalization method [39]. In each microarray the probe with the highest adjusted probe intensities is chosen and averaged over all arrays. This is repeated for the second highest probe intensities, the third highest and so on. Thereby we obtain a set of quantiles, $Q = q_1 \dots q_n$. Then the probe intensities of each array are replaced by the quantiles in a way that the highest quantile replaces the highest probe intensity in each array and so on. This way the distributions of all arrays are equal.

Step 3: Summarization As last step we summarize all probes corresponding to the same gene. The summarization is performed for each of these probe sets separately, using all chips for the calculation. Therefore we apply an additive linear model to the background adjusted, normalized data. The model for a probe value Y_{ij} corresponding to probe i on microchip j can be described as

$$Y_{ij} = a_i + m_j + e_{ij},$$

where a_i denotes the probe affinity effect [38, 40] for probe i and m_j denotes the chip-specific log-scaled expression value for chip j associated with the particular probe set. Median polish [41] is used to estimate m_j. The term e_{ij} determines a random error term.

In this work we use the RMA algorithm implemented in the Affymetrix power tools [42] for the summarization of the used Affymetrix GeneChip arrays.

2.2.2 Gene set enrichment analysis

Gene set enrichment analysis [43] is a general term for computational methods used for analyzing gene sets instead of single genes. As high throughput methods often provide a large number of interesting genes, the manual analysis of the single genes is time consuming and expensive. Therefore the resulting sets of genes are analyzed together.

One method in gene set enrichment analysis is the over-representation analysis, a computational method that analyzes whether a set of genes is

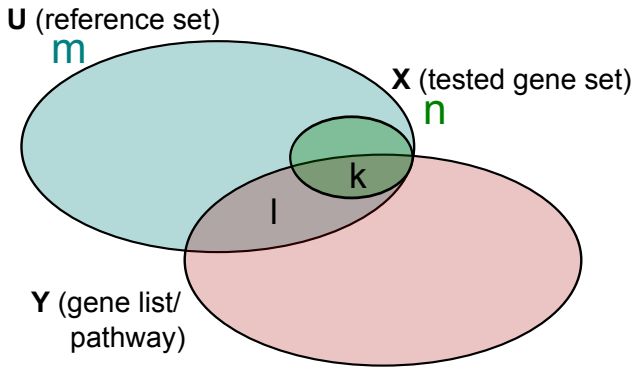


Figure 2.3: Scheme of the three gene sets used for the gene set enrichment analysis (GSEA).

over- or under-represented in another gene set.

Let $X = \{x_1, \ldots, x_n\}$ determine a set of n genes that is tested for over- or under-representation and $U = \{u_1, \ldots, u_m\}$ be a reference set containing m genes. U is a superset of X. We compare X to another gene set Y, for example a set of genes of a specific pathway or regulatory network. From the gene set Y, l genes also belong to the reference set U. The number of genes in the test set X that also belongs to Y is given as k Figure 2.3 shows a Venn diagram of the defined sets.

The probability that a randomly selected gene of the reference set U belong also to Y is l/m. This is the number of genes k' we would expect to find in our test set.

If k, the number of genes actually found, is larger than k', there is enrichment of genes belonging to the gene set Y, otherwise there is a depletion. We then estimate the statistical significance of the enrichment by computing the one-tailed p-value using the hypergeometric distribution as

$$pValue = \begin{cases} \sum_{i=k}^{n} \frac{\binom{l}{i}\binom{m-l}{n-i}}{\binom{m}{n}}, & \text{if } k' < k \\ \sum_{i=0}^{k} \frac{\binom{l}{i}\binom{m-l}{n-i}}{\binom{m}{n}}, & \text{if } k' \geq k. \end{cases}$$

2.3 Machine learning and feature selection methods

In this thesis we use a variety of different machine learning and feature selection methods. For identifying potential biomarkers we apply the three feature selection methods information gain, random forest and a wrapper of

genetic algorithm and support vector machine (GA/SVM) to our data sets. The quality of these algorithms is verified using several classification methods. Additionally, we compare the performance of different classifiers on our two data sets.

After defining some general terms this section gives a deeper insight into the algorithms used in this work. For most of these standard methods we use implementations freely available in the Weka Software Suite [44]. A special role among our methods takes the GA/SVM described in Section 2.3.8. Because it is not implemented in the most popular freely available software packages we implement the algorithm by our own and put it to good use. As the GA/SVM is the most important algorithm in our work we describe it more detailed than the other methods.

2.3.1 Basic definitions

In the following we determine a number of terms useful for understanding the formal description of the classification and feature selection methods in this work. In this thesis we use those methods for the analysis of microarray data and the identification of promising biomarker candidates. For this reason some of the general terms defined below are described more specifically and more intuitively with respect to microarray data and our work in the last part of this section.

Classifier. A classifier is an algorithm or mathematical function that is used for mapping a category to a piece of input data based on its characteristics. We determine a classifier as a function $g(x) : \mathbb{R}^p \to C$ that assigns a category (class) $c_j \in C$ to each sample d_i in a given distribution $D \subseteq \mathbb{R}^p$. In general we split a classification process into the following three steps:

1. **Training.** For training we use a set of n samples $X = \{x_1, \ldots, x_n\}$, called training set. The training set is a subset of the given distribution $X \subset D$. Each sample x_i of the training set consists of p characteristics (features) $x_i = \{f_1, \ldots, f_p\}$ and is explicitly assigned to a class $c_j \in C$. The actual training process depends on the specific classifier. After training we are able to classify new samples using the trained classifier $g(x)$.

2. **Testing.** For determining the quality of a trained classifier we use a set of samples $Y = \{y_1, \ldots, y_m\}$, called test set. The training set X and the test set Y are disjoint. Each sample y_i is an element of the distribution D and the class c_j it belongs to is known. For testing we

use the trained classifier $g(x)$ to assign a class to each sample y_i. The assigned class label is compared to the class y_i is known to belong to.

$$s_i = \begin{cases} 1 & g(y_i) = c_j, \\ 0 & \text{otherwise} \end{cases}$$

As measure for the quality of the classifier we use the classification accuracy that is defined as

$$acc(g) = \frac{\sum_{i=1}^{m} s_i}{m}.$$

3. **Classification of new samples.** In the actual classification step we use the trained classifier $g(x)$ for applying a class c_j to samples $d_i \in D$ not yet associated with a class. We consider the classification of the sample the more trustworthy the higher the accuracy of the trained classifier is.

Cross-validation. Mostly, we have no explicitly given training and test sets. To still be able to evaluate the quality of the trained classifier we dismiss a part of the training set from the training procedure and use it as test set. Usually, this is repeated many times using different parts of the given set for training. A systematic procedure for calculating the mean classification accuracy on a data set is cross-validation. The k-fold cross-validation is following described.

We split the given data set X into k disjoint subsets $X = X_1 \dot\cup \ldots \dot\cup X_k$. The subsets are called folds. Now we train the classifier k-times using $\bigcup_{i \neq k} X_i$ as training and X_k as test set. We calculate the accuracy for each iteration separately and the arithmetic mean of the accuracies over all iterations determines the accuracy of the classifier on the given sample set.

We use a so called stratified cross-validation, that means each fold X_i contains nearly the same number of samples. The most commonly used cross-validation is 10-fold cross-validation [45].

Microarray data in this thesis. Working with microarray data, described in Chapter 3, each sample contains the gene expression values of a single microarray experiment. The features of a sample are the gene expression values of the genes contained in the microarray. In our work the microarrays we analyze belong to one of two classes $C = (c_1, c_2)$. For identifying potential pluripotency biomarkers we work with pluripotent and non-pluripotent samples. For the identification of Alzheimer's disease biomarkers

we work with Alzheimer affected and non-affected samples. In the following we call this samples positive (pluripotent and Alzheimer affected) and negative (non-pluripotent and non-affected) samples. In the following sections describing the principle ideas of the used classifiers and feature selection methods in general. To improve the readability in the result chapters 4 and 5 we than switch to the term of gene instead of feature.

2.3.2 The C4.5 classifier

C4.5 was developed in 1993 by Ross Quinlan [46]. The algorithm builds classifiers in the form of decision trees.

Starting at the root of the tree the training set X is divided into k disjoint subsets $X_1 \dot{\cup} X_2 \dot{\cup} \ldots \dot{\cup} X_k$. Each subset is applied to a new child node. For splitting X we chose feature f with the largest normalized information gain [46] due to classification.

In each node of the tree we repeat the splitting using the associated subset until all samples of the subset belong to the same class. Algorithm 1 illustrates the C4.5 algorithm for growing a decision tree.

Algorithm 1 The C4.5 algorithm

```
1: create root node n_0 containing training set X
2: splitSamples(n_0)
3: procedure splitSamples(n)
4:     if n contains only samples of a class c then
5:         create leaf labeled with c
6:         return
7:     else
8:         for all features i do
9:             calculate the normalized information gain I(f_i)
10:        end for
11:        choose feature i with max_i I(f_i)
12:        split X into k subsets X_1, ..., X_k depending on feature i
13:        for all subsets X_i do
14:            create a new child node n_i and apply X_i to n_i
15:            splitSamples(n_i)
16:        end for
17:    end if
18: end procedure
```

Basically C4.5 is similar to other algorithms used to build decision trees. In contrast to other methods where the trees have to be binary, C4.5 allows

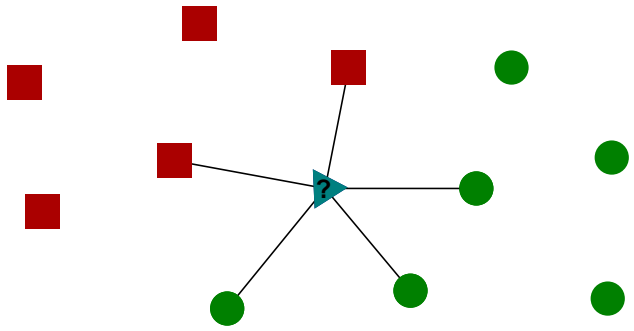


Figure 2.4: Classification of a new sample using the kNN classifier with $k = 5$.

any number of branches on each node. This results in trees having a larger width and a lower depth than the corresponding tress grown with other methods. For this reason the first splits are much more relevant than later ones. As in other decision tree algorithms the initial tree is pruned to avoid overfitting.

2.3.3 The k-nearest neighbor classifier

One of the most intuitive approaches for classification is k-nearest neighbor (kNN) [47]. The idea behind this approach is that two samples more likely belong to the same class the more similar they are.

During the training process, the kNN classifier stores all training samples $X = \{x_1, \ldots, x_n\}$. For the classification of a new sample d we calculate the distance $\delta(x_j, d)$ between a training sample and the new sample for all samples $j = 1, \ldots, n$. Then, the k nearest samples of the training set X are chosen and the majority of the corresponding classes determine the class for d. Figure 2.4 illustrates the classification of a new sample using a kNN classifier.

Although, the principle of kNN is very simple the quality of the algorithm depends on multiple parameters.

The parameter k. The parameter k determines the number of nearest samples used to assign a class label to d_i. On the one hand, for too small values of k, the result is sensible to noise. On the other hand, if k is chosen too large, the influence of neighbored classes distorts the results.

The distance metric. A well chosen distance is determined by the following rule: The smaller the distance between two samples, the greater is the

likelihood that both samples belong to the same class. Although, the Euclidean distance [48] is the most commonly used distance measure for data points, depending on the classification problem other metrics are used es well [48–50].

The combination of the class labels. For determining the class of d using the class labels of the k nearest neighbors the most common approach is the majority voting. For this, the class, most of the neighbors belonging to is assigned to d. Problems occur when the distances between d and its neighbors widely varies and closer neighbors are much more determining than others. Distance-weighted voting methods solve this problem.

2.3.4 The naive Bayes classifier

Based on the Bayes theorem [51] we can express the probability for choosing a specific class $c \in C$ given a sample $d \in D$ as

$$\Pr(c|d) = \frac{\Pr(c) \cdot \Pr(d|c)}{\Pr(d)}.$$

This term consists of the following three parts.

1. $\Pr(c)$ denotes the probability for observing a sample of class c. We estimate $\Pr(c)$ from the training set X.

2. The probability $\Pr(d)$ for observing a specific sample d cannot be estimated. As $\Pr(d)$ is a constant value for all classes c it does not influence the choice of the class and we can rewrite the term above as

$$\Pr(c|d) \propto \Pr(c) \cdot \Pr(d|c).$$

3. The term $\Pr(d|c)$ determines the probability of observing a sample d under the condition of choosing a class c. Under the assumption that the single features of $d = (f_1, \ldots, f_p)$ are independent $\Pr(d|c)$ is defined as

$$\Pr(d|c) = \Pr(f_1, \ldots, f_p|c) = \prod_{k=1}^{p} \Pr(f_k|c),$$

where $\Pr(f_k|c)$ is the probability of observing the feature f_k under the condition of choosing a class c. In general we assume the features in each class as normal distributed, and we estimate the two parameters

μ (mean) and σ^2 (variance) from the training set. Besides the normal distribution other distributions are used as well.

Usually the assumptions we made above are not correct. Even so the model works well on real life problems. We classify a new sample by choosing the class most likely for the given sample d:

$$class(d) = max_c \Pr(c) \prod_{k=1}^{p} \Pr(f_k|c)$$

2.3.5 Support vector machines for classification

A support vector machine (SVM) [52] is a so called large margin classifier as it separates samples of different classes with a large margin between these two classes. Starting with a set of training samples $X \in \mathbb{R}^p$, the SVM calculates a hyperplane separating the samples of two classes. For this the distance of the samples nearest to the hyperplane is maximized. Later, this large margin allows a reliable classification of new samples. For a mathematical exact definition of the hyperplane we use only the nearest samples the so called support vectors. Samples that are further away do not influence the position of the hyperplane. A hyperplane can only separates samples linearly. In real life problems usually the samples are not linearly separable. Therefore the SVM uses a kernel function to separate the classes in a nonlinear manner. Using the kernel function the vector space and therefore the training samples are transferred in a higher dimensional space. If the dimension is high enough the samples become linearly separable. Then the hyperplane is calculated in $\mathbb{R}^m, p < m$. With the re-transformation in the original space the linear hyperplane becomes nonlinear and possibly unconnected. Figure 2.5 displays the principle of the SVM.

Two problems occur when transferring the training samples. First, the transformation in a higher dimensional space is computational expensive and second, the hyperplane in lower dimensional space is very complex.

A well suited kernel function has to describe a hyperplane in high dimensional space that is still easy to define in lower dimensions. Using such a kernel function it is possible to determine the hyperplane in high dimensions without actually applying the transformation to the vector space.

To avoid an overestimation and reduce the number of support vectors the algorithm allows wrong classified samples, so called slack variables, but penalizes them [53].

We use two implementations of SVM. Besides the algorithm implemented in Weka we use the Java implementation LibSVM [54] as part of our wrapper

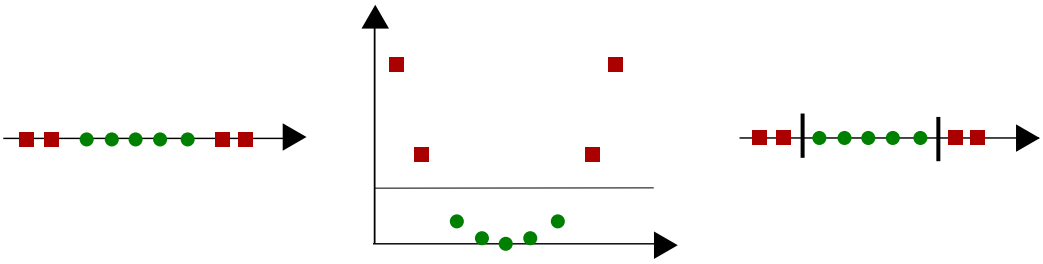


Figure 2.5: Operation breakdown of a support vector machine (SVM). The given samples are non-linearly separable in the 1-dimensional space (left). Using a quadratic kernel function we transfer the data points into a 2-dimensional space. There we separate the two classes using a hyperplane (center). After re-transferring the data to 1-dimensional space the samples are classified (right).

of genetic algorithm and support vector machine.

2.3.6 Random forest for classification and feature selection

First published in 2001 by Leo Breiman [55] the classifier random forest is based on an ensemble of decision trees. Growing the forest of decision trees we start with a training set X of n samples. Each sample contains p features.

Growing a tree in the forest we first select a set X' of n samples out of the training set randomly with replacement. We create a root node that contains X'. Then we choose p' features randomly without replacement, where p' is much smaller than p. Based on these p' features the set X' is split into two subsets X'_1 and X'_2. We create two child nodes containing the two subsets X'_1 and X'_2. Continuing with the selection of $p << p'$ features the process is repeated recursively for all child nodes until all samples of a subset belong to the same class.

This way we grow a fix number of trees that form the random forest. Algorithm 2 displays how the forest is grown.

The training of the random forest classifier is very fast accounted by the short growing times of the single trees. The trees in the forest are usually unpruned. Nevertheless, because of the classification in multiple trees an overfitting is avoided.

As the random forest consists of several decision trees, classification is performed by classifying a sample in each tree separately and determines the final class by majority voting.

Algorithm 2 The random forest algorithm

```
1: for all trees in the forest do
2:     X' ← select n samples of X randomly with replacement
3:     create root node n_0 containing X'
4:     splitSamples(n_0)
5: end for
6: procedure splitSamples(n_0)
7:     if n_0 contains only samples of a class c_j then
8:         create leaf labeled with cj
9:         return
10:    else
11:        select p' features randomly without replacement (p' << p)
12:        split X' into X'_1 and X'_2 based on the selected features
13:        create two nodes n_1 containing X'_1 and n_2 containing X'_2
14:        splitSamples(n_1)
15:        splitSamples(n_2)
16:    end if
17: end procedure
```

Random forest for feature Selection

Because selecting the samples for growing a tree randomly with replacement, some samples are selected more than once and some are never selected. These never selected samples are called out-of-bag instances. To get a measure of the importance of a feature f, the out-of-bag instances are classified and the number of correct classifications is counted as c_{before}. Before classifying the out-of-bag instances again, for the feature f its values in the out-of-bag instances are permuted randomly. The new number of correct classified instances is called c_{after}. The importance of a feature f is given by the difference between c_{before} and c_{after}, averaged over all the trees in the forest. By permuting the values of f within the samples the feature f lose its information content referred to the classification of the samples.

Using this extension of the random forest classifier we are able to evaluate the importance of each feature regarding the classification of the samples. This way we are able to use random forest for feature selection.

2.3.7 Information gain for feature selection

For feature selection we use 'Information Gain Attribute Evaluation' implemented in Weka. Following we refer to this algorithm as information gain.

The features are ranked by the information one gains about a class C on condition of knowing the value of of the feature F. This information gain $I(C, F)$ of a feature F due to a class C is defined as

$$I(C, F) = H(C) - H(C|F).$$

where the entropy $H(C)$ [31] is a measure of the uncertainty connected with a class C and the conditional entropy $H(C|F)$ gives the uncertainty about C if one knows the value of feature F.

This definition is analog to mutual information described in Section 2.1.5. In this work, we use mutual information to quantify the dependency between two genes. For feature selection with information gain we use the information gained about a class C by knowing the value of a feature F for evaluating the importance of F.

2.3.8 GA/SVM for feature selection

A genetic algorithm (GA) is a heuristic optimization procedures belonging to the class of evolutionary algorithms [56]. They are used for solving non-efficient computable problems by generating multiple potential solutions, modifying and combining this solutions with each other and this way create solutions better fitting the given problem.

GAs are inspired by natural evolutionary processes such as inheritance, mutation, crossover and selection. In general, the GA passes through the following steps.

1. **Initialization.** The GA starts with a number of candidate solutions, the so called start population or first generation. The individual candidate solutions are also called individuals or chromosomes.

2. **Evaluation.** We evaluate the quality of each individual using a so called fitness function. This fitness function applies a score to each candidate solution depending on the given problem.

3. **Selection.** The candidate solutions are randomly selected. Here we select better individuals with a higher probability than others.

4. **Recombination.** The selected individuals are permuted and recombined to produce new individuals.

5. **Mutation.** Some of the candidate solutions are randomly modified.

6. **Next generation.** We evaluate all individuals generated in the steps 3 and 4 using the same fitness function as in step 2. The best individuals of the new grown individuals and the actual generation become the new actual generation.

7. **Termination.** We continue with step 2 until a termination criterion is fulfilled. This could be for example the number of generations or the quality of the best individual in the actual generation.

The core of the GA is the fitness function that determines the quality of the results. Only using a suitable fitness function enables the GA to find results that fit the given problem.

Wrapping a GA around a support vector machine (SVM) we are able to select a small set of features that is well suited for classifying a given set of samples. The classification accuracy of an SVM serves as part of the fitness function. As to the best of our knowledge there is no freely available implementation of the GA/SVM we implement the algorithm by our own. Below, we describe some details of the implementation and the used parameters.

Our implementation of a GA/SVM

Our implementation follows the general structure of a GA given above. Starting with an initial population we use chromosomes of this population for recombination and mutation. After evaluating the fitness of each chromosome we use the best individuals as the actual generation. This process is repeated until $it_{max} = 25$ generations are reached. (An analysis of our results shows that on average after 20 generations the fitness no longer increases.)

Algorithm 3 Genetic algorithm

```
1: pop ← Create initial population
2: for i = 1 to it_max do
3:     Recombination(pop)
4:     Mutation(pop)
5:     for all chromosomes do
6:         evaluate fitness
7:     end for
8:     pop ← best individuals of pop
9: end for
```

Create initial population. We start with an initial population of 200 chromosomes ($size_{pop} = 200$). Each chromosome is generated by choosing a random number of features. The chromosomes are binary encoded containing one bit for each of the p features in the sample set (see Section 2.3.1). For each feature a specific bit is allocated with 1 if the chromosome contains the feature and 0 otherwise. Each chromosome of the initial population contains approximately $size_{chr} = 15$ features.

Algorithm 4 Create initial population

```
1: for i = 1 to size_pop do
2:     for all bits j in chromosome i do
3:         with probability = size_chr /p do
4:             j ← 1
5:     end for
6:     evaluate fitness of chromosome i
7: end for
```

As we are interested in sets of features that are as small as possible we start with chromosomes containing only a few features ($size_{chr} \approx 15$). Nevertheless we ensure that the initial population contain most of the features in our data set by creating a large number of chromosomes ($size_{pop} = 200$).

Evaluate Fitness. The fitness function is the core of the genetic algorithm. Its design is crucial for the quality of the obtained results. We are interested in small sets of features fulfilling the following criteria:

1. A high classification capability of the selected features.
2. A small number of selected features.

These two criteria usually compete with each other. We weight both criteria according to our analysis and combine both into a multi-objective fitness function. The fitness $f(chr_i)$ of a chromosome i is determined as

$$f(chr_i) = w \cdot acc(chr_i) + (1 - w) \cdot \frac{p - size_{chr}(chr_i)}{p}.$$

The first part of the fitness function f represents the weighted classification accuracy of an SVM with Gaussian kernel using the features contained in chr_i for training. For calculating the classification accuracy we use the implementation of LibSVM [54] with default parameters ($C = 1$, $\gamma = 1/\ size_{chr}$) and a 6-fold cross-validation.

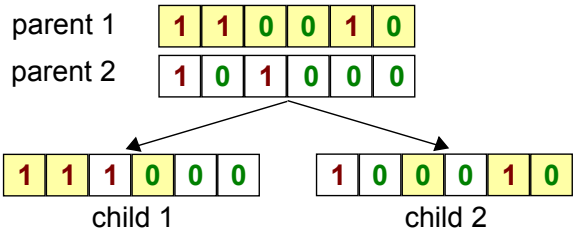


Figure 2.6: Uniform crossover of two binary encoded parent chromosomes resulting in two child chromosomes.

The second part of the function is responsible for minimizing the number of selected features. On condition that p (the number of features in the sample set) as well as $acc(chr_i)$ and w are constant we obtain a higher value for $f(chr_i)$ the fewer features chr_i contains.

The variable w denotes a weight factor. As we are primarily interested in feature sets showing high classification accuracy, we rate this criterion much higher than the number of features. For this reason we choose a large value for $w = 0.8$ [57]. Even so, if the classification accuracy of two chromosomes is equal the fitness the chromosome containing fewer genes is preferred.

Recombination. For recombination we select two different chromosomes of the current population using roulette wheel selection. This function allows to select a chromosome with a probability that directly corresponds to the fitness of the chromosome $f(chr_i)$. The probability $\Pr(chr_i)$to select a chromosome i is determined as

$$\Pr(chr_i) = \frac{f(chr_i)}{\sum_{k=1}^{size_{pop}} f(chr_k)}$$

The two selected parent chromosomes breed two child chromosomes by uniform crossover [58]. For each bit of the parent chromosomes it is chosen randomly which child inherit the characteristic of which parent The process is illustrated in Figure 2.6. As the binary string does not contain any information about the relationship of the features we choose uniform crossover over more popular methods as one- or two-point crossover.

The number of recombination depends on the size of the population. Altogether, we repeat the recombination process $size_{pop}$ times to obtain $2 \cdot size_{pop}$ child chromosomes. As this number of new chromosomes possibly completely covers interesting aspects of the current population we reduce the number of included child chromosomes. Each child chromosome is added to the current population with a probability of $prob_{recomb} = 0.4$.

Algorithm 5 Recombination

```
1: pop ← current population
2: for i = 1 to size_pop do
3:     p1 ← Roulette Wheel Selection(pop)
4:     p2 ← Roulette Wheel Selection(pop)
5:     child[ ] ← Uniform Crossover(parent1,parent2)
6:     with probability = prob_recomb do
7:         add child[1] to pop
8:     with probability = prob_recomb do
9:         add child[2] to pop
10: end for
```

Mutation. Subsequent to the uniform crossover, a mutated version of each chromosome in the current population is added. We flip each bit in the chromosome with a probability depending on the average number of features in the start population ($size_{chr}$), the number of features in the data set (p) and a mutation probability ($prob_{mut} = 0.1$). The probability to flip a bit is determines as

$$prob_{flip} = prob_{mut} \cdot \frac{size_{chr}}{p}.$$

As the genetic algorithm improves the chromosomes in small steps flipping a large number of bits would push the algorithm into a random search in exponential space. For this reason we only choose a small mutation probability. After mutation all chromosomes are added to the current population.

Algorithm 6 Mutation

```
1: pop ← current population
2: for all chromosome j do
3:     for all bit i in j do
4:         with probability = prob_flip do
5:             flip bit i
6:     end for
7: end for
8: add all chromosome to pop
```

Next generation and termination After recombination and mutation each chromosome is evaluated using the fitness function f. We remove the chromosomes with the smallest fitness until the current population contains

the same number of chromosomes as the initial population ($size_{pop} = 200$). The GA/SVM terminates after 25 generations are reached.

GA/SVM200 and GA/SVM500.

In our work we compare our GA/SVM with information gain and random forest. These methods result in a list of all input features ranked by their importance. In contrast our GA/SVM returns a small set of features that is well suited for classification. To obtain a ranked list comparable to those of information gain and random forest, we run the genetic algorithm multiple times and rank each feature by its frequency of occurrence in the small sets selected during the single runs. In this work we compute a combined list from either 200 runs of the GA/SVM (GA/SVM200) or 500 runs (GA/SVM500).

2.4 Software packages

Most of the used methods are implemented in the following freely available software packages.

The R project (v.2.13.0) R [14] is a freely available software environment as well as a programming language. It is is widely used and provides various statistical and graphical methods. Developed in 1992 by Ross Ihaka and Robert Gentleman, it is derived from the statistical programming language S. Since 1993 R is under the GNU General Public License (www.gnu.org).

Some of the statistical methods we use as part of our data preprocessing as well as for the analysis of our results are implemented in R. Specifically we use R for calculating the t-test statistic and the χ^2-test statistics both implemented in the R package 'stat' [27] as well as for calculating the false discovery rate correction implemented in the R package 'qvalue' [30] and the mutual information implemented in the R package 'parmigene' [35]. Furthermore the accuracy diagrams and expression plots as well as the density distributions in the method section are plotted in R.

Affymetrix Power Tools (v.1.14.3) The Affymetrix Power Tools [42] (APT) are a set of open source command-line programs under the GNU General Public License. They implement algorithms for the analysis and the computation on Affymetrix GeneChip arrays.

Besides other things it provides algorithms for the fast preprocessing of large numbers of microarrays. It contains different methods for computing the expression measures of multiple microarray chips, as RMA [37, 38],

PLIER [59, 60] or MAS5.0 [61]. For each of these algorithms various background adjustment methods as well as various normalization and summarization methods can be chosen.

We use the APT to perform the first part of our data preprocessing described in detail in Section 3.2.2, Therefore we use RMA with a background adjustment using only perfect match intensities, quantile normalization and median polish as summarization method.

Weka (v.3) Weka [44] is an open source software under the GNU General Public License that provides a collection of machine learning algorithms for classification, regression, clustering and other data mining tasks. It is implemented in Java and provides interfaces for external Java implementations.

For classification we use naive Bayes classifiers, C4.5 decision trees, k-nearest neighbor, random forest and support vector machines implemented in Weka. Furthermore for feature selection we use an information gain method implemented in Weka as well as a Weka extension of random forest implemented by Livingston [62].

LibSVM (v.2.88) LibSVM [54] provides implementations of support vector machines for regression, classification and distribution estimation. It provides Java as well as C++ sources and various interfaces to other programs like R, Weka, MATLAB or Python. LibSVM is implemented under a free software licenses that is similar to the GNU General Public License.

We use the Java source code of LibSVM as part of our genetic algorithm. It is used for the evaluation of the feature sets selected by the genetic algorithm described in section 2.3.8.

Chapter 3

Microarray data and data preprocessing

In this chapter we first give an introduction to DNA microarrays with a focus on Affymetrix GeneChip arrays. These are special kinds of microarrays analyzed in this thesis. As we do not perform any own microarray experiments we use data freely available at Gene Expression Omnibus.

In the second part of this chapter we describe the data acquisition and preprocessing for our two data sets. In this work we use one data set for the identification of pluripotency biomarkers and the second data set for the identification of Alzheimer biomarkers.

3.1 DNA microarrays

In fact a large number of DNA microarray types exist. In general, they can be classified based on the kind of usage, the spatially disposal of the probes and the manufacturing processes. They are used for measuring the expression level of thousands of genes simultaneously [4, 63]. Other applications are for example the comparison of the genomes of different cells or different organisms [64–66] and the detection of variations of single base pairs within or between populations, so called single nucleotide polymorphisms usually referred to as SNPs [67, 68]. A lot of different fabricators are established whose arrays differ in numerous details.

In the following, we give an introduction into the architecture of microarrays, the design of microarray experiments and the preprocessing of the raw microarray data.

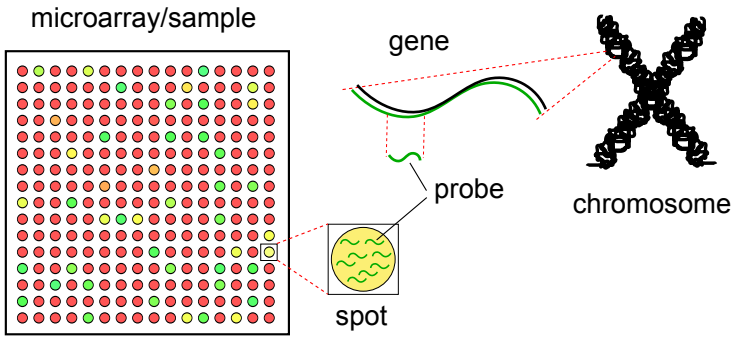


Figure 3.1: Structure of a DNA microarray.

3.1.1 Architecture of DNA microarrays

Figure 3.1 illustrates the structure of a DNA microarray. The basis of an array is a solid substrate, usually a thin glass slide, but other materials like silicon or plastic are used as well. The surface is spotted with small amounts of specific DNA sequences, so called probes. Based on the principle of Watson-Crick base pairing [17], the probes are able to bind to specific complement DNA (cDNA) or complement RNA (cRNA) sequences, called targets. The microscopic DNA spots are ordered in a two-dimensional array and can be distinguished clearly from each other. A DNA microarray may contain thousands of those spots and each spot may contain millions of copies of identical probes.

The core of a DNA microarrays are the DNA sequences of the probes spotted to the array. They determine which cDNAs hybridize to the array. A single probe consists of 25 (Affymetrix GeneChip) to 70 bases. As a gene usually consists of several thousand bases, microarrays designed for gene expression profiling consist of probes that are short subsequences of the examined genes. The DNA sequences have to fulfill several criteria such as a high degree of specificity and sensitivity. There is a whole branch in bioinformatics working on the design of the probe sequences for DNA microarrays [69–73].

Specifics of Affymetrix GeneChip® arrays

Affymetrix GeneChip arrays are DNA microarrays provided by Affymetrix (www.affymetrix.com). These types of microarrays show two characteristics that differ slightly from the general description of DNA microarrays given above.

First, several spots on the array correspond to the same gene. This way the gene expression products can be detected with a higher specificity com-

pared to other DNA arrays. This requires an additional preprocessing step that combines the intensities of multiple spots to obtain the expression level of the corresponding gene.

Second, for each probe that perfect matches a subsequence of the corresponding gene there is a mismatch probe that differs in a small percentage of nucleotides from the perfect match probe. In this way unspecific bonds can be identified and eliminated during an additional preprocessing step.

3.1.2 Steps in a microarray experiment

Microarrays can be used for multiple applications. As we are interested in the expression of genes we use microarrays designed for measuring the quantity of messenger RNA (mRNA) in a biological sample.

If a gene is expressed the DNA sequence of the gene is transcribed to an mRNA. It is the first measurable product of an expressed gene and serves as the indicator for the intensity of gene expression. To detect the expression of a gene the number of specific mRNA molecules is measured using DNA microarrays.

In a first step the samples are purified and the mRNA is separated from other cellular components. Via reverse transcription the mRNA is translated in a cDNA strand. The cDNA strands become labeled in a separate coupling step before or after hybridization depending on the type of DNA microarray. The labeling is either fluorescent or radioactive.

The prepared samples are hybridized to the microarrays by forming hydrogen bonds between the nucleotides of the probes and the complementary nucleotides of the targets. In this process it is possible that targets that are not the exact complement of the probe sequence bind anyway. The strength of hybridization depends on the number of hydrogen bonds formed between the two DNA strands. For this reason those non-specific target bindings are less strong than correctly bound DNA strands and can be washed off.

This way, only targets complementary to the probes stay hybridized. After hybridization and washing the intensity of the fluorescent or radioactive signal of each spot can be scanned using a laser that excites the labeled substances on the array and a detector that measures its intensity. For each spot the intensity, following also referred to as probe intensity, is given by a real number value. The more mRNA of a gene exists in a sample, the more labeled targets hybridized to probes belonging to the specific spot on the array and the higher is the probe intensity of this spot. We illustrate the steps of a DNA microarray experiment in Figure 3.2.

To ensure the integrity of the raw data it has to be quality assessed before further applying several preprocessing steps. Only then the data can be used

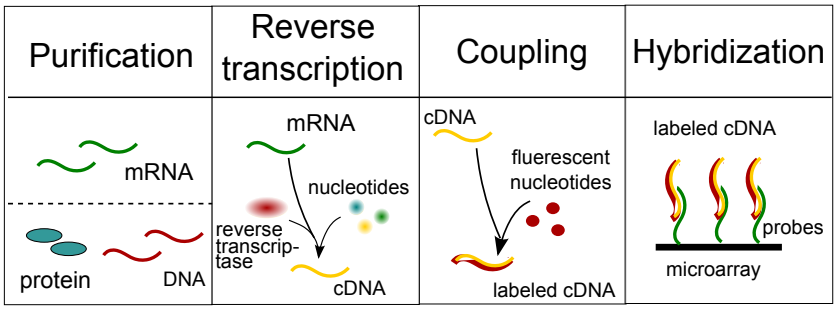


Figure 3.2: Steps in a microarray experiment.

for the actual microarray analysis.

3.1.3 Preprocessing of microarray data

Before gene expression analysis the raw data has to be preprocessed to avoid random variations in the data. The preprocessing of DNA microarrays contains at least background correction and normalization. Also, additional steps as perfect match correction and summarization can be necessary depending on the specific array type.

For each of these steps there are various methods that show good results in different experiments. In the following we describe the principal preprocessing steps applied to Affymetrix GeneChip arrays.

Background correction. Background fluoresce can be a consequences of non-specific bounds of labeled sequences to the microarray surface or of fluorescent residue that remain on the array surface after the washing process. The background correction is performed for each microarray chip independently. The used methods range from simple methods that subtract the noise from the single probe intensities to complex convolution models of signal and noise distributions.

Normalization. Normalization is needed when combining multiple arrays for a later analysis. There will always be minor variances in the hybridization process used for two different arrays. A different amount of RNA in the examined sample or the time used for hybridization may influence the overall intensity of an array. Even the usage of different scanners for reading the fluorescence intensities on the array can influence the result. The most popular methods for the normalization of microarrays is the quantile normalization [39].

Perfect match correction The architecture of Affymetrix GeneChip arrays allows an additional correction step. Besides the probe sequences that perfectly match subsequences (pm probes) of the corresponding genes the microarrays contain probes that differ in a small number of nucleotides, so called mismatch (mm) probes. The mm probes give a measurement for the non-specific bounds of the associated pm probe. As shown in various studies including mm probes may cause several problems [38]. For this reason most often the mm probes are ignored using just the pm probes for analysis.

Summarization Because of the design of the array with multiple probes that correspond to a single gene, the summarization is an elementary step to get the actual gene expression levels. For each gene the intensities of all associated probes are combined to obtain the actual expression value for each gene on the array.

A large number of data preprocessing algorithms for Affymetrix GeneChip arrays are published, including RMA [37,38], PLIER [59,60], dChip [40] and MAS5.0 [61]. All these methods are accepted standards for microarray data preprocessing.

3.2 Data acquisition and preprocessing

As we are interested in biomarkers for pluripotency and Alzheimer's disease, we assemble two data sets for our analysis. The data of all microarray experiments included in the two data sets are freely available at Gene Expression Omnibus (GEO), a data repository for microarray data.

In the following section we describe the selection of our data sets in detail. As we include only raw data of microarray experiments, the data sets have to be preprocessed before we are able to analyze them. Section 3.2.2 gives a detailed description of the preprocessing we apply to each of our data sets.

3.2.1 Data acquisition

We use two different data sets to find potential biomarkers for pluripotency on the one hand and Alzheimer's disease on the other hand. In the following we refer to these two data sets as PLURI data set and AD data set.

Both data sets consists of two kinds of samples. The PLURI data set contains samples of pluripotent and non-pluripotent cells whereas the AD data set contains samples of Alzheimer affected patients and a healthy control group. For a better understanding we use the terms positive samples or

positive labeled samples for the pluripotent and Alzheimer effected samples in the respective data set and negative samples or negative labeled samples for the non-pluripotent samples and the samples of the control group for Alzheimer.

Studying pluripotency we usually have to deal with studies containing only a few micro arrays, so they are not well suited for machine learning methods. As we see in Table 3.1 in contrast to pluripotency for Alzheimer's disease many large gene expression studies exist. These studies consisting of more than 100 micro arrays and so they are useful for machine learning. In the following we want to explain the selection of data series for the PLURI data set and the AD data set in detail.

The PLURI data set

Starting our work on pluripotency-related data in Mai 2010, the according data series available at GEO consisted of no more than 20 samples. For this reason we collect the data of several series and combine them.

To ensure a reliable merging of the data series we assume that all data are based on an identical set of probes. Therefore we decide to use only data series processed on the same gene chip. We use the Affymetrix mouse 430.2 oligonucleotide chip [74] which is one of the most popular platforms available and is used to perform a large amount of microarray studies related to pluripotency.

Based on the description of the single samples we manually identify samples as pluripotent and non-pluripotent. Samples we cannot clearly classify we dismiss rather than taking the risk of adding wrongly labeled samples. To include a GEO series to our data set, it has to contain at least one sample that we can identify as pluripotent. Usually the included series consists of a mixture of pluripotent and non-pluripotent samples. That includes embryonic stem (ES) cells (up to embryonic day 3.5) as well as pluripotent germline stem cells and induced pluripotent stem (iPS) cells. In contrast, the non-pluripotent samples arise from all kinds of differentiated cells. We choose samples of embryoid bodies (day 5 and older), germline stem cells and partially reprogrammed iPS cells as well as cells of completely differentiated tissues as liver, lung and brain. For a detailed list of the included GEO data series see Appendix A.1.

The resulting data set consists of 146 pluripotent (positive labeled) and 140 non-pluripotent samples (negative labeled). As the data set contains of multiple biological and technical replicates the samples are partially correlated. The partitioning of the data set in three independent folds is described in Section 3.2.2.

GEO Series	publication	microarray	samples	reference
GSE29676	2011-07-13	Invitrogen ProtoArray v5.0	609	[75]
GSE15222	2009-04-10	Sentrix HumanRef-8 Expression BeadChip	363	[76]
GSE5281	2006-07-10	Affymetrix Human Genome U133 Plus 2.0 Array	161	[77, 78]
GSE26927	2011-01-29	Illumina humanRef-8 v2.0 expression beadchip	118	
GSE6834	2007-01-23	Ion Channel Splice Array	60	[79]

Table 3.1: Large microarray studies for Alzheimer's diseases sorted by the number of containing samples (retrieved 2011-09-14).

The AD data set

Other than for embryonic stem cells, some large microarray studies exists for Alzheimer's disease. Table 3.1 gives an overview over the five largest microarry studies available at GEO (retrieved 2011-09-14). As we started analyzing the AD data set in January 2011, GSE29676 was not yet available. For the series GSE15222 which was at the time the largest series for Alzheimer's disease no raw data was provided. For this reason we use the series GSE5281 as one of the most exhaustive data sets with provided raw data. The microarray experiments are done on an Affymetrix human genome U133 Plus 2.0 array [80]. This data series consists of 161 samples of six different brain regions. 87 samples are of patients diagnosed with Alzheimer's diseases (positive labeled) whereas the other 74 samples come from a healthy control group (negative labeled) [78]. As the samples partially come from the same donors we assume the samples to be correlated. Such a correlation cannot be precluded for the Alzheimer disease affected samples as well. In Section 3.2.2 we describe the partition of the data set in three independent parts to enable a proper cross-validation.

The two Affymetrix GeneChips®

Affymetrix provides microarray techniques widely used in medical and biological research [81]. We work with gene expression data of two Affymetrix GeneChip arrays [82], the Mouse Genome 430 2.0 Array [74] and the Human Genome U133 Plus 2.0 Array [80]. Both arrays are single-channel

DNA microarrays with in situ synthesized probes. For the transcription level measurement of each target sequence eleven pairs of probes are used. Each probe pair consist of a perfect match probe (pm) and a mismatch probe (mm) with 25 oligonucleotides each. The pm probe is a complemented subsequence of the target DNA transcript. The mm probe distinguishes from the pm probe in one oligonucleotide. The middle base is changed to the Watson-Crick complement. So the intensities of the mm probes can be taken as measure for non-specific hybridization. The sequences used to design the arrays are collected from GenBank® [83], dbEST [84], and RefSeq (www.ncbi.nlm.nih.gov/RefSeq/).

The Mouse Genome 430 2.0 Array [74] consists of over 45,000 probe sets representing over 34,000 mouse genes. The Human Genome U133 Plus 2.0 Array [80] allows the simultaneous investigation of more than 47,000 transcripts and variants, including more than 38,500 genes.

3.2.2 From raw data to preprocessed data sets

After selecting the raw microarray data for our data sets, we preprocess the two data sets separately.

We apply the same preprocessing steps on both data sets. In the following section we describe the data preprocessing. Besides the preprocessing absolutely necessary for Affymetrix GeneChip arrays, we also use additional filtering procedures to reduce the number of genes before machine learning. As classical machine learning methods are known to show poor results dealing with a large number of unimportant genes we previously remove the less promising genes of the data sets. Furthermore we prepare the data sets for cross-validation that allows us to make a clear statement about the quality of our feature selection algorithms.

Preprocessing

For the preprocessing of Affymetrix GeneChips a large number of algorithms exist. The most commonly used methods include robust multichip average (RMA) [37, 38], probe logarithmic intensity error (PLIER) [59, 60], DNA-chip analyzer (dChip) [40] and Affymetrix microarray suite 5 (MAS5.0) [61]. Recent studies compare the performance of several methods [85–87], but it is shown that all methods have advantages as well as disadvantages. As RMA in many cases outperforms other methods we use this algorithm for the data preprocessing.

We use Affymetrix Power Tools [42] that include an RMA implementation provided by Affymetrix (www.affymetrix.com). The background adjustment

is performed using only unmodified perfect match intensities. After quantile normalization and quantification with median polish the resulting expression values are $\log_2$-scaled. The whole RMA algorithm is described in detail in Section 2.2.1. Figure 3.3 shows the density curves and the boxplots of both resulting data sets.

After preprocessing we obtain a matrix containing the expression levels for each gene on the array. In fact, for some genes we obtain more than one expression value on each array. These results from different splice variants of the same gene measured simultaneously on the array. In our work we are primarily interested in the expression of different genes independent of their variants. For this reason we combine all variants that correspond to the same gene symbol of the UniGene record [88], by calculating the mean value of their expression levels. At this point, each of the two data sets consists of more than 15,000 genes.

Partitioning

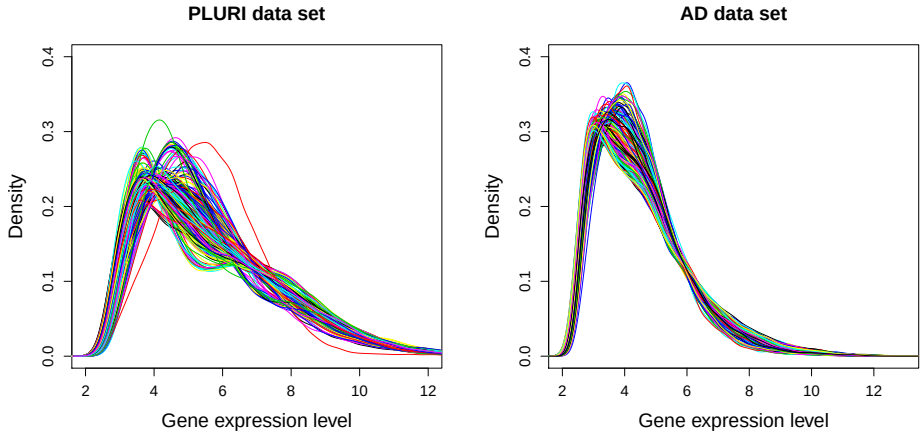
For the performance analysis of different classification and feature selection methods in Chapter 4, we use a 3-fold cross-validation. For this we split the samples into three disjoint subsets. The classifiers and feature selection methods are trained using two of the subsets and tested on the remaining one. This procedure is repeated three times using each subset for testing exactly once.

In case of independent samples we can randomly split the data into three subsets. Unfortunately, the samples of our data sets are partially correlated, as mentioned in Section 3.2.1. For this reason when partitioning the data we have to ensure that correlated samples are classified into the same subset.

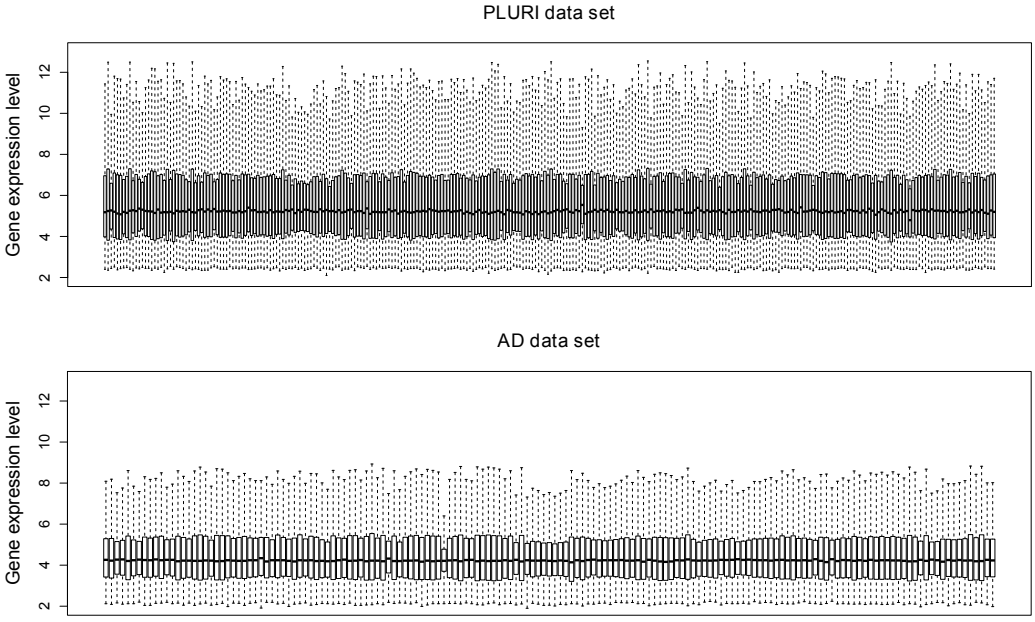
PLURI data set. The PLURI data set contains plenty of biological and technical replicates. We assume that all replicates of the same sample are correlated. For this reason we sort all replicates of a sample into the same subset. The partitioning of the samples in three disjoint subsets is listed in Appendix A.2.

AD data set. For each array the gender and the age of the tissue donor are recorded. Samples with the same gender/age combination are always sorted into the same subset. This way we use the two characteristics to classify the samples into disjoint subsets. The partition of the samples can be found in the Appendix A.3.

The following filtering process is done for each training set separately. This



(a) Density curves: Illustrate the distribution of the gene expression values on each array.



(b) Boxplot: Illustrate the range and the median of the gene expression levels on each array.

Figure 3.3: Visualization of the two preprocessed data sets.

way we can test the performance of our algorithms on completely independent test sets. Furthermore we use the same subsets for all feature selection and classification methods. This ensures a high quality comparison between the methods. Consequently, for our analysis in Chapter 4 using a 3-fold cross-validation we have to apply following filtering process separately to all three training sets for the PLURI data set as well as for the AD data set.

For the actual identification of potential biomarkers in Chapter 5 we do not validate the results using cross-validation. So we are able to use a data set containing all samples for feature selection. For this reason the following filtering process is also applied to the complete data set.

Filtering

In machine learning dealing with a huge number of features while only having a small amount of labeled samples often decrease the performance of the algorithms. In our case the features are genes and the labeled samples are the positive and negative microarray samples. So, instead of using more than 15,000 genes and less than 300 samples for training and testing, we reduce the number of genes to 1,000.

The idea of the filtering procedure is the elimination of those genes that are probably less important for distinguishing positive and negative samples. We assume that genes with no differences in expression values between the two groups of samples are less important than genes with a large difference between positive and negative samples.

We want to identify genes differentially expressed in the two classes of each data set. We select those genes having similar expression values within these two groups, but different expression values between. Hence, we apply a two tailed t-test for samples with unequal variances to the data and test the difference in mean expression for each gene. The t-test is described in Section 2.1.2. Figure 3.4 shows the distribution of the p-values for the two data sets containing all samples.

As we apply the t-test multiple times on the same data set we adjust the p-values using false discovery rate correction (FDR). This reduces the number of false negatives. The concept of FDR is described in Section 2.1.4. As we expect an error of 5% we dismiss all genes with a corrected p-value (q-value) larger than 0.05.

We are interested in genes with a high difference in expression levels between the groups even if the variance within these groups is relatively high. So in a second step we use the fold change (FC) to get the best 1,000 genes out of the remaining data set. We sort the genes by their FC calculated as described in Section 2.1.1. The first 1,000 genes serve as our data set.

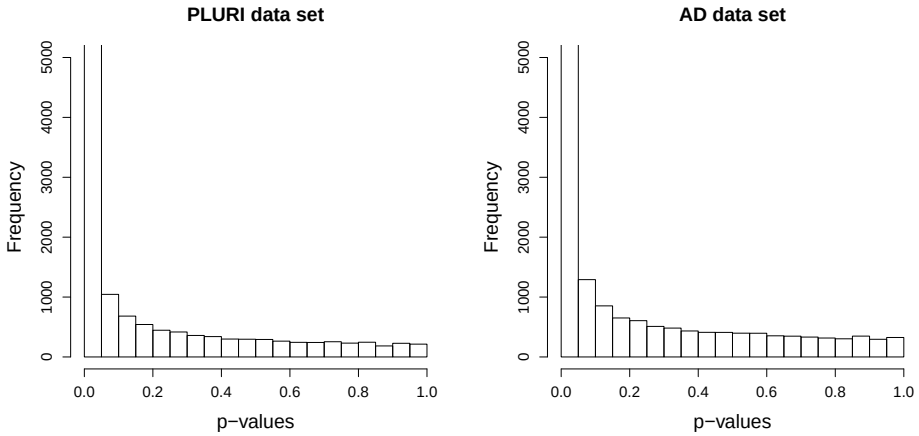


Figure 3.4: The p-value distribution for the two data sets.

The filtering process results in a gene expression matrix containing the expression values of the 1,000 most differentially expressed genes of each array.

Chapter 4

Analyzing feature selection on microarray data

In this chapter we compare the quality of different feature selection methods with each other. We use two microarray data sets selecting genes that play an important role in the maintaining of the pluripotent cell state, on the one hand, and in Alzheimer's disease affected brain tissue, on the other hand.

As microarray data contain a large number of gene expression values statistical methods as well as machine learning and feature selection algorithms are widely used for analyses [17].

After given an introduction to common statistical a machine learning approaches for the analysis of microarray data we focus on feature selection for the identification of important genes.

Comparing the quality of three feature selection methods in Section 4.2, we find that, independent of the used data set, our wrapper of genetic algorithm and support vector machine (GA/SVM200) outperforms information gain and random forest by far. We trace the quality of the GA/SVM200 to the small gene sets selected within the algorithm described in Section 2.3.8. As, to the best of our knowledge, those small gene sets have not yet been investigated in-depth in Section 4.3 we examine the dependencies of genes occurring together in a small set.

4.1 Methods for the identification of important genes from microarray data

In the following we give an introduction to different statistical, machine learning and feature selection techniques for analyzing microarray data. In particular we pay attention to the usability of these methods for the identification

of important genes.

Additionally, we focus on methods for feature selection that show good results identifying potential biomarkers described in Section 4.1.3.

4.1.1 Statistical methods

Statistical methods are widely used to identify genes differentially expressed in different groups of samples [89]. For the identification of those differentially expressed genes calculating the fold change, as described in Section 2.1.1, is the most intuitive method. As the fold change gives no estimation for the statistical significance it is agreed that the fold change alone is not a valid method [90].

To estimation the statistical significance of a differentially expressed gene we use statistical methods that estimate how likely the observed difference is due to chance. Usually we consider a gene as significantly differentially expressed if the probability of observing this difference by chance is below 5% (p-value < 0.05). Popular methods are for example the t-test, described in Section 2.1.2, MaxT [91] , SAM [20] or ANOVA [92].

Applying a test statistic to the each gene of a microarray each time we have a 5% chance to identify a gene as differentially expressed even if it is not. As a microarray usually contains thousands of genes we expect a large number of genes incorrectly identified as differentially expressed. To reduce the number of those incorrectly identified genes two methods are widely used. False discovery rate (see Section 2.1.4) correction as well as the more conservative family-wise error rate correction [93] show good results in adjusting the p-values.

Various software implementations offer a large number of statistical approaches for the identification of differentially expressed genes in microarray data. Besides many specific packages for the free software environment R commercial tools such as MATLAB or SAS provide packages for the analysis of microarray data.

Although statistical methods are able to identify differentially expressed genes, often the classification accuracy of those genes is not very high. For this reason we use feature selection methods for the identification of genes best suited for distinguishing different samples. Even so, we use statistical methods, more concrete fold change, t-test and false discovery rate correction, in a previous filtering step to reduce the number of genes to 1,000. The filtering process is described in Section 3.2.2.

4.1.2 Machine learning methods

Classic machine learning methods as support vector machines (SVM) or random forests are very popular for analyzing microarray data [94,95]. Machine learning methods are used to reveal correlations between samples and classify new samples into known relationships. The identification of genes allowing this clustering or classification is not directly implicated. Extending those algorithms or combining them with additional methods allows the application for biomarker discovery.

Learning algorithms can be divided into unsupervised and supervised methods. In the following we give a brief introduction to these two kinds of machine learning techniques.

Unsupervised methods

Unsupervised data analysis does not include any prior knowledge about the data. It encompasses clustering techniques that are used to group similar objects into the same cluster. There are two possibilities for clustering analyses. On the one hand we can cluster similar genes and on the other hand we can also focus on the similarities between different samples. A not very popular but promising hybrid of these two approaches is the so called biclustering. Here one groups genes at the same time as the samples [96]. In the following we describe the clustering approaches for clustering similar genes but the same methods can be used analogously for sample clustering.

Probably, the most popular clustering technique in microarray data analysis is the hierarchical cluster analysis. Starting with a distance matrix that gives a pair-wise similarity measure for all genes the two genes showing the highest similarity in gene expression levels are grouped in a cluster. The distance of the new cluster to all remaining genes is calculated and the process is repeated until only one cluster remains. A problem of hierarchical clustering is that suboptimal decisions cannot be corrected in later progress.

Non-hierarchical methods avoid this problem by classifying the genes into a predefined number of clusters using a replacement function to optimize a given objective function.

The most sensitive decision is the choice of the distance metric that gives a measurement for the similarity of two objects. Popular metrics are the Euclidean [48] as well as the Manhattan [48] distance. As no clear guidelines for the choice of the metric exist and the kind of metric is crucial for the resulting clusters the solutions are usually controversial.

Clustering methods always generate patterns but it is not clear if the patterns observed on the sample data also represent patterns in new sam-

ples. Although there are resampling based methods that could solve those problems tests raise doubts that many clustering methods are not able to generate clusters from sample data that reflect patterns in new samples [97].

It is assumed that clustering methods are used far too often for microarray data analysis [90]. Besides the difficulties in evaluating the results cluster analysis does not primary answers the question which genes are most important for the clustering decisions. Instead it is used for grouping genes showing the same patterns in expression level or grouping samples showing similar expression level patterns. Nevertheless there is a high number of widely used clustering techniques such as principal component analysis [98], k-means clustering [99] or self-organizing maps [100].

Supervised methods

Supervised or classification methods encompass techniques that are usually used to assign a class label to a given sample. In microarray analysis a sample contains the measured intensities of a microarray, in our case gene expression levels. The classifier is trained using a set of samples that is independent of the sample set used for testing. For the training set as well as for the test set the class labels are known. Classifying the test set with the trained classifier we are able to determine the classification capability of the classifier before applying it to new samples with unknown class labels.

There are dozens of classification methods available showing good results in solving multiple problems [101]. The choice of a classifier mainly bases on the number of samples and the complexity of the classification model. It is assumed, that small sample sizes require simple models with only few parameters or even the use of less complex strategies as statistical or clustering techniques. In old days, microarray studies only contained few experiments, but meanwhile the microarray chips become cheaper and complex studies with more than 100 arrays are not a curiosity any longer. So, complex classifiers are usable as well as simple ones.

Probably, the most popular supervised learning algorithm used for microarray data analysis is the support vector machine (SVM) [102, 103], detailed described in Section 2.3.5. Other widely used approaches for classifying microarray data are implementations of decision trees and random forests described in Section 2.3.2 and Section 2.3.6. Extensions of those classifiers such as SVM for recursive feature elimination (SVM-RFE) [104] show good results in feature selection.

4.1.3 Feature selection methods – Identification of potential biomarkers

Microarray data series usually contain thousands of genes but only a small number of samples. Machine learning methods are originally not designed for dealing with a large number of unimportant features. In combination with feature selection those methods can exploit their full potential.

Feature selection methods can be applied to unsupervised as well as to supervised methods. Although traditionally more attention is drawn to the supervised applications unsupervised methods show good results on various problems, too [105–107].

In supervised classification feature selection can be used to avoid overfitting and improve the performance of classifiers as well as for building more efficient models. In unsupervised learning the elimination of unimportant features increases the probability that patterns found by the clustering algorithm also represent for actual patterns. Despite all benefits feature selection offers there is no guaranty that the selected subset of features is optimal. Besides the errors of the learning methods additional errors may arise from the experimental design of the data.

Feature selection methods are widely used to reduce the number of features in a data set [6, 108–110]. Compared to other techniques as projection [98] and compression [111] that also decrease the number of features, feature selection methods do not modify the representation of the data. They conserve the content of the single features and so the results are more intuitively interpretable. When analyzing microarray data as features we use the genes measured by the array. Hence, feature selection in microarray data means the selection of a subset of important genes.

Feature selection methods

Feature selection methods can be categorized into the following classes based on the assembly of feature selection search and machine learning model.

Filtering approach. A filter method usually applies an important score to each gene. Then, the genes with the highest importance are selected as feature subset. Most filtering approaches are univariat what means that the importance of a gene only depends on the expression value of the single gene and does not include any correlations to other genes. Even if there are some multivariate approaches as Markov blanket filter [112] or fast correlation-based feature selection [113] the widely used methods are univariate. Those univariate filter methods contain statistical methods such as t-test statistic

[22], Wilcoxon rank sum test [114] or χ^2-test statistics [22, 115] as well as methods inspired by information theory as for example information gain, described in Section 2.3.7.

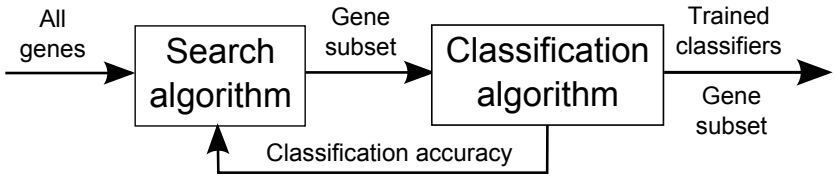
Embedded approach. Embedded methods use all genes of the data set to build a classifier. Then the classifier is analyzed to determine the importance of the single genes. The exact implementation of the analysis depends on the particular classification model. The importance of a gene does not only depend on the single gene but also considers correlations to other genes. Additionally, the importance of a gene is specific to the used classifier. Similar to filtering approaches the size of an optimal subset has to be chosen arbitrarily. Extensions of support vector machines (SVM-RFE) [104] as well as of random forests (described in Section 2.3.6) are popular embedded methods for feature selection.

Wrapper approach. Wrapper approaches select a subset of genes in dependency of a specific learning approach. Therefore, a search method is wrapped around a classifier. There are small differences between supervised and unsupervised methods. After a subset of genes is selected in supervised wrapper approaches the gene subset is evaluated by training and testing the classifier on the subset of genes. In contrast, in unsupervised approaches the subset is clustered and the resulting patterns are evaluated by an additional criterion function. Figure 4.1 illustrate the wrapper approach for supervised and unsupervised learning methods.

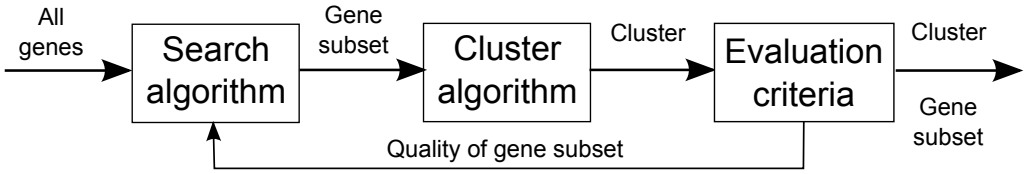
We classify wrapper approaches into two categories based on the search method wrapped around the classifier.

Greedy approaches as sequential backward elimination [116] and sequential forward elimination [117] use greedy search strategies as for example hill climbing [118] algorithms. Sequential backward elimination starts with a set containing all genes of the data set and stepwise eliminates the least promising genes. Therefore, a gene can only be eliminated if this does not decrease classification capability of the inner classifier. As the elimination of a gene is irreversible the algorithm has a high probability coming to a standstill in local (not global) optima. Analog to sequential backward elimination, sequential forward elimination starts with an empty gene set and recursively increases the number of genes in the set.

In contrast, random or stochastic approaches use heuristic search strategies designed for large scale combinatorial problems. Besides the genetic algorithm described in Section 2.3.8, ant colony optimization [119] and simulated annealing [120] are popular methods. Random search methods are



(a) Wrapper for supervised learning.



(b) Wrapper for unsupervised learning.

Figure 4.1: Wrapper approaches for supervised and unsupervised learning.

computational expensive but efficiently include gene interactions as well as redundancies for determining the importance of a gene. In contrast to wrapper methods using greedy search algorithms they are more robust against local optima.

Compared to filter or embedded methods an advantage of wrapper approaches is the fix size of the selected subsets. As filtering and embedded methods result in a ranked list of genes to obtain a subset of genes we have to arbitrarily chose a cutoff.

Identifying potential biomarkers

The definition of biological markers, so called biomarker, is not yet consistent [121]. Most often the definition given by the biomarker definition working group of the national institutes of health (NIH) (www.nih.gov) is used. The NIH officially define the term biomarker as

> a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention [3].

In our work we select important genes from microarray data using different feature selection methods. Microarrays are used for measuring the expression level of thousands of genes in a specific biological sample. In this work we use two data sets, one containing pluripotent and non-pluripotent samples the other containing samples of Alzheimer's disease affected brain

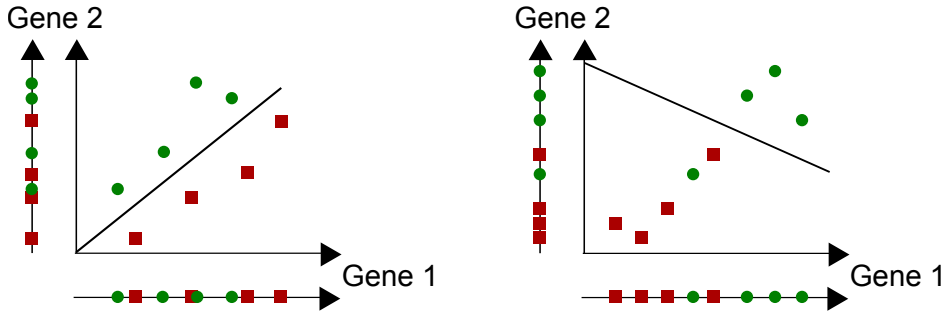


Figure 4.2: Problems occurring from combining biomarkers. Combining two genes with a low classification capability may increase the accuracy (left). In contrast, combining two genes that are both well suited for distinguishing the samples does not necessarily improve the results (right).

tissue as well as samples of non-affected brain tissue. As our data sets contain the measurements of samples from different biological states we rate a gene as important if it is well suited for distinguishing between the different states. This way the selected genes fulfill the given definition of biomarkers with the expression level as an objective measurable characteristic for a normal biological process in case of pluripotency and a pathogenic process in case of Alzheimer's disease.

As biological and medical issues, for example the processes in regulatory networks or the diagnosis of a particular disease, usually are very complex a combination of biomarkers often shows much better results than using a single biomarker. For this reason combining the best ranked genes of a feature selection method to a small set better suited for separating two classes of samples is a general approach [122–124]. Several studies deal with the topic of combining biomarkers under various aspects. Especially in clinical usage combining biomarkers is very important to get a reliable diagnosis [125–129].

However, combining single genes that show good discriminatory abilities does not necessarily lead to a gene set better usable for classification. If two biomarkers characterize more or less the same aspect of the classification both genes may have a high classification accuracy individually, but there is little gain when combining the two genes [130]. This problem is illustrated in Figure 4.2.

Another problem is that combining single biomarkers increases the complexity and therewith the costs of biological and clinical trials. For this reason we are looking for biomarker sets which are as small as possible and still have a high classification capability.

In this work, we use information gain (filtering approach), random forest (embedded approach) and a wrapper of genetic algorithm and support vector machine (GA/SVM, wrapper approach) for feature selection. All methods are described in Chapter 2. Filtering as well as embedded methods result in a list of genes ranked by their importance whereas wrapper methods usually select a close to optima subset of genes. To make the methods comparable we slightly modify our GA/SVM.

Nevertheless, the GA/SVM takes a special place among our methods. Besides the genes selected in the ranked list we analyze the small gene set selected by the GA/SVM. Using the GA/SVM for feature selection in microarray data is a common approach. Even so, to the best of our knowledge the quality of the particular small sets is not yet deeply investigated. In the following section we outline recent publications for feature selection in microarray data using wrapper approaches of genetic algorithm and support vector machine.

GA/SVM a promising approach for biomarker selection

The genetic algorithm (GA) was introduced by David Goldberg in 1989 [56]. It is used to find near to optimal solutions in large/exponential search spaces. The heuristic search method mimics natural processes of evolution such as inheritance, mutation, crossover and selection. The GA is known to show good results on a variety of different optimization problems.

The GA starts with a population of candidate solutions of an optimization problem. Through multiple cycles this candidate solutions are improved to get solutions that are close to optimum. During each cycle additional solutions are breed by slightly changing and combining the candidate solutions. The quality of each candidate solution according to specific criteria is determined by the so called fitness function. Using this function the single solutions can be compared to each other and the best solutions form a new population of candidate solutions that serve as starting point for the next cycle. Usually, the algorithm terminates if either the solutions show a satisfying fitness or a fix number of cycles has been executed. For more details refer to Section 2.3.8.

The quality of the solutions found by the GA highly depends on the nature of the fitness function. As feature selection is used to select a subset of genes that is important for a good classification of samples a suitable measurement for the quality of a gene subset is the classification accuracy of a supervised classifier using the subset as input.

For feature selection in microarray data wrapper methods of GAs and different supervised classifiers such as k-nearest neighbor (kNN) [131, 132],

perceptrons [133] or maximum likelihood classifiers (MLHD) [134] usually show good results. Also, a wrapper of a genetic algorithm and a support vector machine (SVM) is not a completely new idea and over the years it has shown good results in various applications [135–141].

This wrapper approach is not commonly used for feature selection and there are no freely available implementations of the algorithms in standard software packages for microarray data analysis such as R [14] or Weka [44] described in Section 2.4. For this reason we use our own implementation of a GA/SVM wrapper described in Section 2.3.8.

During the last years GA/SVM has been mainly used for feature selection in cancer using benchmark data sets [142] for leukemia cancer, colon cancer and lymphoma cancer [139, 143]. Besides the problem of distinguishing cancerous and non cancerous samples in cancer diagnostics we have to differentiate between different types of cancer, for example the 4 different types of leukemia: Acute lymphoblastic leukemia, chronic lymphocytic leukemia, acute myelogenous leukemia and chronic myelogenous leukemia. The GA/SVM also show good results on those multiclass classification problems [141, 144].

Besides cancer classification GA/SVM is also used for other problems. In 2010 Li et al. use a hybrid of GA and SVM for the classification of G-protein coupled receptors (GPCR) [145] and in 2009 Pourbasheer et al. use the algorithm for the prediction of BK-channel activity [146].

Another point of interest is not only finding subsets of features with a high classification accuracy but also removing redundant features and decreasing the size of the feature subset as far as possible. Recent studies show that the GA/SVM algorithm is also well suited for those multi-objective optimization tasks [138, 147].

4.2 Quality of different classification and feature selection methods.

Besides the identification of potential biomarkers for pluripotency and Alzheimer's disease the most important issues of this thesis is the evaluation of the three different feature selection methods, information gain, random forest and a wrapper of genetic algorithm and support vector machine (GA/SVM200). Each of these algorithms results in a list of genes ranked by their importance to the investigated biological state. Looking for potential biomarkers we are interested in genes best suited for distinguishing samples representing this state and samples not representing this state. For this reason we use the

classification capability of the top-ranked genes as quality characteristic for a particular feature selection method.

In our work we analyze two different data sets described in Section 3.2.1. Usually we perform the same analyses on both data sets even though we do not explicitly mention this. The figures and tables in the following sections used to illustrate our results always show the results of the two data sets side by side. We always find the outcomes for the PLURI data set and the left and the outcomes for the AD data set on the right side of the figure.

4.2.1 Classification performance of different classifiers

Before comparing the performance of different feature selection methods with each other we apply multiple commonly used classifiers to our data sets. This gives insights to the general usability of the two data sets for machine learning and feature selection. Additionally we are able to make a well-founded decision for a classifier later used for evaluating the quality of our feature selection methods.

Quality of six classifiers

We use six classification methods, namely C4.5 decision tree, naive Bayes, random forest, k-nearest neighbor and support vector machine (SVM) with Gaussian and linear kernel for a prior investigation of the general usability of the two data sets for machine learning. The methods are described in Section 2.3. We use the algorithms as implemented in Weka [44] and do not tune any parameters besides two exceptions. First we build 1000 trees for classification with random forest to reduce accidental variance. And second we use the LibSVM [54] default parameters ($C = 1$, $\gamma = 1/\#\text{genes}$) for the SVM with Gaussian kernel because the LibSVM implementation is used inside the GA/SVM200 for feature selection as described in Section 2.3.8.

We use the classification accuracy as a measurement for the classification performance of each classifier. The classification accuracy is given by the number of correctly classified samples divided by the total number of samples in the data set, as described in Section 2.3.1. We use a 3-fold cross-validation for calculating the classification accuracy, using the same three folds for all six methods. The partitioning of the data sets is described in Section 3.2.2.

In Table 4.1 we see the accuracy of the six classification methods obtained on the two data sets. On both data sets, the accuracy reached by the six classifiers varies by more than 11%. The classification accuracy on the PLURI data set lies between 81.1% (naive Bayes) and 99.0% (SVM with linear kernel). The lowest accuracy observed on the AD data set is 78.9%

	PLURI	AD
Naive Bayes	87.1%	81.4%
C4.5 decision tree	95.1%	78.9%
Nearest neighbor	96.5%	87.0%
Random Forest	97.2%	87.0%
SVM + Gaussian kernel	97.9%	85.7%
SVM + linear kernel	99.0%	91.9%

Table 4.1: Classification accuracies for the PLURI and the AD data set resulting from classification using six different classifiers. The accuracy is computed by a 3-fold cross-validation.

(C4.5) while the highest classification accuracy is 91.9% reached by the SVM with linear kernel. On both data sets, the two methods with the lowest accuracies are naive Bayes and C4.5, whereas the SVM with linear kernel shows the highest accuracies by far. The quality of the remaining three methods varies depending on the data set.

Also, we observe a very large difference between the two data sets. For each of the six classifiers the accuracy observed on the AD data set is at least 5% lower than the accuracy on the PLURI data set. The mean accuracies over all methods reaches only 85.3% on the AD data set whereas we obtain an accuracy of 95.5% on the PLURI data set.

SVM: Best classifier on both data sets

A classification accuracy of over 80% for most of the classifiers strengthens our decision to use the two data sets for machine learning and feature selection. All analyzed classifiers reached reasonable results. Even so, the SVM with linear kernel show the best results by far. For this reason we use this classifier for the evaluation of the genes top-ranked by our feature selection methods.

We have to avoid the danger of overreaching a single feature selection method by the choice of the classifier. As the SVM is used inside one of our feature selection methods (even if this refers to an SVM with Gaussian kernel) we use two additional classifiers for verifying our results in Section 4.2.2. We use the SVM with Gaussian kernel that is part of one feature selection method and random forest that is part of a second feature selection method. By comparing the results of the three classifiers we are able to detect possible preferences of single feature selection methods by a particular classifier.

We discuss the large differences between the two data sets separately in

Section 4.4.

4.2.2 Classification performance of selected features

In the following we estimate the quality of the genes top-ranked by our three feature selection methods, namely information gain, random forest and a wrapper of genetic algorithm and support vector machine (GA/SVM200). For this we use a number of different classifiers discussed in Section 4.2.1. Besides the classification capability of the top-ranked genes we are also interested in the biological dependencies of those genes. This allows a deeper insight into the correlations between the top-ranked genes. We discuss the obtained results in respect to the functioning of the single feature selection methods.

Quality of the top-ranked genes

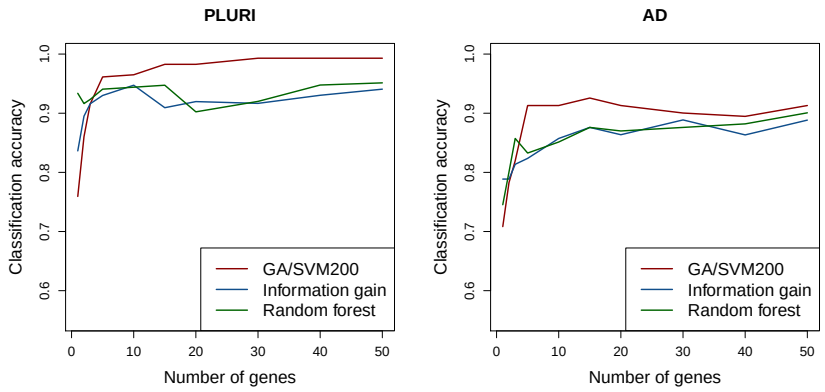
Each of the three feature selection methods results in a list of genes sorted by their importance. In order to compare the quality of the ranking made by the different algorithms, we use the expression values of the selected features for training several classifiers. As comparative value we use the classification accuracy of those classifiers that is calculated as described in Section 2.3.1.

To enable a validation of our results we use a three-fold cross-validation using the same three folds, described in Section 3.2.2, for each feature selection method. On each of the three training sets we run each feature selection method exactly once. The resulted gene lists are validated using the particular training set for training the classifiers and the test set to estimate the classification accuracy. The overall classification accuracy is averaged over all three folds.

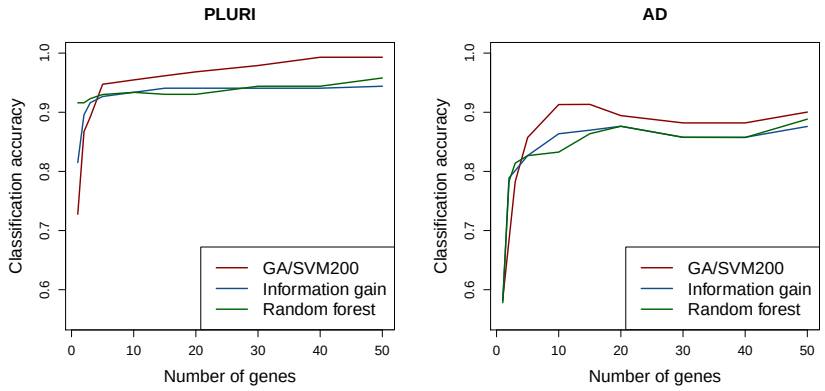
As the feature selection algorithms do not provide any information about the optimal size of the gene subset we use data sets containing incrementally larger gene sets for training the classifiers. We use sets of 1, 2, 3, 5, 10, 15, 20, 30, 40 and 50 genes best ranked by our feature selection algorithms. As classifier we use SVM with linear kernel, as it shows best results for classification on both data sets (see Section 4.2.1). Additionally we use the random forest classifier and an SVM with Gaussian kernel.

In Figure 4.3 we illustrate the classification accuracy of SVM with linear and Gaussian kernel as well as of random forest.

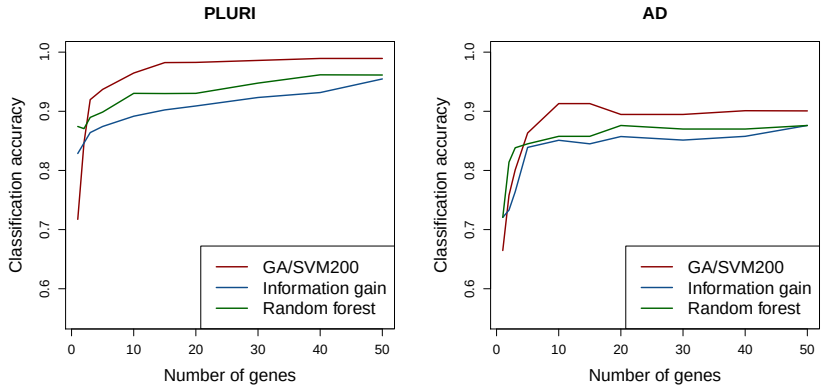
Comparing the three classifiers, we do not find any notable differences in the classification accuracy curves. For this reason following we only describe the results obtained for the SVM with linear kernel in Figure 4.3(a) representative for the other two classifiers.



(a) Classifier: SVM with linear kernel



(b) Classifier: SVM with Gaussian kernel



(c) Classifier: Random forest

Figure 4.3: Classification accuracy of three classifiers using incrementally larger sets of genes for training selected by our three feature selection methods. The accuracy is evaluated by a 3-fold cross-validation.

Independent of the particular feature selection algorithm, we assess that using only few genes for training the classifier results in lower accuracies than using a large number of genes. In general, the highest accuracies are obtained when using at least ten genes for training. Using more than ten genes does not strongly increase the obtained accuracies any more.

The observed classification accuracies for information gain and random forest are nearly identical. Nevertheless, they differ a lot from the accuracies for our GA/SVM200.

Using only single genes for training the genes selected by information gain or random forest show a 4% to 17% higher classification accuracy than those selected by the GA/SVM200. The exact difference depends on the data set as well as on the particular classifier. Using gene sets consisting of at least five genes the genes selected by our GA/SVM200 reach a higher accuracy than those of information gain and random forest. We observe a difference of at least 3% between the maximum accuracy of GA/SVM200 and the maximum accuracy of the other two algorithms.

As for the different classifiers in Section 4.2.1 we again observe a large difference in the absolute classification accuracies obtained on the two data sets. Independent of the used feature selection method the classification accuracy reach on the AD data set is much lower than on the PLURI data set. In average we obtain a difference of about 8%.

Mutual information of the top-ranked genes

To allow a quantitative comparison of the redundancies occurring among the 50 genes top-ranked by information gain, random forest and our GA/SVM200 we use mutual information described in Section 2.1.5.

We run each of our three feature selection methods on the whole set of samples (not divided into training and test set). Using the gene expression values of all samples we calculate the pairwise mutual information for each combination of the 50 genes top-ranked by one of our methods. This way for each feature selection method we obtain a (50×50)-matrix containing the pairwise mutual information of the top 50 ranked genes.

Figure 4.4 shows the density of mutual information for the 50 genes selected by information gain, random forest and our GA/SVM200.

The peak of a probability density function in Figure 4.4 gives the mutual information we most likely observe for two genes of the top 50 gene list. The lower the mutual information of a gene pair the less depends one gene of the pair on the other.

For both data sets the mutual information curve for our GA/SVM peaks at a lower value as the curves for information gain and random forest. On

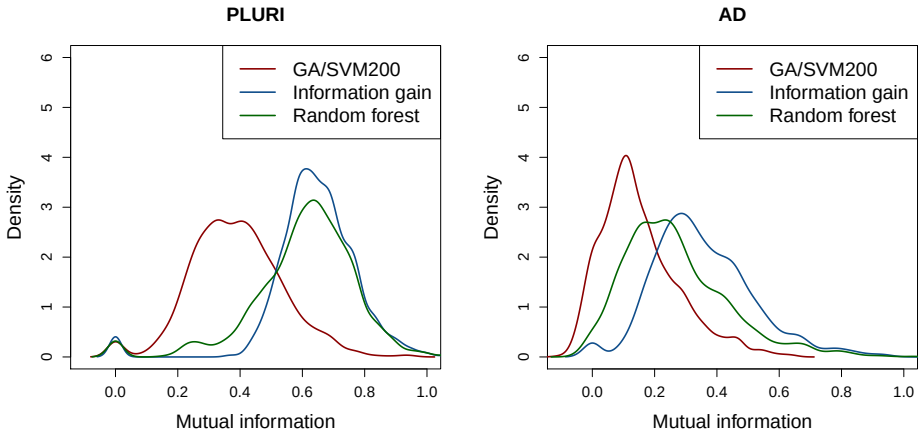


Figure 4.4: Density of mutual information of the 50 genes top-ranked by information gain, random forest and our GA/SVM200.

the AD data set we also observe a difference between information gain and random forest, where the curve for random forest peaks on a lower value than the curve for information gain.

The most noticeable difference between the two data sets is that independent of the particular feature selection methods we observe the peaks on the PLURI data set at higher mutual information than for the AD data set.

Comparison of the three feature selection algorithms

We use three different classifiers to evaluate the quality of the genes selected by information gain, random forest and our GA/SVM. Figure 4.3 shows the results of our analyses.

As we obtain no significant differences in the classification accuracy of the three classifiers we conclude that the classification capability of the selected genes is mostly independent of the classifier used for evaluation. Even if we apply a classifier that is also part of a particular feature selection algorithm the results of this algorithm are not explicitly preferred by the classifier. For this reason we assume that the genes selected by our feature selection methods are usable for distinguishing different samples in the most general sense. For this reason we following refer to the results of the SVM with linear kernel (Figure 4.3(a)) representative for all three classifiers we use.

In the following we discuss similarities and differences between the three feature selection methods information gain, random forest and GA/SVM200, focusing especially on the quality of the selected genes.

First, we assess that the classification capability of the selected genes increases the more of the top-ranked genes we use for training. This observation is independent of the particular feature selection method and the data set. Obviously, the inclusion of more information given by the gene expression level of additional genes improves the classification accuracy of the classifier. We observe saturation at about ten genes. A further increasing of the gene number does not strongly advance the classification accuracy any more. For the AD data set we even find a slight decrease in classification accuracy when using more than 10 genes. We interpret this as evidence for a slight over-fitting of the classifier using too many genes for training. As each gene we include into a set of potential biomarkers also increases the complexity of biological and medical experiments we suggest to use gene sets with no more than ten genes for further analyses.

While the classification accuracies of information gain and random forest are very similar we observe a large difference between these two methods and our GA/SVM200. Using only the top-ranked gene for classification the genes selected by information gain and random forest are better suited than the gene top-ranked by our GA/SVM200. However, combining five or more genes GA/SVM200 outperforms the other two algorithms.

To explain the differences between our GA/SVM200 and the other two methods and enable a better understanding of the obtained results we following expand on the basic principles of the three algorithms.

As described in Section 4.1.3, information gain is a so called filtering approach and by far the fastest of the three algorithms. It deterministically applies an importance score to each gene considering only a Shannon entropy [31] based score of the single gene (see Section 2.3.7). Therefore, the score is not based on a particular classification process and independent of other genes. As the algorithm does not include any information about gene interactions using information gain we are not able to eliminate redundant genes. For this reason, we expect a lot of redundancies under the top-ranked genes. Combining multiple top-ranked genes for classification we only obtain a small gain of accuracy. As combining redundant genes for classification does not necessarily increase the classification accuracy this supports our assumption that the top-ranked genes of information gain contain many redundancies. Even so, the entropy seems to be a suitable measurement for the classification capability of a single gene.

The accuracy curve of random forest is similar to information gain even if the basic principles of the two algorithms are completely different. The random forest belongs to the class of embedded feature selection methods described in Section 4.1.3. We build a random forest classifier by growing multiple decision trees. For each tree we use a subset of samples drawn

randomly with replacement. In each node of the tree we use a very small subset of randomly selected genes and the best suited gene is used for dividing the samples into two subsets. For applying a score to each gene we first use a decision tree to classify all samples not used for growing this tree. Subsequently, we permute the expression values of the particular gene in the classified samples and repeat the classification. The score is given by the difference between the two resulting classification accuracies, and it is averaged over all trees in the forest. So, other than information gain random forest applies an importance score to each gene in dependency of multiple other genes. Additionally, the score depends directly on the classification capability of the gene in the random forest.

In each decision tree a gene that is redundant to a gene already chosen to split the sample set is probably not selected in another node of the tree. So inside a tree redundant genes are eliminated. Growing a random forest in each node we use only a small number of genes to make a decision for splitting the sample set. This way, redundant genes are seldom used in the same tree of the forest, but they probably occur in the different trees. So, we assume that, similar to information gain, redundancies will not be eliminated. This is supported by the similar curve shape of the two algorithms in Figure 4.3. Even if combining the top-ranked genes does not lead to a strong increase of accuracy we already observe high classification accuracy for the top-ranked single gene. Therefore we assume that random forest is well suited for identifying a single biomarker.

The results obtained for our GA/SVM200 differ a lot from those of information gain and random forest. Using only a single gene for classification we obtain a lower accuracy than for information gain and random forest. Increasing the number of genes also strongly increases the classification accuracy of the classifier. So, using five or more genes for classification the genes top-ranked by our GA/SVM outperform the other two methods by far.

The importance score applied to a gene by the wrapper of genetic algorithm and support vector machine is based on 200 small gene sets, each containing genes that are together well suited for separating the samples. The more often a gene is selected in these small sets the higher is the applied importance score. If several genes are redundant and consequently fulfill the same function in the investigated biological state one of those genes is selected randomly whenever the function is included into one of the small sets. This way, low importance scores are applied to genes that have redundant partners. We assume that the genes top-ranked by our GA/SVM200 do not contain many redundancies. This assumption is supported by our results that show a distinct increase of classification accuracy when combining multiple

top-ranked genes.

Our assumption that there are more redundancies among the genes ranked top 50 by information gene and random forest than among the same number of genes identified by our GA/SVM is also supported by the results displayed in Figure 4.4. The figure shows the distribution of pairwise mutual information of the top 50 ranked genes for information gain, random forest and our GA/SVM200. On average, the mutual information obtained for information gain or random forest is higher than the mutual information for our GA/SVM200. As described in Section 2.1.5 the mutual information of two genes is usable as measurement of the contained redundancies. Large mutual information of two genes implies a large dependency of these two genes. More exact, the larger the mutual information the more similar information are carried by the two genes and the larger is the redundancy of these genes.

Compared to information gain and random forest our GA/SVM200 is the only algorithm that eliminates redundancies among the top-ranked genes. As regulatory biological processes usually are very complex and single genes are barely able to explain the whole process the selection of good genes sets is a promising advantage of the GA/SVM and not yet well investigated. As the elimination of redundant genes mainly depends on the structure of the algorithm and therewith on the small sets selected in the single runs of the GA/SVM in the next sections we will focus on the analysis those small sets.

As the differences in absolute classification accuracy as well as in the averaged mutual information are not the only differences we observe between the two data sets we discuss these observations separately in Section 4.4.

4.3 The true potential of our GA/SVM

The feature selection methods described in Section 4.1.3 typical result in a ranked list of genes. The top-ranked genes can be used as biomarkers to extend the understanding of the molecular mechanisms of the examined processes. Usually these processes are very complex so that a single biomarker cannot explain the whole process. For this reason we are interested in biomarker combinations well suited for explaining the investigated biological state.

In Section 4.2.2 we show that combinations of genes top-ranked by our wrapper of genetic algorithm and support vector machine (GA/SVM200) are better suited for classification than the same number of genes top-ranked by information gain or random forest. We trace this to the elimination of redundancies in the list of top-ranked genes that is founded in the small gene sets selected by our GA/SVM in each of the 200 single runs. In the following

the term small set explicitly refers to those small gene sets selected in a single run. In order to make the results of our GA/SVM comparable to information gain and random forest we mostly ignore the specific structure of these small sets. Nevertheless, we assume that those small sets are promising biomarker sets containing genes working together very well for distinguishing different samples. To the best of our knowledge this aspect of the GA/SVM is not yet investigated in-depth.

Again we perform our analysis on two data sets describe in Section 3.2.1. We show the outcomes for both data sets beside each other (the PLURI data set on the left side, the AD data set set on the right side).

In the following we examine the small sets in the PLURI data set as well as in the AD data set. We compare their classification capability to the quality of those genes top-ranked by the GA/SVM200. Further we examine particular gene pairs that occur in the small sets more- often or less often than expected.

4.3.1 Classification capability of small biomarker sets

In its basic form the GA/SVM results in a small set of genes. Based on the design of the algorithm, detailed described in Section 2.3.8, we assume that such a small set consists of genes together very well suited for classification. Also, we assume that the small sets are as small as possible which means that they do not contain any redundant genes.

In the following, we compare the classification capability of the small gene sets selected by our GA/SVM to the list of genes ranked by the GA/SVM200 used for feature selection.

Quality of small gene sets

Similar to the procedure in Section 4.2.2, we use a 3-fold cross-validation for analyzing the quality of the small gene sets. The partitioning of the PLURI data set and the AD data set is described in Section 3.2.2.

First we run the core algorithm of the GA/SVM 200 times using two of the three subsets for feature selection. This way, for each fold we obtain 200 small gene sets of variable size as well as a list of genes ranked by our GA/SVM200.

For each of the resulting small sets we determine its classification accuracy using a support vector machine (SVM) with Gaussian kernel. The SVM is trained on the same fold on which the small set is selected using the genes of the small set for training. The accuracy of a small set is then given by the obtained accuracy of the SVM on the remaining test set.

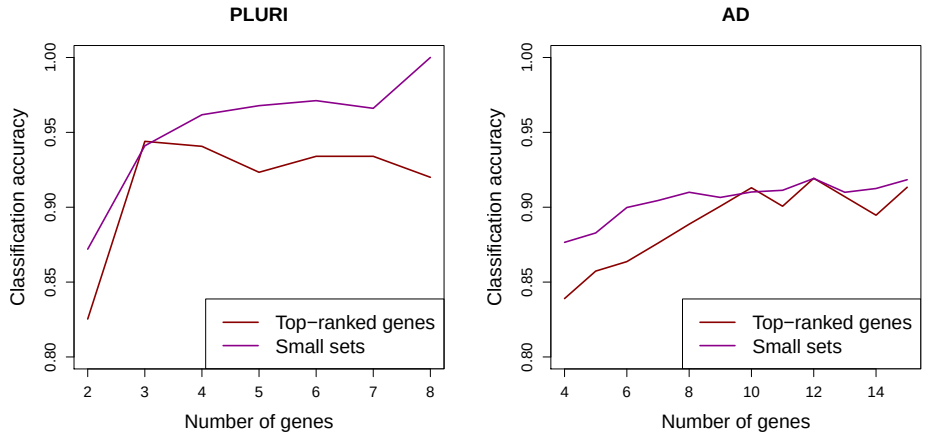


Figure 4.5: Classification accuracy of an SVM with Gaussian kernel using incrementally larger gene sets for training. The gene sets are obtained either from our GA/SVM during a single run or from the list of gene top-ranked by our GA/SVM200. The classification accuracy is computed by a 3-fold cross-validation.

In each fold we average the classification accuracies of all small sets of the same size. To obtain the overall accuracy for the small sets of a specific size we average the results over all three folds.

To enable a comparison to the ranked list of genes we use the same number of top-ranked genes as input for an SVM with Gaussian kernel and calculate the mean classification accuracy over all three folds.

In Figure 4.5 we compare the classification accuracy of the top-ranked genes to the accuracy of small sets of the same size.

We clearly see that the classification accuracy of the small sets is at least as high (in most cases much higher) than the classification accuracy of the same number of genes top-ranked by our GA/SVM200.

Again we find some differences between the two data sets. First, the classification accuracy is always lower on the AD data set then on the PLURI data set. Additionally the size of the small sets varies between the two data sets. For the PLURI data set the small sets consist of 2 to 8 genes, whereas the small sets selected on the AD data set contain between 4 and 15 genes. On the PLURI data set we observe a large difference in accuracies between using 2 and using 3 genes for classification (especially for the genes top-ranked by GA/SVM200, but also for the small sets). In contrast, on the AD data set the accuracy slowly but steadily increases and reaches a saturation using approximately 10 genes for classification.

Comparing small gene sets to the top-ranked genes

Within the GA/SVM the selection of a small set is determined by a multi-objective optimization function that mainly favours small gene sets with a large classification accuracy regarding an SVM with Gaussian kernel. For this reason we assume that the small sets are very well suited for distinguishing different samples. This is supported by our results shown in Figure 4.5 where the trained classifier shows a distinct higher accuracy using the small gene sets for training than using the same number of genes top-ranked by our GA/SVM200.

The second objective of the optimization function within the GA/SVM is minimizing the size of the small sets. As we weight this objective much lower than the classification accuracy we favour a small set only if the classification accuracy of this set is at least the same as the classification accuracy of a larger small set. This way, we suppose to force the GA/SVM to eliminate all redundant genes within a single small set.

As discussed in Section 4.2.2 combining the genes top-ranked by our GA/SVM200 increases the accuracy of multiple classifiers and outperforms assemblies of genes top-ranked by information gain or random forest. Even so, the small sets selected during single runs of our GA/SVM outperform sets of genes top-ranked by the GA/SVM200. This supports our assumption that the small gene sets contain only few redundant genes. Even for very small gene sets (containing 2 genes in case of pluripotency and 4 genes in case of Alzheimer) we obtain an averaged classification accuracy of more than 87% on both data sets. This indicates that the small sets in general are very well suited for classifcation.

We assume that the assembly of genes in a single small set is essential for its classification capacity. A small set well suited for distinguishing different groups of samples contains genes that play an important role in the investigated biological state. As the small set does not contain redundancies the genes cover multiple aspects of the biological state. This way, small sets are well suited for explaining multiple parts of the underlying molecular processes.

This results show the real strength of the algorithm. Besides the popular approach using the GA/SVM200 for the identification of potential biomarkers by ranking the genes by their frequency of occurrence in the single small sets the small sets itself are very well suited as potential biomarker sets.

Compared to greedy wrapper approaches (described in Section 4.1.3) that also results in a gene subset of optimal size the small gene sets selected by our GA/SVM differ strongly from run to run. This way, we find many close to optima small sets when running the GA/SVM multiple times. From

a biological point of view we expect to find more than one set of genes explaining the biological state. Most biological processes are regulated by many different genes and the change in the expression of a single gene may already change the biological state of a cell or tissue. For example, it is shown that multiple gene combinations can be used for reprogramming somatic into pluripotent cells (described in Section 5.1.3).

Even if, the biological evaluation of our results is challenging small sets offer promising opportunities. An interesting attempt for further research would be the investigation of small gene sets with regard to known relationships between the single genes in the set. This way we could identify functional similarities and verify the usability of the small sets. We suggest that the information contained in the large number of different small sets considering the same biological state is usable for modelling regulatory networks.

Notable, even if the small sets of multiple runs differ from each other, running the GA/SVM multiple times and combining the genes contained in the small sets to a ranked list as described in Section 2.3.8, the ranking is nearly perfect reproducible. So, some genes are always selected more frequently than others.

To enable a deeper insight to the structure of the small sets selected by our GA/SVM we following investigate pairs of genes occurring together in the small sets more often or less often than expected.

The differences between the two data sets we discuss in detail in Section 4.4 along with other observed differences between the data sets.

4.3.2 Analysis of gene pairs in small sets

As we discuss in the last section we consider the small gene sets selected by the GA/SVM during a single run to be the true strength of this algorithm. As the biological evaluation of our small sets is expensive we following focus on gene pairs occurring in the small sets. This gives a deeper insight into the structure of the small sets as well as a better understanding of the quality of our GA/SVM. Additionally we are able to nominate particular gene pairs interesting for a biological examination. We start this section defining some specific terms used for the analysis of the gene pairs.

Specific definitions for the analysis of gene pairs

In the following we define some specific terms used for our analysis. The terms refer to the small gene sets selected by our GA/SVM during single runs.

Gene occurrence. The occurrence of gene i in a small set S is defined as

$$geneOcc_i(S) = \begin{cases} 1, & \text{if } g_i \in S, \\ 0, & \text{otherwise.} \end{cases}$$

In a set of n small sets $U = \{S_1, S_2, \ldots, S_n\}$, $geneOcc_i(U)$ is given by

$$geneOcc_i(U) = \sum_{k=1}^{n} geneOcc_i(S_k).$$

Joint occurrence. We further analyze gene pairs occurring jointly in a small set. Similar to the gene occurrence *geneOcc*, the joint occurrence of gene i and gene j in a small set S is determined as

$$jointOcc_{i,j}(S) = \begin{cases} 1, & \text{if } g_i \in S \text{ and } g_j \in S, \\ 0, & \text{otherwise.} \end{cases}$$

Further, the joint occurrence of two genes i and j in a set of n small sets $U = \{S_1, S_2, \ldots, S_n\}$ can be calculated as

$$jointOcc_{i,j}(U) = \sum_{k=1}^{n} jointOcc_{i,j}(S_k).$$

Expected joint occurrence. Assuming independence of the gene occurrences $geneOcc_i$ and $geneOcc_j$, the expected value for the joint occurrence of the genes i and j is defined as

$$E\left[jointOcc_{i,j}(U)\right] = \frac{geneOcc_i(U) \cdot geneOcc_j(U)}{n},$$

where n is the number of small sets in $U = S_1, \ldots, S_n$.

Importance of a joint occurrence. To enable a ranking of the gene pairs based on the relevance for our analysis we define the importance of a joint occurrence of two genes i and j as

$$score_{i,j}(U) = \log\left(\frac{jointOcc_{i,j}(U)}{E\left[jointOcc_{i,j}(U)\right]}\right),$$

where $U = \{S_1, S_2, \ldots, S_n\}$ is the set of small sets. In the following we refer to a gene pair (i, j) as over-represented in U if $score_{i,j}(U) > 0$. Analog, we refer to the gene pair as under-represented if $score_{i,j}(U) \leq 0$. Consequently, the gene pair with the largest value for $score_{i,j}(U)$ is the gene pair most over-represented in U, and the lowest value for $score_{i,j}(U)$ is applied to the most under-represented gene pair.

Analysis of the most over- and most under-represented gene pairs

We run the GA/SVM 3,000 times using the whole 1,000 gene containing data set. This way we achieve a set of $n = 3000$ small sets. For each gene i and each gene j we compute the joint occurrences $jointOcc_{i,j}$ and the expected joint occurrence $E\left[jointOcc_{i,j}\right]$ as described in the previous section.

We are only interested in pairs of genes where the occurrence of one gene in a small set influences the occurrence of the other gene in the same set. This means we are interested in gene pairs whose joint occurrence in the small sets is statistically independent. We use the χ^2-test described in Section 2.1.3 to determine the statistical dependencies for each gene pair. As one of the χ^2-test requirements $E\left[jointOcc_{i,j}\right]$ has to be larger than five. For this reason we dismiss all gene pairs that do not fulfill this criterion. Subsequently, for the correction of the p-values we use false discovery rate correction with a cutoff of 0.05. This method is described in Section 2.1.4.

For all gene pairs consisting of genes that depend on each other we compute the importance of the joint occurrence $score_{i,j}$ (described in the previous section) and split the gene pairs into over- and under-represented pairs. We sort the over-represented gene pairs by their importance $score_{i,j}$ in descending order starting with the most over-represented gene pair. Analog, we sort the under-represented gene pairs in ascending order starting with the most under-represented gene pair.

For each over- and under-represented gene pair (i, j) we determine three reference scores based on the classification accuracy of a support vector machine (SVM) with Gaussian kernel. In the most general sense we compute the accuracy $SVMacc$ using a single gene or a gene pair for performing a 10-fold cross-validation on the whole sample set as described in Section 2.3.1. The three reference scores are defined as follows.

- Classification accuracy:
$$SVMacc_{i,j}\,.$$
- Mean gain of accuracy:
$$\text{SVMgainMean}_{i,j} = SVMacc_{i,j} - \frac{SVMacc_i + SVMacc_j}{2}.$$
.
- Minmal gain of accuracy:
$$SVMgainMin_{i,j} = SVMacc_{i,j} - \max(SVMacc_i, SVMacc_j).$$
.

The gain of accuracy describes the difference in classification accuracy between using two genes and using only one of those genes for classification. Whereas for $SVMgainMin_{i,j}$ we use the gene with the higher accuracy for $\text{SVMgainMean}_{i,j}$ we consider the mean accuracy of both genes of the pair.

Figure 4.6 displays the classification accuracy (4.6(a)), the mean gain of accuracy (4.6(b)) as well as the minimal gain of accuracy (4.6(c)) of the most over- and the under-represented gene pairs. More precisely we display the reference accuracies averaged over incrementally larger sets of the 3, 6, 9 ..., 75 most over-represented gene pairs. We proceed similarly for incrementally larger sets of the 3, 6, 9 ..., 75 most under-represented gene pairs. For each reference measurement we show the results for the most over- and under-represented gene pairs in the same chart.

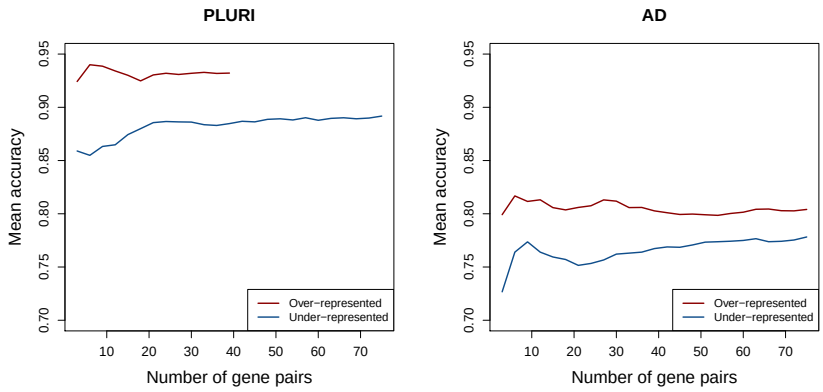
Note that we average the reference accuracies over multiple gene pairs to avoid random variations. The step-width of 3 is chosen arbitrarily but ensures we still have enough data points for an analysis. For the PLURI data set we only find 38 significantly over-represented gene pairs. For this reason the curves displaying the most over-represented gene pairs in the charts for the PLURI data set are much shorter than the curves for the most under-represented gene pairs.

Figure 4.6(a) shows the mean classification accuracy of the most over- and under-represented gene pairs. We find a large difference between the over- and under-represented gene pairs. The accuracy of the gene pairs occurring in the small sets more often than expected is always higher than the accuracy of the under-represented gene pairs. Further, as we also observe in other analysis we again find a large difference in accuracy between the PLURI data set and the AD data set.

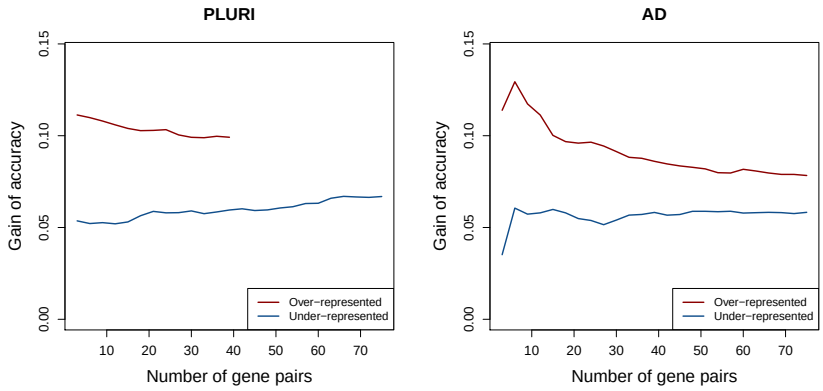
Figure 4.6(b) and Figure 4.6(c) display the mean and minimal gain of classification accuracy for the most over- and under-represented gene pairs. Again we observe a large difference between the over- and the under-represented pairs, finding a higher gain of accuracy for the over-represented gene pairs than for the pairs occurring in the small sets less often than expected.

For the over-represented gene pairs we see a slight decrease of accuracy gain the more gene pairs we add to determine the average gain of accuracy. This implicates, that on average with an decrease of the importance of the joint occurrence the classification accuracy also decreases. Similar, we observe a slight increase of accuracy the more of the under-represented gene pairs we use for calculating the average gain of accuracy. So, on average the larger the important score of an under-represented gene pair the higher is its classification capability.

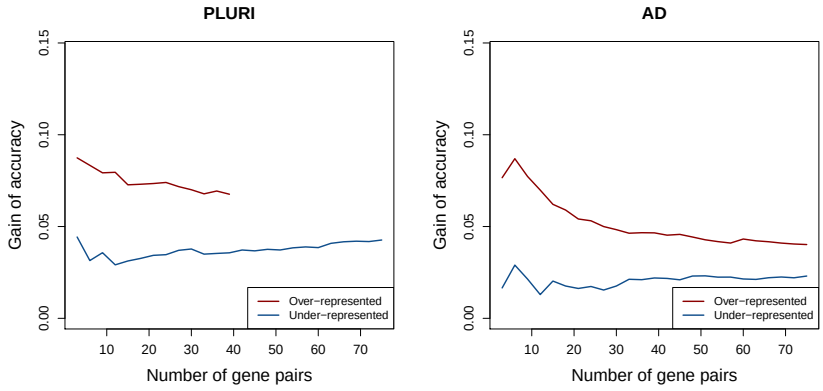
We illustrate the gain of accuracy obtained from combining two genes for classification in Figure 4.7. For each data set we show the gene expres-



(a) SVM accuracy



(b) Mean gain of accuracy



(c) Minimal gain of accuracy

Figure 4.6: Gain of accuracy of the most over- and under-represented gene pairs found by the genetic algorithm. The accuracy gain is calculated by an SVM with Gaussian kernel.

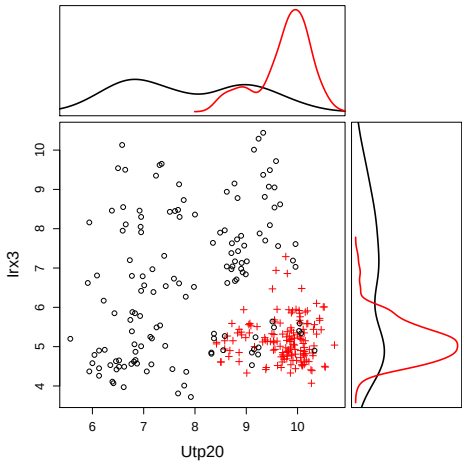
gene	accuracy	data set
Utp20	0.81	PLURI
Irx3	0.77	PLURI
Otx2	0.83	PLURI
Gbx2	0.85	PLURI
SLC39A12	0.70	AD
LAP3	0.63	AD
GEM	0.69	AD

gene pair	accuracy	acc. gain (mean)	acc. gain (min)
Utp20/Irx3 (↑)	0.92	0.13	0.11
Otx2/Gbx2 (↓)	0.88	0.04	0.03
SLC39A12/LAP3 (↑)	0.80	0.14	0.10
SLC39A12/GEM (↓)	0.71	0.02	0.01

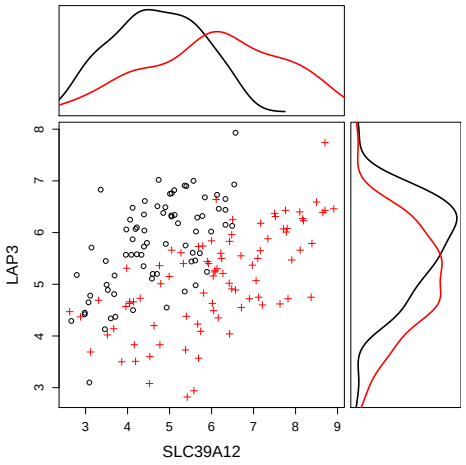
Table 4.2: SVM accuracy of the genes contained in the most over- and the most under-represented gene pair of each data set (top). SVM accuracy as well as the mean and minimal gain of accuracy of the most over- (↑) and the most under-represented (↓) gene pair of each data set (bottom).

sion pattern of the most over- and the most under-represented gene pair. These are also the gene pairs with the highest and the lowest mean gain of accuracy. The scatterplots show the distribution of the samples in a two-dimensional space. To get a vision of the classification power of the two genes we differentiate between positive and negative samples (pluripotent and the non-pluripotent, Alzheimer affected and non-affected). For each gene of the pair we add a density chart showing the distribution of positive and negative samples for the single gene. Further, to support the charts, in Table 4.2 we list the SVM classification accuracy and the gain of accuracy for all gene pairs and single genes shown in Figure 4.7.

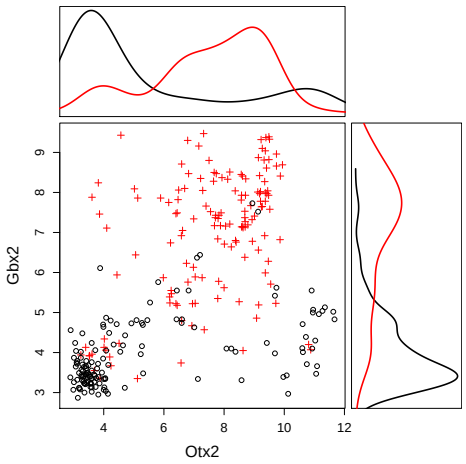
Looking at the density charts of the single genes we find an obviously smaller overlap of positive and negative samples in the PLURI data set than in the AD data set. Also the overlap of the samples in the two dimensional space is much smaller for the PLURI data set. Comparing the separability of samples in the two dimensional space given by the most over- and the most under-represented gene pair of each data set we obtain a higher classification capability of the most over- than for the most under-represented gene pair.



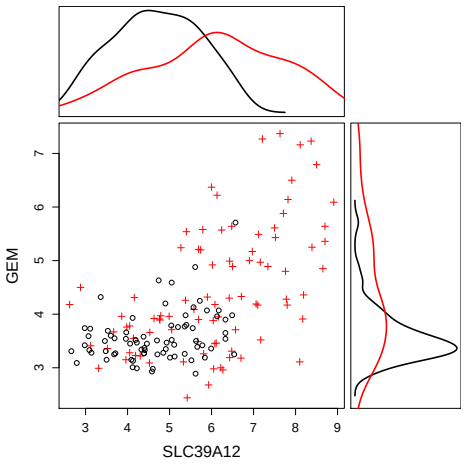
(a) Most over-represented gene pair of the PLURI data set



(b) Most over-represented gene pair of the AD data set



(c) Most under-represented gene pair of the PLURI data set



(d) Most under-represented gene pair of the AD data set

Figure 4.7: Gene expression plot of the most over- and most under-represented gene pair of each data set. The density diagrams show the distribution of the samples of the two classes for single genes. The scatter plots show the distribution of the samples for the gene pairs. (red: pluripotent/Alzheimer affected samples, black: non-pluripotent/non-affected samples)

Gene pairs preferred by our GA/SVM

Our results in Figure 4.6(a) show, that gene pairs often selected together in a small set are well suited for classification. In contrast the classification capability of gene pairs rarely occurring in the small sets is relative low. We assume that the GA/SVM preferably combines genes that jointly show a high classification capability.

Nevertheless, combining two genes individually good for classification it is more likely to obtain large classification accuracy than by combining two genes with a low classification accuracy. For this reason the results in Figure 4.6(a) could also be explained by the preference of genes individually good for classification and not specifically in combination. Under this assumption the small sets would still be usable as close to optima biomarker sets discussed in Section 4.3.1. Even so, under this condition we would expect many redundant genes in a small set.

For this reason, besides the absolute classification accuracy of the over- and under-represented gene pairs we investigate the gain of accuracy obtained for those gene pairs in relation to the included single gene accuracies. In Figure 4.6(b) and Figure 4.6(c) we find a much higher gain of accuracy for gene pairs preferred by the GA/SVM than for those gene pairs discriminated by the GA/SVM. We conclude that using the GA/SVM we find close to optima small sets of potential biomarkers that are together much better suited as indicator for a specific biological state than the individual biomarkers. The large gain of accuracy observed for the most over-represented genes supports the assumption that the small sets selected by our GA/SVM only contain few redundancies.

From a biological point of view under-represented gene pairs contain two genes with a close relation. Among others, this could be a co-regulation or co-expression, a direct or indirect correlation or a functional consensus of the two genes. In contrast, based on our results we assume over-represented genes to fulfill different biological functions that are both essential for the examined processes.

Analysing gene pairs instead of small gene sets of variable size that are usually contains much more genes reduces the complexity of future biological experiments. This enables a validation of our results in a biological context. Nevertheless, to understand all mechanisms the genes of a small set are involved in, future, we have to consider combinations of three or more genes and finally of all genes contained in a small set.

In the following section we illustrate the usability of the most over- and most under-represented gene pairs for separating samples of different classes.

The differences between the two data sets are discussed in detail in Section

4.4.

The most over- and most under-represented gene pairs

Figure 4.7 visualize the most over- and the most under-represented gene pair of each data set.

For the PLURI data set the most over-represented gene pair is *Utp20/Irx3* displayed in Figure 4.7(a). Looking at the density distribution of the single genes we only find a small overlap between the positive and the negative samples. This implies a high classification capability of the individual genes. In fact, in Table 4.2 we find the accuracy for using the single genes for classification is 81% for *Utp20* and 77% for *Irx3* combining the two genes for classification leads to an accuracy of even 92%. Giving a rule for separating the samples in the two-dimensional space we suggest to classify a sample as pluripotent if the gene expression value of $Utp20 > 8$ and $Irx3 \leq 6$. Using such an easy rule we classify 262 samples correctly (out of 286) and reach the obtained accuracy of 92%.

Figure 4.7(c) shows the gene pair *Otx2/Gbx2* most under-represented in the small sets. Again we observe only a small overlap regarding the single genes, but the overlap of different sample types in two-dimensional space is much larger then for *Utp20/Irx3*. As we see in the two-dimensional chart, a rule for separating the two classes is difficult to determine and many samples will probably be incorrectly classified. This results in a low classification accuracy of only 84%.

In Figure 4.7(b) we find *SLC39A12/LAP3* the preferred gene pair of our GA/SVM on the AD data set. Compared to the PLURI data set the overlap between the positive and negative samples regarding the single genes is much higher leading to much lower classification accuracies that is 70% for *SLC39A12* and 63% *LAP*. Even so, combining the two genes improves the classification accuracy to 80%. Again we give a simple rule for classifying a sample as Alzheimer's disease affected if the expression value for *LAP3* is smaller than the value for *SLC39A12*. This way we classify 133 out of 161 samples correctly reaching with 83% nearly the same accuracy as calculated by the SVM.

SLC39A12/GEM is the most under-represented gene pair in the small sets of the AD data set (Figure 4.7(d)). Compared to the PLURI data set we again observe a large overlap of the two sample classes regarding the single genes. Similar to the most under-represented gene pair for the PLURI data set we still observe a large overlap between positive and negative samples considering the two-dimensional space. This results in a relative low accuracy of only 71%.

Considering our results from a biological point of view we assume that the genes of the pairs most over-represent in the small sets both have important functions concerning the investigated biological state. Even so, we suppose that the genes are regulated differently, fulfill different task or take part in different regulatory networks. In contrast, the most under-represented gene pairs are assumed to have much more in common. Combining the two genes the gain of classification capability is very low. For this reason we assume that one gene of the pair explains more or less the same aspects of the investigated biological process as the other gene. Possible reasons for this could be a co-expression of the two genes, a mutual dependency in a regulatory network or the involvement in the same signaling pathway. Note, that even if those gene pairs are rarely chosen, the single genes of the pairs occurring among the top-ranked genes of our GA/SVM200. So, we assume that the single genes still play an important role in the particular biological state.

4.4 Comparison of our data sets

In this chapter we analyze the performance of different classifiers and feature selection methods. Supported by our results we also analyze small gene sets selected by a wrapper of genetic algorithm and support vector machine (GA/SVM). For the analyses we use two microarray data sets referred to as PLURI data set and AD data set as described in Section 3.2.1. Besides a comparison between different methods and an evaluation of the quality of the selected genes we compare the two data sets with each other. More explicitly we are able to compare differences in the performance of equal algorithms obtained on the two data sets.

In Section 4.2 and Section 4.3 we describe a number of differences between the two data sets following discussed.

In Section 5.5.2 we discuss some more differences between the data set from a more biological view concerning our results in Chapter 5.

Absolute classification accuracy

The obtained absolute classification accuracies are the most obvious difference between the two data sets. In Table 4.1 we list the classification accuracies of six classifiers using a 3-fold cross-validation using all 1,000 genes for training. Without exception, all classifiers show better results on the PLURI data set than on the AD data set. The reached classification accuracy is at least 5% higher on the PLURI data set.

In Figure 4.3 we display the classification accuracies of three different classifiers using the top-ranked genes for training. Again the observed accuracy is at least 5% lower on the AD data set. Also, the classification capability of the small sets selected by our GA/SVM (Figure 4.5) and of the containing gene pairs (Figure 4.6(a)) are lower using the AD data set for our analyses than using the PLURI data set. The large difference in classification accuracy is also visible in the two dimensional arrays of Figure 4.7. We see a much larger overlap for the AD data set than for the PLURI data set considering the sample distribution of the single genes as well as the distribution in the two-dimensional space given by the gene pairs.

As we analyze the two data sets in exactly the same way there has to be a difference between the data explaining the different classification capabilities.

The most obvious difference is the size of the two data sets. Both sets contain the expression values of 1,000 genes, even so, whereas the PLURI data set contains 286 samples the AD data set only consists of 161 samples. On the PLURI data set the algorithms include more samples and therewith information in the processes of classification and feature selection. We suppose analyzing larger data sets concerning Alzheimer's disease may result in better classification accuracies.

The second difference is the investigated biological state. Of course, the concrete regulatory processes involved in pluripotency differ a lot of those involved in Alzheimer's disease. For this reason we assume that the genes involved in Alzheimer are detectable much more difficultly than genes involved in pluripotency. The difference in the number of genes involved in the regulatory processes could be one possible explanation. The more genes are involved the easier they are detectable.

Note that we also cannot preclude noise and other factors impair the AD data set leading to a worse performance of our algorithms compared to the PLURI data set.

Redundancies in the top-ranked genes

Besides the classification accuracy we observe a difference in the mutual information contained in the top-ranked genes of the three feature selection methods. In Figure 4.4 we see the distribution of the mutual information of the to 50 ranked genes. Whereas using the PLURI data set the probability density functions peak at a mutual information of about 0.3 (GA/SVM200) and 0.6 (information gain and random forest), the functions peaks at about 0.1 (GA/SVM200), 0.2 (random forest) and 0.3 (information gain) using the AD data set. Larger mutual information means a larger dependency between the top-ranked genes and therewith more redundancies. So, we assume, that

independent of the feature selection methods the top 50 ranked genes contain more redundancies for the PLURI data set than for the AD data set.

We assume that more redundancies are found for the PLURI data set because more genes are involved in the biological state of pluripotency than in Alzheimer. This supports that the complexity of the underlying molecular mechanisms of pluripotency and Alzheimer's disease differs a lot.

Size of the small sets

A third issue that supports our assumption of the different complexity of the regulatory networks responsible for pluripotency and Alzheimer is the observed size of the small sets. Whereas the size of the small sets selected in the AD data set varies between 4 and 15 (9 genes on average) the small sets selected on the PLURI data set are much smaller. There, the small sets vary between 2 and 8 with an average size of 4 genes. For the sets with the smallest size we obtain the same classification accuracy (87%) on both data sets. Even so, the smallest gene sets for the PLURI data set contain 2 genes whereas the smallest gene sets selected on the AD data set contain 4 genes. We assume that we need much more genes to be able to distinguish Alzheimer affected and non-affected samples than classifying pluripotent and non-pluripotent ones.

Chapter 5

Promising biomarker candidates

From a molecular biological point of view, there are much more differences between pluripotent and Alzheimer's disease affected tissues than similarities. While pluripotent cells play an important role in the embryonic development and may have great potential for therapeutic use in stem cell therapy [7, 8] Alzheimer is a disease usually diagnosed with patients older than 65. The neurodegenerative disease of the brain is, at this time, not efficient treatable and supposed to cause large substantial costs in the future [9, 10].

Despite all differences, both biological states is united by the lack of knowledge about the underlying molecular mechanisms. Especially the identification of new genes involved in the biological states are able to give new starting points for further research.

In this work, besides evaluating different feature selection methods extensive discussed in Chapter 4, we identify genes well suited as potential new biomarkers for pluripotency and Alzheimer's disease. For this, we use three feature selection methods, namely information gain, random forest and a wrapper of genetic algorithm and support vector machine (GA/SVM) showing good results in our analysis. In Chapter 4 we show that GA/SVM outperforms information gain and random forest in respect to the classification capability of the selected gene set. For this reason in our discussion we mainly focus on biomarkers selected by our GA/SVM.

For feature selection we run the three algorithms on the whole data sets without spare samples for an independent testing. This way we are able to use maximum information for selecting potential biomarkers. In this chapter we validate our results from a biological point of view using gene set enrichment analysis and discuss the selected genes with respect to recent publications.

We split the following chapter into three parts. First we give an intro-

duction to pluripotency and present our selected biomarkers. We discuss the relevance of our results with respect to current research in the field of pluripotency. In the second part of this chapter we focus on Alzheimer's diseases. Similar to pluripotency we start with an introduction to Alzheimer and then present our results with respect to current research. In the last part we compare the two data sets with each other and discuss some problems we observe in microarray data analysis.

5.1 Pluripotency

Pluripotency is derived from the Latin words for 'very many' (plurima) and 'power' (potentia). In cell biology pluripotent cells are defined by the potential to differentiate into any of the three germ layers: endoderm, mesoderm and ectoderm.

The only natural incidence of pluripotent cells are the cells of the inner cells mass of the blastocyst from which the embryonic stem (ES) cells are extracted. Besides pluripotency, ES cells have the possibility to go through numerous cycles of cell division without losing their undifferentiated states. As the processes in ES cells are essential for the individual development researchers need to completely understand the complex underlying mechanisms. Besides, the pluripotent state and the potential for self-renewal of ES cells offer interesting possibilities for regenerative medicine.

Using human ES cells for research and medicine is a controversial discussed subject [148]. On the one hand scientists expect to achieve a high benefit for the medical treatment of yet incurable diseases. On the other hand extraction of human ES cells requires a destruction of human embryos at an early phase of development. Stem cell therapies using adult stem cells (ASC) are already used for treating specific types of cancer but using ES cells could offer more flexibility.

First successes in inducing pluripotency in somatic non-pluripotent cells show the value of the current knowledge. Expanded research on this topic and a better understanding of ES cells may supersede the use of natural human ES cells for further medical stem cell therapies.

5.1.1 Embryonic and adult stem cells

Naturally, pluripotent cells occur only as (ES) cells. The development of an organism, illustrated in Figure 5.1, starts with a fertilized ovum, a single totipotent cell. This kind of cell has the potential to differentiate into any tissue, even extra-embryonic tissue that is essential for the development of

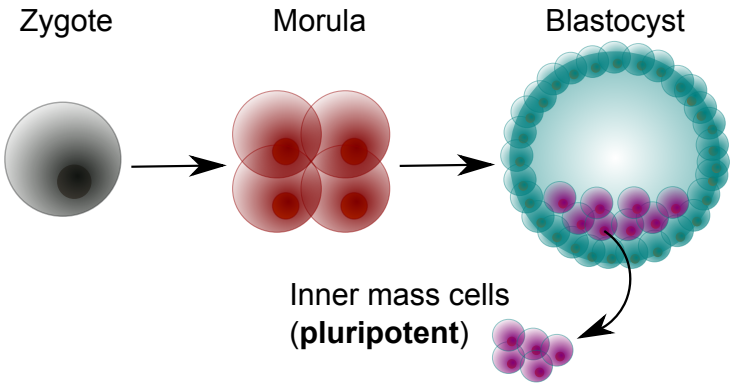


Figure 5.1: Embryonic development from the zygote to the blastocyst state at which the pluripotent cells are extracted.

an organism. In the first hours after fertilization the cell divides into 16 totipotent cells that form the morula. The cells of the morula differentiate either into the pluripotent cells of the blastocyst inner cell mass or in cells of the trophoblast, that later forms large part of the placenta. The cells of the inner cell mass pass through multiple division cycles keeping their pluripotent character until they start specializing after three to four days. Then they start differentiating into cells of the different germ layers and form various different tissues.

ES cells can be extracted during a short time frame in the stage of the blastocyst from the inner cell mass. In vitro, the so gained ES cells can be multiplied any number of times without loss of pluripotency. Further they can specifically differentiated into several cell types such as neural or blood cells.

After embryonic development there are still undifferentiated cells occurring in the organisms. These cells are called adult stem (AS) cells. Often there is no precise distinction between adult end embryonic stem cells. Even so the differences are crucially. Like ES cells, AS cells also have the ability to go through numerous circles of cell division without losing their undifferentiated states. In contrast to ES cells, AS cells can differentiate only into cells of a specific organ. For example hematopoietic stem cells that can be found in the bone marrow are able to differentiate into different types of blood cells. In vivo, there are two types of AS cell divisions. The symmetrical division breed two identical daughter cells, both multipotent and with the potential for self-renewal. The asymmetrical division gives rise of a cell with all characteristics of the AS cell and a progenitor cell with only limited self-renewal potential. The progenitor cell can still go through various cell cycles before it is finally differentiated and form a new cell of the specific tissue.

5.1.2 Molecular mechanisms of pluripotency

Although, we already know a lot about embryonic stem cells and pluripotency, there is still a need for a more precise investigation of various aspects. The core proteins inducing pluripotency are well defined but the exact regulation mechanisms of these proteins have to be particularly specified. In recent studies the knowledge about pluripotent cells was already used for inducing pluripotency in somatic non-pluripotent cells [149–151]. Even so, the high tumor genesis of those induced pluripotent cells is still challenging.

Oct4, Sox2 and Nanog - The key proteins in pluripotency

Meanwhile it is generally accepted that OCT4 and NANOG are the key proteins for the maintenance of pluripotency [152–156]. As it is an important interactive partner for both proteins often SOX2 is added to these key proteins [157]. All three proteins are involved in the embryonic development and the determination of the cell fate.

OCT4 (octamer-binding transcription factor 4) is encoded by the gene *Pou5f1*. In adult organism it is only expressed in germ cells. The role of OCT4 for the self-renewal of undifferentiated embryonic stem (ES) cells was discovered in 2002 by Hans Schöler et al. [158] In pluripotent cells the expression level of the gene *Pou5f1* is kept within certain limits. Too much as well as too little of the protein cause the differentiation of the cell [159]. It is also shown that OCT4 is involved in the tumor genesis of adult human germ cell tumors [160].

SOX2 (sex determining region Y-box 2) is a transcription factor that is assumed to regulate the expression of OCT4 [161]. There are evidences that SOX2 is involved in the pathological differentiation of intestinal cancer cells [162].

It is shown that the absence of NANOG in mouse embryonic stem cells causes the differentiation of the cells into endodermal tissue [153, 156].

The three proteins OCT4, SOX2 and NANOG regulate each other [163]. A heterodimer formed by OCT4 and SOX2 is an important factor regulating NANOG [164]. It is also shown that OCT4 induces the differentiation of ES cells by suppressing NANOG via a negative feedback loop [165]. Further OCT4 binds to the promoter of *Pou5f1* and so also suppresses its own expression. Other studies show a positive auto-regulation of NANOG [166]. The regulatory network of the three proteins is shown in Figure 5.2.

It is assumed that the three regulators have distinct functions in pluripotent cells but work through related pathways [167]. The regulator proteins co-occupy the promoter regions of a large number of genes. Many of those

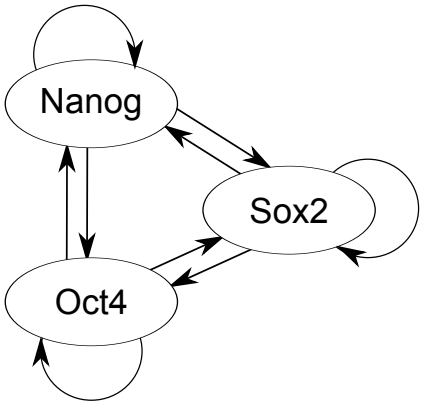


Figure 5.2: The regulatory interactions of Oct4, Nanog and Sox2, the main proteins for the maintenance of the pluripotent cell state.

genes encodes transcription factors that belong to several regulatory circuits in ES cells [157]. Even if the functions of OCT4, SOX2 and NANOG are partially overlapping different studies have shown that the correct expression of OCT4 as well as of NANOG is essential for the pluripotent state of a cell [165].

Peripheral proteins

Besides the three transcription factors OCT4, SOX2 and NANOG, a large number of proteins involved in the pluripotent state of ES cells are identified. Further their regulation and regulatory influence to the key proteins is well investigated. During the last years several regulatory networks of pluripotency-associated genes have been published [168–170]. Differing in the size, depending on the number of included peripheral genes, the largest network contains about 270 genes.

Some of those genes encode proteins having only a small influence on pluripotency and often the mechanisms of protein interactions are not yet specified in detail. But every investigated protein leads to a better understanding of the complex processes taking place in ES cells.

Although, many proteins play a very important role in cell differentiation we want to emphasize the three factors LIN28, C-MYC and KLF4. Even if we do not count them to the key proteins of pluripotency they play a very important role in current research. In different combinations together with OCT4, SOX2 and NANOG this genes allowed to induce pluripotency into non-pluripotent somatic cells [171].

Signaling pathways

As there are a lot of well examined processes in pluripotent cells a number of signaling pathways probably play an important role in pluripotency. Similar to the involved genes there are signaling pathways that are closer connected to the pluripotent cell state and wider accepted than other. In the following we describe the five most recent pathways [172–175].

JAK-STAT signaling pathway. Through the pathway signals from outside the cell are forwarded to the promoters of target genes in the core of the cell. The pathway regulates processes during cell development such as determining the cell fade, regulating the growth control or the apoptosis.

The JAK-STAT pathway is also known to play a role in maintaining the pluripotent state in mouse ES cells. It is assumed that the signaling process that includes STAT3 is sufficient but not necessary for pluripotency [176]. In human ES cells the pathway seems not to support pluripotency at all [177]. This is an evidence that in different species different pathways are involved in pluripotency.

TGFβ signaling pathway. The pathway is involved in a large number of cellular processes in the developing embryo as well as in the adult organism. First discovered in tumor cells it plays a key role in arresting cellular growth in response to extracellular signals. TGFβ belongs to a large family of related growth and differentiation factors. Other pluripotency relevant factors as Activin, Nodal and BMP4 belong to this family, too [178].

Evolutionary, the components of the TGFβ signaling pathway are well conserved and are shown to play an important role for the determination of cell fate in several organisms [179–181]. There are also evidences that TGFβ signaling supports the pluripotent state in human ES cells [182–184].

WNT signaling pathway. The gene *Wnt* is best characterized in Drosophila, where it was discovered as a recessive mutation affecting the wings [185]. The WNT signaling pathway is one of the evolutionary most conserved pathways found in many animals from C. elegans to humans. It plays an important role in embryogenesis and is also observed in several cancer tumor cells [186].

It is shown that WNT plays an important role in embryonic as well as adult stem cells sustaining the potential for self-renewal [187].

MAPK signaling pathway. Through the MAPK signaling pathway signals from the cell membrane are forwarded to the core of the cell. This pro-

cess is based on the stepwise phosphorylation of three MAP kinases (MAP3K, MAP2K and MAPK). The MAP kinases can be divided into three groups: ERKs (extracellular signal-related kinases), PAKs (p38 activated protein kinases) and JNKs (c-Jun-N-terminal kinases).

When mouse ES cells start differentiation they show a high concentrations of ERKs. The suppression of ERK keep the cell in an undifferentiated state [188]. In human, there are several genes of the MAPK pathway shown to be up- or down-regulated in ES cells. Additionally, FGF (growth promoting factor) that signals through the ERK/MAPK signaling pathway is required in ES cells [189]. For this we can conclude that the ERK/MAPK signaling pathway is active in ES cells.

PI3K-AKT signaling pathway. PI3K-AKT signaling is known to be involved in apoptosis and found in different types of cancer [190–192]. In many tumors the over-activation of the PI3K-AKT signaling reduces apoptosis and allows an unchecked growth of the cell.

In mouse ES cells PI3K support the cell proliferation [193]. It is activated by ERAS that is specifically expressed in ES cells [194]. It is also involved in the maintenance of the pluripotent cell state. As in human ES cells some components of the PI3K-AKT signaling pathway are up-regulated [189] it is connected to pluripotency in human, too.

5.1.3 iPS cells - first successes in ESC research

Induced pluripotent stem cells (iPS cells) are artificially derived from non-pluripotent cells. Therefore, adult somatic cells are forced to express gene patterns similar to those in pluripotent cells and have most of the characteristics of embryonic stem (ES) cells.

In 2006, K. Takahashi and S. Yamanaka [149] have successfully reprogrammed differentiated mouse cells into a pluripotent state. They used a retroviral transduction of the four transcription factors OCT3/4, SOX2, KLF4, and C-MYC, become known as Yamanaka factors. Among other things, the iPS cells show similarities to ES cells in morphology, proliferation, gene expression and telomerase activity. Additionally, using this artificial induced pluripotent cells it is possible to grow tissues of all three germ layers in vitro.

One problem of the method introduced by K. Takahashi and S. Yamanaka is the high potential for tumor growth observed in iPS cells in animal testing. The most likely reason for this are the two transcription factors, KLF4 and C-MYC, well-known as oncogenes. For this reason researchers are looking for alternative methods for medical practice.

In 2007, M. Nakagawa et al. introduced alternative genes for the reprogramming of somatic cells. Besides OCT3/4, SOX2 and KLF4 they use NANOG and so eliminate the oncogene C-MYC [150]. Additionally in 2008, M. Stadtfeld et al. use adeno viral transfection for the migration of genes in the cell [195].

Furthermore, J. B. Kim et al. induced a pluripotent state in adult neural stem cells using only OCT3/4 [151]. The other genes are naturally expressed in adult neural stem cells.

5.1.4 Stem cell therapies - A glace to the future of medicine

Stem cell therapies are known for more than 30 years. Today, the treatments of several cancer types as leukemia [196] and lymphoma [197] using adult stem cells are very popular in medicine. Even so, after 30 years the transplantation of hematopoietic stem cells is still a risky procedure with various serious complication [198]. Before transplanting the bone marrow cells of a healthy donor chemotherapy is used to kill the cancerous stem cells. So the transplanted hematopoietic cells can breed new healthy blood cells without being influenced by the pathological cells.

Although this technique is well established, currently it is the only clinical usage of stem cell therapy. Many other potential fields are still under investigation and subject of current research. So, possibilities for the treatment of neurodegenerative brain diseases as Alzheimer's or Parkinson's disease are under investigation [7,199–201]. Furthermore stem cell therapies show a high potential for the treatment of cardiac diseases by regenerating damaged heart muscle tissues [8]. Also the usage for the treatment of blindness and visual impairment is under investigation. Recent studies show that the transplantation of retinal stem cells into damaged eyes may restore the function of the eyes [202].

Until now embryonic stem (ES) cells have not been used for stem cell therapies.

Because of the pluripotent state ES cells are more flexible than adult stem cells. As they can be differentiated into any possible tissue occurring in the human body it is imaginable that ES cells can be used to produce tissues for medical therapies. They offer a renewable source of highly diverse cells and tissues that can be used for the treatment of a large number of diseases.

Before using new stem cell therapies in clinical practice researchers have to face a lot of sophisticated challenges. The first step is a complete understanding of the molecular mechanisms in pluripotent cells and of the differentiation

GA/SVM500	Information gain	Random forest
Fam134b	*Dppa5a*	*Ottmusg00000010173*
Pam	*Gdf3*	*Dppa5a*
Dub1	*Mybl2*	*Gdf3*
F2rl1	*Dppa2*	*Mybl2*
Gldc	*Dppa4*	*2610305d13rik*
Spp1	*Ottmusg00000010173*	*Bb001228*
Dazl	*2610305d13rik*	*Au019176*
Ccnd2	*Rex2*	*Esrrb*
100043292	*Zfp42*	*Gtsf1l*
Otx2	*Bb001228*	*Dppa4*
Utp20	*Tdgf1*	*Tdgf1*
Jam2	*Esrrb*	*Dppa2*
Gjb5	*2410004a20rik*	*Rex2*
Foxc1	*Calcoco2*	*Trap1a*
Bb001228	*Spp1*	*E130012a19rik*
Calcoco2	*Gart*	*Gart*
Crim1	*E130012a19rik*	*Morc1*
Irs1	*Gtsf1l*	*2410004a20rik*
Mal	*F2rl1*	*Zfp42*
Col4a5	*Ttr*	*Dnmt3l*

Table 5.1: The top 20 genes oft the PLURI data set selected by the three feature selection methods. The genes occurring in all three lists are underlined.

process of ES cells.

5.2 Identified biomarkers for pluripotency

We use three different feature selection methods for the identification of potential biomarkers for pluripotency. In contrast to our results in Chapter 4 we do not validate our results using an independent test. We validate our results from a biological point of view using gene set enrichment analysis and discuss the obtained results in regard to other recent publications. As we do not need an independent test set we use all samples for feature selection. As feature selection methods we use information gain, random forest and our wrapper of genetic algorithm and support vector machine (GA/SVM500) resulting in three ranked gene lists The top 20 genes of these lists are listed in Table 5.1

5.2.1 Gene set enrichment analysis

We evaluate the biological relevance of the identified potential biomarkers using gene set enrichment analysis with a hyper geometric function as described in Section 2.2.2. As reference set serves a list of all genes of the microarrays contained in the PLURI data set. As test sets, we use extracts of the best ranked genes of information gain, random forest and our GA/SVM500. In the absence of a pre-defined number of most important genes we use increasingly larger sets of genes starting with the best 40 genes of each list and increase the number up to 200 in steps of 20. This way we get 9 lists for each feature selection method.

In recent work we find plenty of networks describing the complex processes in pluripotent cells [168–170, 203]. Although those regulatory networks are already very large we assume more genes to be involved in pluripotency than these networks contain. To give a first clue of the quality of the identified biomarkers use already verified networks and pathways associated with pluripotency for gene set enrichment analysis. Additionally, we compare our gene lists to 190 pathways collected by Ingenuity Systems (www.ingenuity.com).

Pluripotency networks

In a first step we compare our 27 test lists to the genes of several well established pluripotency networks. As shown in Table 5.2, a significant enrichment (p-value < 0.05) is found for nearly all 27 tested gene lists in the pluripotency networks published by Som et al. [168], MacArthur et al. [170] and Müller et al. [169] For two other networks ('PluriUp' and 'PluriPlus' [203]) that are much larger than the other three networks our GA/SVM500 does not show a significant enrichment but information gain and random forest do. Additionally, we analyze two gene lists referred to as 'Tissue+' and 'Tissue-' that contain genes that are known as enriched respectively depleted in embryonic tissues [204]. In 'Tissue+' information gain as well as random forest show a significant enrichment. As expected, in 'Tissue-' no enrichment is found by any of the three methods.

Pluripotency-related pathways

After considering different surveys of pathways playing an important role in pluripotent cells, we use the five most important pathways as reference for another enrichment analysis that is shown in Table 5.3 [172–175]. As reference we use the KEGG [205] pathways for JAK-STAT, WNT, TGFβ and MAPK signaling. For PI3K-AKT we use the pathway provided by Ingenuity

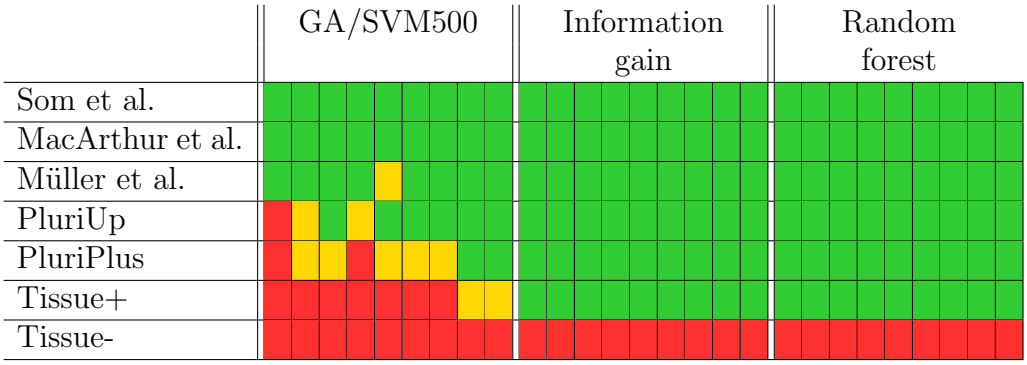


Table 5.2: Results of the gene set enrichment analysis in pluripotency-related networks. We use incrementally larger sets of genes found by our three feature selection methods. Green: significant enrichment (p-value < 0.05). Yellow: enrichment. Red: no enrichment.

Systems. The gene lists selected by our GA/SVM500 show a significant enrichment (p-value < 0.05) in the WNT and the JAK-STAT pathway. We find enrichment in the PI3K-AKT pathway as well. For some gene lists of the information gain method we find a significant enrichment in the MAPK signaling pathway. The list selected by random forest shows a non-significant enrichment in the JAK-STAT, MAPK and PI3K-AKT pathways.

Pathways provided by Ingenuity Systems

In Table 5.4 we give a conclusion of the results of the gene set enrichment analysis using 190 pathways collected by Ingenuity Systems. We define the genes selected by one of our feature selection methods as significant enriched (p-value < 0.05) in a pathway, if at least six of the nine sets show a significant enrichment. First we list all pathways that show a significant enrichment for all three feature selection methods. As we mainly focus on the genes selected by GA/SVM500 in the second part of the table we list, those pathways show only a significant enrichment in the lists selected by GA/SVM500. The listed pathways are associated with pluripotency. The references are also listed in the table.

5.2.2 Biological relevance of the selected biomarkers

In the following we discuss the results of the gene set enrichment analysis and discuss the selected biomarkers in respect to current research. As we show in Chapter 4 our GA/SVM200 outperforms information gain as well as random forest referred to the quality of selected genes. For this reason in our

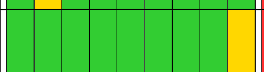
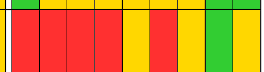
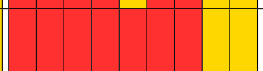
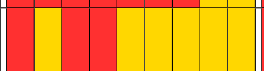
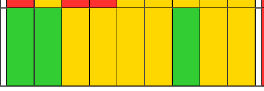
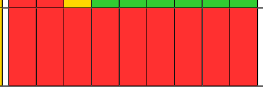
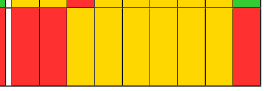
	GA/SVM500	Information gain	Random forest
JAK-STAT signaling pathway			
WNT signaling pathway			
TGFβ signaling pathway			
MAPK signaling pathway			
PI3K-AKT-Signaling			

Table 5.3: Results of the gene set enrichment analysis in pluripotency-related pathways. We use incrementally larger sets of genes found by our three feature selection methods. Green: significant enrichment (p-value < 0.05). Yellow: enrichment. Red: no enrichment.

discussion we focus on the results of the GA/SVM500.

Gene set enrichment analysis

Table 5.2 shows the enrichment of gene sets selected by information gain, random forest and our GA/SVM500 in pluripotency-associated regulatory networks. Each of the three networks Som et al. [168], MacArthur et al. [170] and Müller et al. [169] contains between 120 and 270 genes. In these networks we find a significant enrichment for all three methods what shows that all of those algorithms find a large number of genes already known to play an important role in pluripotency.

The lists 'PluriUp' and 'PluriPlus' are derived from a cluster analysis using self-organizing maps [203]. For these approximately 2,600 and 1,000 genes containing lists the gene lists selected by information gain and random forest show a significant enrichment. We also find a non-significant enrichment in the gene lists selected by our GA/SVM500.

'Tissue+' and 'Tissue-' are lists of genes known as enriched respectively depleted in embryonic tissue [204]. Again the genes selected by information gain and random forest show a significant enrichment for 'Tissue+' whereas the gene set selected by GA/SVM500 does not show enrichment. 'Tissue+' contains nearly 5,000 genes with a role in embryonic development. The list 'Tissue-' is added as a kind of control list. We do not expect enrichment in genes depleted during the embryonic development and none of the three

Pathways significant enriched for all methods	Pluripotency relevance
Cell-Cycle-G1-S-Checkpoint-Regulation	[206]
Glycin-Seronin-and-Threonine-Metabolism	[207]
Pathways significant enriched for GA/SVM500	**Pluripotency relevance**
BMP-Signaling-Pathway	[208–211]
Aminoacyl-tRNA-Biosynthesis	[212]
EGF-Signaling	[213]
IGF-1-Signaling	[214, 215]
IL-2-Signaling	[216, 217]
Neutrophin-TRK-Signaling	[218]
Starch-and-Sucrose-Metabolism	[219]
Tight-Junction-Signaling	[210]

Table 5.4: List of pathways that show a significant enrichment in the results of all three feature selection methods (top). List of pathways that show only a significant enrichment in the results found by our GA/SVM500 (bottom). For all pathways we add references that show the connection between the pathways and pluripotency.

algorithms found enrichment.

The enrichment we found for the lists of all feature selection methods punctuates our assumption that the algorithms are able to identify genes associated with pluripotency. As discussed in Chapter 4 the genes selected by our GA/SVM500 are less redundant then the gene lists of information gain and random forest. Even though, our GA/SVM500 finds a significant enrichment in most of the gene regulatory networks of which we assume that they contain many redundancies. Especially the two lists 'PluriUp' and 'PluriPlus' that are identified by cluster analysis contain a large number of closely related genes. We do not find enrichment for the GA/SVM500 genes in these two lists.

As there are a lot of pathways known to be involved in the maintenance of the pluripotent cell state we also perform a gene set enrichment analysis using the five most popular of those pathways [172–175]. Table 5.3 shows the results of this analysis. We see that the genes found by the GA/SVM500 are enriched in four of the five pathways. Furthermore, in two of these pathways we can consider the enrichment as significant. The TGFβ signaling pathway is the only pathway where we do not find enrichment of genes selected by the GA/SVM500. Also the gene sets identified by the other two methods do not show enrichment in this pathway. The genes selected by information gain are only significantly enriched in the MAPK signaling pathway. The

genes identified by random forest are enriched in three of the five pathways but none of these enrichments is considered to be significant. We conclude that the genes selected by our GA/SVM500 are involved in more different pluripotency-related signaling processes than genes found by the other two methods.

In a last step we perform a gene set enrichment analysis using 190 pathways collected by Ingenuity Systems (www.ingenuity.com) . In Table 5.4 we list the pathways with a significant enrichment for all methods as well as pathways that show only a significant enrichment in the gene sets selected by the GA/SVM500. All listed pathways are associated with pluripotency even if not all of these connections are finally verified.

In the following we want describe four pathways with a direct link to the five most popular pluripotency related-pathways. For these pathways only the gene lists selected by our GA/SVM500 show a significant enrichment. This underlines our conclusion that the genes selected by GA/SVM500 are involved in many different pathways linked to pluripotency.

- The most interesting pathway is the BMP signaling pathway. BMP is a member of the same family of growth and differentiation factors TGFβ belongs to. *TGFβ* is often used as prototype for genes of this family. In fact for an involvement in pluripotency *BMP* is the most promising candidate of this family [172–175].

- The IL-2 signaling pathway is directly linked to two well-known pluripotency-related pathways. Interleukin 2 activates the JAK-STAT pathway and at the same time it is inhibited by derivatives of the MAPK signaling pathway [216].

- Also the Neutrophin-TRK signaling pathway is of great interest. PI3K signals through the TRK receptor and this way mediates the activation of Neutrophin [217, 218].

- As a last pathway we want to mention IGF-1 signaling. IGF-1 is known as the most potent natural activator for the AKT signaling pathway. It stimulates the cell growth and the proliferation and additionally inhibits the programmed cell death. All of these processes are involved in the maintenance of pluripotent cells [215].

As discussed in Chapter 4 the genes selected by our GA/SVM contain less redundant genes than the gene lists selected by information gain and random forest. For this reason they are able to cover a wider range of

pluripotency-associated processes. This is supported by the observed enrichment of GA/SVM genes in a large number of pluripotency-related signaling pathways.

Recent work

The gene highest ranked by our GA/SVM500 is *Fam134b*. It is selected in about 35% of the small gene sets the GA/SVM500 results in. Currently, the exact function of the genes belonging to the Fam134 family is unknown. Besides *Fam134b*, that is primarily expressed in sensory and autonomic ganglia, this family contains two other members, *Fam134a* and *Fam134c*. The human orthologous gene *FAM134B* (also called *JK-1*) is supposed to promote cell proliferation and is possibly linked with specific kinds of carcinoma [220]. Mutations of the gene *FAM134b* cause several sensory and autonomic neuropathies [221]. It also encodes a newly identified cis-Golgi protein [222]. Knockdown experiments in mouse show that *Fam134b* is jointly responsible for the structure of the cis-Golgi compartment. Loss-of-function mutations induce apoptosis in specific types of ganglion neurons. In the context of embryonic stem cells, *FAM134B* possibly influences the two cell surface markers SSEA1 (stage-specific embryonic antigen 1) and AP (alkaline phosphatase). Both proteins transported to the cell membrane are localized in the Golgi [223]. Although, in human it is shown that the two markers are expressed in embryonic stem cells [224], the exact function of the two proteins is still unknown.

The second most important gene selected by the GA/SVM500 is *Pam* (peptidylglycine alpha-amidating monooxygenase) that encodes a multifunctional enzyme [225]. At the moment, *Pam* is not closely associated with pluripotency.

The third ranked gene is *Dub1* (deubiquitinating enzyme 1), an enzyme with growth-suppressing activity [226]. As it is shown that ubiquitination plays an important role in repressing developmental control genes in embryonic stem cells, *Dub1* is a promising new candidate gene for regulating pluripotency [227].

Furthermore we examine gene four to ten of the list of genes selected by our GA/SVM500. More or less strong, most of these genes are also associated with pluripotency. In the following we give a short summary of these genes.

- *F2rl1* (coagulation factor II (thrombin) receptor-like 1, also known as *Par2*): Acting as a G-protein-coupled receptor it plays an important role in the mouse blastocyst [228]. It is also supposed to be involved in the distinction between multi- and pluripotent cells [229].

- *Gldc* (glycine dehydrogenase (decarboxylation)): *Gldc* belongs to a number of genes known as consistently up-regulated in induced pluripotent and embryonic stem cells [230].
- *Spp1* (secreted phosphoprotein 1, also known as osteopontin): It is regulated by a heterodimer of *Oct4* and *Sox2* two of the most important genes in pluripotent stem cells (see also Section 5.1.2) [231].
- *Dazl* (deleted in azoospermia-like): It plays a role in the maintenance of the pluripotent cell state in primordial germ cells [232]. It also acts as a translational factor during self-renewal and differentiation in mouse embryonic stem cells [233].
- *Ccnd2* (cyclin D2): It is repressed by the WNT transcription factor TCF3 that is closely related to pluripotency [234].
- *10043292* (also known as *GM4340* (predicted gene 4340) and THO complex 4-like): Its function is unknown.
- *Otx2* (orthodenticle homolog 2): *Otx2* is a well-known part of different pluripotency networks [168–170, 203].

Among the top ten genes selected by information gain and random forest are three members of the Dppa family (developmental pluripotency-associated), namely *Dppa5a*, *Dppa2* and *Dppa4*. These genes are already included in the gene regulatory networks of Som et al. [168], MacArthur et al. [170], Müller et al. [169] and other collections of pluripotency-associated genes [203]. Other genes selected by these two algorithms are also established in at least one of those gene lists. This concerns *Gdf3* [235, 236] and *Mybl2* [237] both ranked top ten by information gain and random forest, as well as the genes *Bb001228*, *Esrrb* [238], *Rex2* [229] and *Zfp42* [239]. Besides *Gtsf1l* that is also linked to pluripotency [235, 236], information gain lists 3 genes with unknown function, *Ottmus00000010173*, *26100305d13rik* and *Au019176*. *Ottmus00000010173* and *26100305d13rik* are also selected by random forest.

Finally, we want to bring up the gene *Bb001228* (also known as *Tet1*) that is ranked top 20 by all three methods. Knockdown experiments showed that *Bb001228* is involved in the regulation of pluripotency and developmental factors in embryonic stem cells [240, 241]. A knockdown of *Bb001228* impairs the self-renewal capability of the cell [242].

Even if most of the genes selected by the GA/SVM500 can already be linked to pluripotency they are not yet established in current pluripotency networks. For this reason we assume GA/SVM500 well suited for finding new

potential biomarkers for pluripotency. Information gain and random forest mainly selecting well established genes with a well-known function in pluripotency. Together with the observed higher classification accuracy of the genes selected by our GA/SVM500 shown in Chapter 4 we prefer GA/SVM500 for biomarkers selection over information gain and random forest.

5.3 Alzheimer's disease

The examination of the brain of people died from Alzheimer shows reductions in the size of specific brain regions. It is assumed that this cell death is initiated by excitotoxic processes. At this, an excessive stimulation by neurotransmitters causes the death of neural cells.

During the last years recent studies explore some underlying molecular mechanisms of the disease that advance the basic understanding of Alzheimer. Still, our knowledge about the causes of Alzheimer's disease and the possibilities of treatment are restricted.

In the brain of Alzheimer diseased patient we can find two abnormalities, plaques and tangles. These microscopic molecules can also be observed in normal aged brain, but the concentration in Alzheimer disease affected brain is much higher. Senile plaques are deposits in the grey matter of the brain. The main component of the plaques are Aβ (amyloid β) peptides, that are considered to act as a neurotoxin. The second abnormality found in the affected brains is neurofibrillary tangles. Tangles are formed by a hyper phosphorylation of the microtubule-associated protein TAU (taurine) that causes TAU to group in an insoluble form. This involves a loss of the function of TAU whose main function is the stabilization of the microtubules. Tangles can be associated with numerous diseases summarized as tauopathies. Combined with a large number of plaques Alzheimer can be diagnosed with a high probability. The number and distribution of the tangles allow classifying the progress of the disease.

5.3.1 Molecular mechanisms of Alzheimer

Alzheimer's disease is characterized by the death of neural cells, initiated by the excessive stimulation of the cells by neurotransmitters. This process is called excitotoxicity.

Amyloid plaques and neurofibrillary tangles are the most obvious pathological alterations in Alzheimer diseased brain. Their composition, formation and effect on the disease is already well examined but not yet fully understood. Recent studies show, that APP (amyloid precursor protein) specifi-

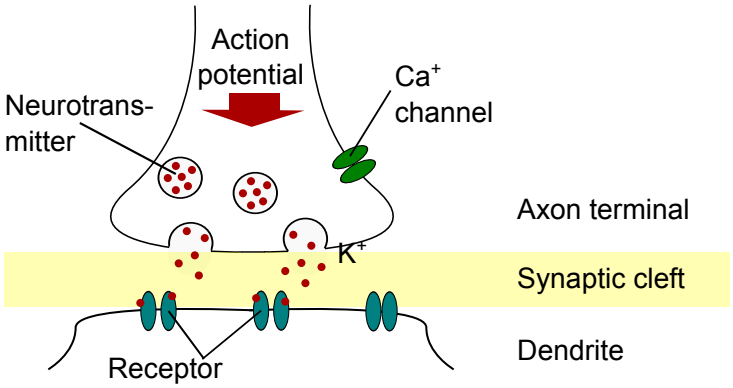


Figure 5.3: Chemical signal transmission between two neurons. Stimulated by an electrical signal the pre-synaptic cell releases chemical molecules, so called neurotransmitter. These transmitter bind to receptors of the post-synaptic neuron. This way secondary messenger pathways or an electrical responses may be initiated.

cally its metabolite Aβ (amyloid β) causes senile plaques. The neurofibrillary tangles are caused by a hyper phosphorylation of the microtubule-associated protein TAU.

Structure and signal transmission in neural cells

A neuron is a cell that is able to transmit information by electrical and chemical signals. It usually consists of dendrites, a cell body and an axon. Dendrites are thin structures arising from the cell body. Each neural cell contains hundreds of those dendrites that receive the signals sent by other neural cells. The Axon is a fibroid extension of the cell body that forward signals to other neural cells.

The transmission of information inside a neuron is processed via electrical signals that pass through the axon. The terminal of the axon is related to dendrites of other neural cells via synapses. At the synapses a chemical transmission of signals takes place. Induced by the electrical impulse passed through the axon, the synapses release neurotransmitter. These react to specific receptors located on the postsynaptic membrane of dendrites of the signal receiving neuron. Figure 5.3 shows a schema of signal transmission in neural cells.

The role of APP and Aβ oligomers

APP (amyloid precursor protein) is an integral membrane protein that probably plays a role in the regulation of synapse formation [243] and neural plasticity [244]. The manly studied function of APP is its role as precursor protein for Aβ (amyloid β). The proteolysis of APP by the two enzymes β- and γ-secretase generates a 39 to 42 amino acids long peptide that is the main component of the amyloid plaques [245] found in the brain of Alzheimer diseased patients.

As Aβ is considered to be neurotoxic it is likely that a reduction of the plaques can help to improve the symptoms of Alzheimer [246]. At the moment the development of new drugs and therapies for the treatment of Alzheimer's diseases focus on the reduction, removal and prevention of amyloid plaques.

Recent studies show that a small amount of Aβ is needed for a normal information transfer to the neurons [247]. In solution, the secondary structure of Aβ usually contains short regions of β-sheets and α-helices, while it forms large alpha helical structures in membranes. In a high concentration the Aβ molecules form a β-sheet rich tertiary structure. These amyloid fibrils deposits outside neurons and form the senile plaques in the grey matter of the brain.

The role of TAU

TAU belongs to the class of microtubule-associated proteins (MAPs). It modulates the stabilization of axonal microtubules. TAU ensures the stability as well as the flexibility of microtubules. For building the filamentary structure of the microtubules, several TAU proteins are connected using recurring segments. Mainly expressed in neurons, there are six known isoforms in human organisms. They result of alternative splicing of the *MAPT* gene. Besides occurring in various isoforms, TAU is a phosphoprotein with several phosphorylation sites.

Among other things, the phosphorylation of TAU is developmentally regulated. In embryonic tissue TAU is much more phosphorylated than in adult [248]. A hyper phosphorylation of the paired helical filaments that are formed by TAU inhibit the binding of TAU to microtubules. As a result the microtubules become unstable and start to disintegrate. The unbound TAU proteins start to clump and form the tangles that can be found in neurons of Alzheimer affected brains.

Those tangles can be found in other pathological brains as well. Diseases characterized by the presence of tangles are classified as Tauopathies.

The connection between Aβ and TAU

Senile plaques as well as tangles are indicators for several diseases. Together they can be found only in the brains of Alzheimer patients. While amyloid plaques are extracellular depositions formed by Aβ, tangles are intracellular depositions of hyper phosphorylated TAU. But how do these two components interplay to cause the death of neural cells?

In 2001 two independent groups found evidences that Aβ induces the pathological phosphorylation of the TAU protein in mouse [249,250]. Further, it is shown that a reduction of TAU decreases the excitotoxicity in vivo [251]. The main function of TAU is the stabilization of the microtubule of neural cells [252–254]. As it also occurs in small quantities in dendrites of neural cells, an additional function is identified in targeting the kinase FYN [255]. In Alzheimer the hyper phosphorylated TAU accumulates in the dendrites [256] where Aβ likely attacks the postsynaptic receptors of the cell [257–259]. This can explain the influence of TAU on the Aβ neurotoxicity when Alzheimer's disease is initiated.

Familial clustering in Alzheimer - the role of APOE and PSEN1/2

In cases of Alzheimer's disease that show a familial clustering Aβ seems to play the key role. Several genetic predispositions that lead to the overproduction of Aβ are identified.

- The gene encoding for APP (amyloid precursor protein) was the first familiar early-onset Alzheimer gene that has been identified. Alzheimer can be associated with several mutations on the gene [260–263], but only some of them lead to an early-onset form.

- Various missense mutations are detected in *PS1* (presenilin-1) that can be associated with an early-onset form of Alzheimer's disease [264]. In fact mutations on *PS1* are responsible for most of the familiar early-onset forms [265, 266]. The gene *PS2* (presenilin-2) is a sequence homolog of *PS1* containing some missense mutations also associated with Alzheimer, but seldom with an early-onset form [267, 268]. The genes *PS1* and *PS2* encode for the proteins PSEN1 and PSEN2 that are subunits of the γ-secretase.

- APOE (apolipoprotein E) is part of certain lipoproteins and has an important function in lipid metabolism [269]. In human, there are three known isoforms of APOE (APOE2, APOE3, APOE4). In recent studies it is shown that the existence of at least one APOE4 allele increases the risked of Alzheimer by 40% to 60%, whereas the existence

of APOE2 decreases the risk [270–272]. The exact effect of APOE on Aβ is still not fully understood and part of current research [273]. As APOE is also expressed in blood, the concentration of APOE4 is used for the detection of a higher Alzheimer risk.

At the moment only a small percentage ($< 5\%$) of the Alzheimer's diseases can be associated with a familial clustering. In the past, dementia is not always clearly diagnosed as Alzheimer and people not always reach a high age so the number of cases with a familial clustering could be much higher.

Signaling processes assumed to be involved in Alzheimer

The mechanisms that lead to neurodegeneration in Alzheimer are not yet fully understood. As previously described Aβ and APP are identified as two key proteins in this process. Some cases of Alzheimer occur with a familial clustering so mutations of genes as APP, PSEN1, PSEN2 or APOE are identified as cause. The question how Aβ leads to synaptic injury and neurogenesis defects is not yet answered, but a number of possibilities are under investigation.

As previously described hyper phosphorylated TAU in the dendrites is assumed to reinforce the exitotoxic effect of Aβ. The membrane-associated kinase FYN interacts with TAU [274–276] and could be responsible for the aggregation of hyper phosphorylated tau proteins in the dendrites.

The influence of other signaling proteins to Alzheimer, such as CDK5 [277, 278] and CSK3β [279, 280] has also be shown in different studies.

Further it is examined what role alterations in glutamate receptors play in excitotoxicity in Alzheimer [258, 281–283]. Moreover the role of mitochondrial dysfunction [284, 285], lysosomal failure [286] and signaling pathways related to synaptic plasticity and neuronal cell death are under investigation. Particular interest is also attached to specific signaling pathways of the MAPK family such as ERK [287–289], JNK [290,291] or PAK pathways [292].

5.3.2 Diagnosis and treatment

In the diagnostic of Alzheimer's diseases there have been significant improvements during the last years. But still, there is no possibility to diagnose Alzheimer with a guaranty of 100%. The differential diagnosis for Alzheimer includes plenty of measurements and focus on the exclusion of all other possible causes for the observed symptoms. At the moment, an exact diagnosis is only possible post-mortem by examining the brain tissues.

In April 2011, the National Institute of Aging and the Alzheimer's Association published guidelines for the diagnosis of Alzheimer's disease [293].

They classify the progress of the disease into three stages that merge into each other:

- The asymptomatic, preclinical phase [294]
- The symptomatic, pre-dementia phase [295]
- The dementia phase [296]

The diagnosis of Alzheimer in the asymptomatic, preclinical phase is not yet possible and is very restricted in the symptomatic, pre-dementia phase [297]. Only in the dementia phase Alzheimer can be diagnosed with a high probability.

In clinical practice the diagnosis of Alzheimer usually starts with a patients and familiar history, cognitive tests and behavioral assessments. To prove cognitive impairments a 30-point questionnaire test, known as minimental state examination (MMSE), is widely used [298]. To exclude other diseases that can cause the observable symptoms of cognitive impairment, methods as X-ray computed tomography and magnetic resonance imaging are used. Brains of Alzheimer patients show shrinkage in specific regions that can be detected using magnetic resonance imaging. Further the test of the cerebrospinal fluid can show abnormalities such as a low level of Aβ (amyloid β) peptides and a high level of TAU proteins [299]. The measurement of Aβ, TAU and other potential liquor biomarkers requires a lumbar puncture, an invasive procedure with possible harmful side effects. Biomarkers in blood are not yet know.

Several studies deal with the prevention of Alzheimer's disease. They show that enough exercise and a balanced diet decrease the risked of Alzheimer [300, 301]. Especially, diseases of civilization as high blood pressure and an increased cholesterol level seem to be negative factors [302–305].

At the moment there is no effective treatment for Alzheimer's disease. Although a number of drugs for the treatment of the symptoms have been developed, their influence on the disease is relative small. The available drugs are not able to stop the progression of the disease.

5.3.3 Perspective

Based on an increased age of the world population recent studies estimate that the number of people living with Alzheimer's disease will double every 20 years without an effective treatment [9,10]. This would produce enormous costs for the health systems, especially in developing countries.

Since the last 100 years, the Alzheimer's disease is known but the research is still at the beginning. For the development of new approaches for diagnosis and therapy we need to completely understand the underlying molecular mechanism of the disease.

Current approaches for new Alzheimer therapies are mainly based on the insights about Aβ (amyloid β) and TAU. The neurotoxin Aβ can be found in Alzheimer affected brains in form of senile plaques. Aβ is assumed to cause the death of neural cells by excitotoxic processes. So, one approach for the treatment of Alzheimer is the reduction of the amyloid plaques in the grey matter of the brain. The hyper phosphorylation of the protein TAU that deposits in the dendrites of neural cells, so called tangles, is a second characteristic of Alzheimer's disease. The hyper phosphorylated TAU is assumed to reinforce the toxic effect of Aβ during the initiation of the disease. Therefore, a further possibility for the therapy of Alzheimer is the reduction of the tangles formed by hyper phosphorylated tau proteins.

As the death of cells in Alzheimer's disease affected brains is irreversible, a successful treatment of Alzheimer starts with an early diagnosis of the disease. Even if brain scans can detect evidences for Alzheimer years before they cause any problems, this methods are toe expensive for routine precaution. For this reason, easily identifiable biomarkers, that enable an effective diagnosis before any symptoms occur, are needed.

Some interesting approaches for the early diagnosis of Alzheimer focus on the eye. Using laser technique previous studies prove a higher concentration of Aβ in the eyes of Alzheimer patients than in healthy peoples [306, 307] . On the Alzheimer's Association International Conference 2011 Shaun Frost et al. [308] introduced a promising method for the early detection of Alzheimer. Using retinal scans they show that the ratio between the arterial and the venous diameter is associated with the density of amyloid plaques in the brain. This could also be proven for participants without any symptoms. Although it is only evaluated on a small numbers of people, it opens the possibility for a non-invasive systematic scanning for Alzheimer. The eye test is not valid on its own but needs an additional blood analysis scanning for Alzheimer specific markers.

5.4 Identified biomarkers for Alzheimer's disease

For the identification of potential biomarkers for Alzheimer's disease we use three feature selection methods showing good results in our analyses in Chap-

GA/SVM500	Information gain	Random forest
LOC642711	*FLJ11903*	*LOC283345*
PRKXP1	*TNCRNA*	*FLJ11903*
LOC283345	*LOC283755*	*LOC642711*
SST	*PTPN3*	*GPRASP2*
LY6H	*PCYOX1L*	*TNCRNA*
ERCC3	*PPIH*	*PCYOX1L*
LOC643287	*HSD17B7*	*MID1IP1*
TNNI3K	*GPRASP2*	*PDZD11*
CDK2AP1	*LOC283345*	*MAFF*
FBXO16	*C6ORF151*	*LOC283755*
GEM	*METTL7A*	*FLJ25477*
TAF3	*MRPS22*	*BCL6*
ZNF415	*CDC37*	*PALLD*
LOC285927	*NFKBIA*	*LOC645352*
MAEL	*FAM63A*	*EIF3S12*
SUPV3L1	*RAD51C*	*SLC12A7*
C6ORF151	*ANP32B*	*MGC12488*
FAM54B	*UBXD4*	*NFKBIA*
PCYOX1L	*TERF2IP*	*C6ORF151*
EP300	*BCL6*	*ATP5B*

Table 5.5: The top 20 genes of the AD data set selected by the three feature selection methods. The genes occurring in all three lists are underlined.

ter 4. As we do not use an independent test set for the validation of our results we run the feature selection methods on the whole AD data set containing all available samples. In this chapter we perform the evaluation of the selected biomarkers from a biological point of view using gene set enrichment analysis as well as recent studies for validation. Our three feature selection algorithms are information gain, random forest and a wrapper of genetic algorithm and support vector machine (GA/SVM500). Table 5.5 lists the 20 top-ranked genes of each algorithm using the AD data set for feature selection.

5.4.1 Gene set enrichment analysis

We use a gene set enrichment analysis (described in Section 2.2.2) with a hyper geometric function to evaluate the biological relevance of our results. We use the best ranked genes of our three feature selection methods for modeling our test set. As the number of the most important genes is not determined we use incrementally larger sets of 40, 60, ..., 200 genes. This

way we create nine test sets for the list of each feature selection algorithm. As reference set serves the list of all genes contained in the microarrays forming the AD data set.

For the enrichment analysis of our test sets we use the well-known Alzheimer pathway provided by KEGG [205] as well as various collections of Alzheimer-related genes. As the research in Alzheimer's disease is a late breaking topic and time is needed to verify results before including them to regulatory networks and other gene collections we also compare our results to the findings of two actual gene expression studies identifying Alzheimer-associated genes. Other than in pluripotency the signaling pathways important in Alzheimer's disease are not yet well defined. For this reason we use a list of 190 pathways provided by Ingenuity Systems (www.ingenuity.com) for our gene set enrichment analysis.

Alzheimer-related gene collections

Table 5.6 shows the results of our analysis for Alzheimer-related networks and gene collections. The only regulatory network we use, is the KEGG network for Alzheimer's disease ('KEGG AD') [205]. For this the enrichment analysis shows a significant enrichment (p-value < 0.05) for the genes found by information gain and random forest, but no enrichment for GA/SVM500. We observe nearly the same results analyzing two well-known collections of Alzheimer genes: 'AlzGene' [309] and 'Genotator' [310]. Here we find the lists identified by GA/SVM500 at least non-significant enriched. Furthermore we analyze the enrichment of a large collection of genes from 'GeneCards' (www.genecards.org). For this collection we find a significant enrichment for all three methods.

Besides these gene collections we use two lists created by Soler et al. [311] and Goni et al. [312] These lists are results of gene expression studies for the detection of new biomarkers. The GA/SVM500 shows a significant enrichment for both of these new studies, whereas information gain and random forest only show a significant enrichment in the list of Soler et al. As second part Goni et al. [312] analyze biomarkers in blood. As expected, for this gene set none of the three methods show any enrichment.

Pathways provided by Ingenuity Systems

An analysis of the 190 pathways collected by Ingenuity Systems reveals pathways that are significant enriched (p-value < 0.05) in the gene lists found by our three feature selection methods. We consider a pathway as significant enriched for a method if at least six sets show a significantly enrichment.

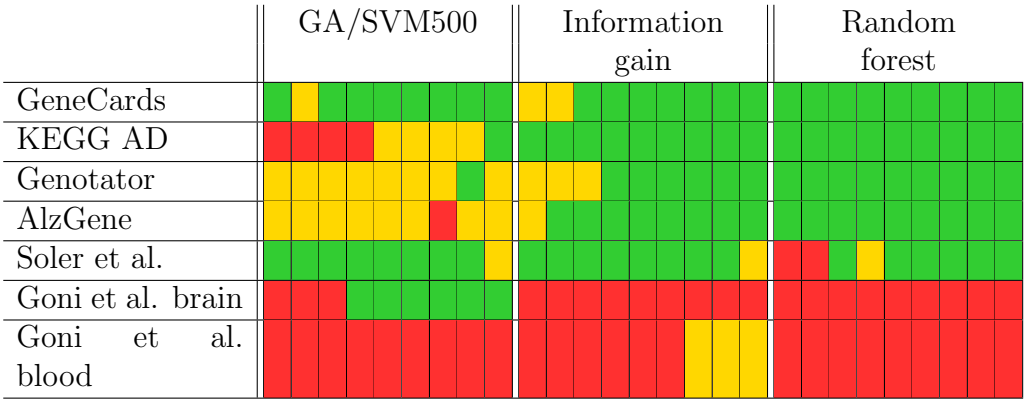


Table 5.6: Results of the gene set enrichment analysis in Alzheimer's disease-related networks. We use incrementally larger sets of genes found by our three feature selection methods. Green: significant enrichment (p-value < 0.05). Yellow: enrichment. Red: no enrichment.

In Table 5.7, we list the pathways showing a significant enrichment for all three methods as well as those pathways showing a significant enrichment for GA/SVM500 only. We also list references that link the found pathways to Alzheimer's disease.

5.4.2 Biological relevance of the selected biomarkers

In the following we discuss the results of the gene set enrichment analysis for the AD data set. Furthermore we take a closer look to the genes top-ranked by information gain, random forest and our GA/SVM500. We present evidences that support an involvement of those genes in processes of Alzheimer's disease.

Gene set enrichment analysis

The gene set enrichment analysis enables us to estimate the quality of our results without performing biological experiments. In a first step we compare the resulting gene lists to a number of lists containing of Alzheimer-associated genes. In Table 5.6 we see the enrichment analysis for seven Alzheimer-associated gene sets. 'GeneCards' is the most widespread list among these gene collections. It contains nearly 2,000 genes associated with Alzheimer. Besides the well investigated genes such as APP, APOE, PSEN1 and PSEN2 it also includes genes of the wider periphery with more indirect influence to the disease. We observe a significant enrichment for all three methods in this gene collection. For this reason we assume that all three methods are

Pathways significant enriched for all methods	Alzheimer relevance
Alanine-and-Aspartate-Metabolism	[313]
Androgen-and-Estrogen-Metabolism	[314]
Cell-cycle-G2-M-DNA-Damage-Checkpoint-Regulation	[315]
PPAR-Signaling	[316]
Synaptic-Long-Term-Potentiation	[257, 317]
TGFβ-Signaling	[318]
Pathways significant enriched for GA/SVM500 only	**Alzheimer relevance**
Aryl-Hydrocarbon-Receptor-Signaling	[319]
beta-Alanine-Metabolism	[320]
Glutamate-Metabolism	[321]
Glycerophospholipid-Metabolism	[322, 323]
Methane-Metabolism	—
NRF2-mediated-Oxidative-Stress-Response	[324]
Phenylalanine-Metabolism	[325]
taurine-and-Hypotaurine-Metabolism	[326, 327]
VEGF-Signaling	[328]

Table 5.7: List of pathways that show a significant enrichment in the results of all three feature selection methods (top). List of pathways that show only a significant enrichment in the results found by our GA/SVM500 (bottom). For all pathways we add references that show the relevance of the pathways to Alzheimer's disease.

usable for the identification of Alzheimer-associated genes and thereforee for the selection of potential biomarkers.

For the three gene lists 'KEGG AD' [205], 'Genotator' [310] and 'AlzGene' [309] we find a significant enrichment in the gene sets selected by information gain and random forest. We also observe enrichment in the resulting sets of the GA/SVM500 but it is not considered to be significant. Whereas 'KEGG AD' contains 150 genes of a small regulatory network for Alzheimer, 'Genotator' and 'AlzGene' are expert-supported collections of Alzheimer-related genes. As the wider periphery is not yet well investigated, the three applied lists contain only genes directly associated with Alzheimer. We expect the genes contained in these lists to be highly redundant. The fact that the gene sets selected by GA/SVM500 show a significant enrichment in a gene collection including a widespread periphery whereas the enrichment found in more specific gene lists is not significant supports the assumption that the algorithm primarily selects genes of the periphery of the disease. A reason for this is the strict elimination of redundant genes discussed in

Chapter 4 as well as the close relation of the genes in the analyzed Alzheimer networks.

At about the same time we were working on the identification of Alzheimer biomarkers, Soler et al. [311] and Goni et al. [312] also used microarray data analysis to examine underlying mechanisms in Alzheimer diseased brain. They published a number of genes they identified to play a role in this processes. This experimental data is not yet fully verified but they include new genes with a possible relation to Alzheimer disease. For this reason we also use this data as reference for our enrichment analysis. For the list of Soler et al. all three methods show a significant enrichment. For the list of Goni et al. (brain) we only find such a significant enrichment in the sets selected by the GA/SVM500. This suggests that the algorithm is well suited to find new potential biomarkers for Alzheimer's disease.

Furthermore, we use a list of genes identified as Alzheimer biomarkers in blood. As we work on microarray samples of brain tissue we do not expect enrichment in the blood-related biomarker set. For this reason, none of the algorithms showing enrichment in the list of Goni et al. (blood) fulfills our expectations.

To strengthen our hypothesis that the GA/SVM500 finds genes that play a role in the periphery of the regulatory Alzheimer network, we also perform a gene set enrichment analysis on 190 pathways provided by Ingenuity Systems. In Table 5.7 we list the pathways showing enrichment for the gene sets of all three methods as well as the pathways showing enrichment for the GA/SVM500 selected gene sets only. For all of these pathways we find evidences in literature that support a link to Alzheimer's disease. The most interesting of the enriched pathways is the taurine-hypotaurine metabolism, because the protein taurine (TAU) forms the so called tangles that can be found in brains of Alzheimer diseased patients. Even if this pathway plays an important role in Alzheimer's diseases it is only enriched in gene sets found by our GA/SVM500. Similar to our results for the PLURI data set (see Section 5.2.1) we observe enrichment in many important signaling pathways. This suggests that the gene selected by our GA/SVM cover many processes associated with Alzheimer and underlines the biological relevance of the results of our GA/SVM500.

Recent work

In the AD data set*LOC642711* is the gene highest ranked by GA/SVM500. It is selected in about 86% of the resulting small gene sets, what is a large frequency. The runner up genes *PRKXP1* and *LOC283345* appear in more than 50% of the small sets.

Formerly, it was assumed that *LOC642711* is similar to *JMY* (junction-mediating and regulatory protein). However, the sequence status in the NCBI Reference Sequence (RefSeq) database [329] is currently withdrawn. But even if this chromosomal sequence officially not corresponds to a gene, the microarray data shows that at least parts of the sequence are expressed and we observe a different expression level in healthy and Alzheimer affected brain tissues.

The former gene *LOC642711* is allocated on the chromosome 15 and contains 29,364 bases. An analysis reveals that about two third of the DNA sequence consists of long interspersed nuclear elements (LINE) and long terminal repeat elements (LTR). The remaining part with fewer repeats is well conserved in rhesus and some other mammalian. We find the mRNA AK125576 coded on this part of the sequence what underlines the probably transcription of the chromosomal sequence [330].

The second highest rated gene by our GA/SVM500 is the pseudogene *PRKXP1* (protein kinase, X-linked, pseudogene 1). For the originated gene *PRKX* there are orthologous genes in mouse (*Pkare*) and drosophila (*PKA-C3*). The mouse orthologous gene is expressed in many parts of the central nervous system [331] and in drosophila *PKA-C3* plays an important role in progressive neurodegeneration [332]. In human *PRKX* is known to be highly expressed in fetal brain [333].

The third gene in the list is *LOC283345*, better known as *RPL13P5* (ribosomal protein L13 pseudogene 5). There are 13 processed pseudogenes of *RPL13* listed by Zhang et al [334]. The pseudogene is differentially expressed regarding Alzheimer's disease but its function is still unknown. Interestingly, the parent gene *RPL13* is one of the best housekeeper genes in Alzheimer brain [335].

We could none of the three highest ranked genes directly associate with Alzheimer's disease. The same is also true for the genes ranked top three by the other two feature selection methods, information gain and random forest. In fact random forest and GA/SVM500 share two of the three genes and besides *LOC642711* and *LOC283345* there is also no evidences for an involvement of *FLJ11903* in the mechanisms of Alzheimer. Also for the genes identified by information gain, *FLJ11903*, *TNCRNA* (nuclear paraspeckle assembly transcript 1) and *LOC283755*, there is no proven Alzheimer relation.

Recently, M.Squillario and A.Barla identified *PRKXP1* and *TNCRNA* as two of 39 genes possibly involved in the underlying processes of Alzheimer's disease [336]. This supports our assumption that the selected genes play an important role even if we cannot verify this proposition at the moment.

An interesting fact is that all three algorithms highly consider transcribes of non-protein-coding areas. Only the gene *FLJ11903* encodes a protein simi-

lar to the hypothetical protein MGC40405. The dysfunctional copies of genes, known as pseudogenes, are characterized by losing their protein-coding ability [337, 338]. Although, it is estimated that 2% to 20% of the pseudogenes are transcribed [339–342]. It has been shown that the expression level of some pseudogenes varies depending on the physiological conditions for example in diseased cells. Among others, those pseudogenes have been identified for diabetes [343] and cancer [344–346]. This indicates that the pseudogenes identified by our three algorithms are possibly usable as biomarkers for Alzheimer. Furthermore it is known that long non-coding RNA (ncRNA) sequences play a role in alternative splicing and many other regulatory processes [347–349]. So it is also thinkable that the identified pseudogenes encode such regulatory RNAs, what has to be investigated more precisely.

5.5 Focusing on our data sets

Besides the identification of potential biomarkers and their biological relevance we have to consider two other aspects of the resulting gene lists presented in Table 5.1 and Table 5.5. We use three different feature selection methods for our analysis. Although, we use the same two data sets for all methods, a comparison of the resulting gene lists shows only a small overlap. A second interesting point is that genes generally known as biomarkers for pluripotency and Alzheimer's disease are strongly under-represented in the obtained gene lists.

Furthermore we compare the two data sets with each other based on the results of the gene set enrichment analysis and the overlap observed in the lists selected by different feature selection methods.

5.5.1 Problems in gene expression data analysis

Microarray data are widely used for the identification of differentially expressed genes and biomarkers. Nevertheless, we observe some problems commonly known. Even so the results supplied by microarray data analyses often provide a good edge for further investigations and biological experiments.

Small overlap

In 2006, Jeffrey et al. analyze the intersections between gene lists found by 10 different feature selection methods. Across these lists they observe an overlap of only 8% to 21% in a list of 100 genes [89].

Using pairs of lists containing the 100 highest ranked genes, we observe an overlap of 24% to 68% on the PLURI data set and 19% to 36% on the AD data

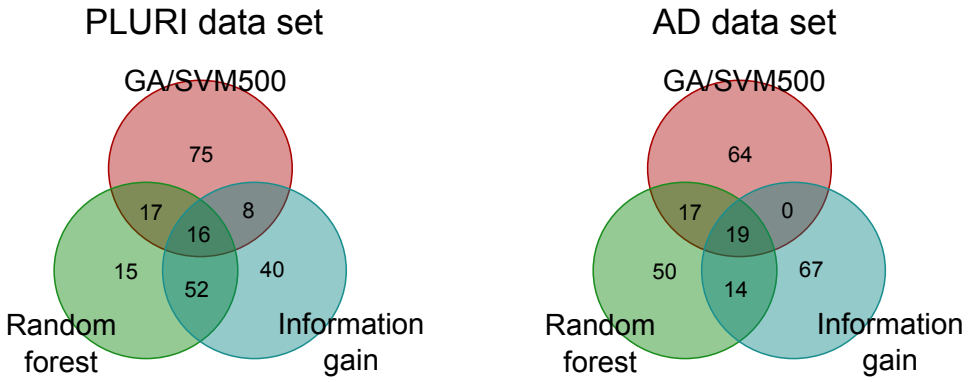


Figure 5.4: Overlap of the gene lists containing the 100 top-ranked genes selected by information gain, random forest and GA/SVM500.

set. This is at least twice as much as the expectation value for completely independent list. So even if an overlap of 19% is not that large, we observe the gene lists selected by the different algorithms have more similarities as randomly selected lists.

For the PLURI data set the 68% of the genes selected by information gain and random forest overlap. Second most similar are the gene lists found by our GA/SVM500 and random forest with an overlap of 33%. The lists of information gain and GA/SVM500 show the smallest overlap with only 24%. For the AD data set there is nearly no difference between the overlap of GA/SVM500/random forest and random forest/information gain gene lists. This time the lists of GA/SVM500 and random forest are slightly more similar (36%) than the lists of random forest and information gain (33%). Again the smallest overlap we observe between GA/SVM500 and information gain. Figure 5.4 displays the overlap of the genes selected by random forest, information gain and GA/SVM500.

As we discuss in Chapter 4, the functioning of the three feature selection methods is very different. Especially, GA/SVM differs from the other two methods regarding the elimination of redundant genes. For this reason, we expect larger overlap between information gain and random forest than between these two methods and our GA/SVM.

Reproducibility

It is also known that the reproducibility of the results is very low [350, 351]. The results mainly depend on the used data set and using other data sets for the same research issue mostly leads to new results. As it was challenging to find large data sets we analyze only a single data set for each of the

	Canonical genes			Genetic algorithm		
	Pou5f1	*Sox2*	*Nanog*	*Fam134b*	*Pam*	*Dub1*
non-pluripotent	6.02	6.43	4.67	7.38	5.82	4.19
Pluripotent	11.71	11.59	10.91	5.58	4.39	5.69
Fold change	5.69	5.16	6.24	1.79	1.43	1.5

Table 5.8: Fold change of the three well known biomarkers *Pou5f1*, *Sox2* and *Nanog* as well as the three genes *Fam134b*, *Pam* and *Dub1* top ranked by our GA/SVM500.

two biological aspects. For this reason we cannot make any proposition concerning this problem based on completely independent data sets.

Although, we have the possibility to compare the resulting list of each algorithm using only parts of the data set. We compare the lists of the top 20 genes (Table 5.1 and 5.5) to lists selected by the same algorithm using only two third of the samples. We only observe an overlap of around 72% on the PLURI data set and even 42% on the AD data set. In general this overlap is independent of the method we use.

Lack of canonical biomarkers

For pluripotency [352, 353] as well as for Alzheimer's diseases [312] many microarray studies result in lists or networks lacking the most canonical genes. In both data sets we found some generally known genes under the top 20 selected genes for example *Dazl* [232], *Gdf3* [235,236] and *Mybl2* [237] (PLURI data set) or *SST* [309], *BFKBIA* [309] and *EP300* [310] (AD data set). Even so we do not find any of the most recent biomarkers in any of the two data sets.

In the PLURI data set we examine the three genes *Pou5f1*, *Sox2* and *Nanog* as the core proteins for pluripotency as described in Section 5.1.2. Table 5.8 shows the fold changes of this genes. For comparison we also list the fold changes of the first three genes found by the GA/SVM500. As we see the fold changes of the canonical genes are much higher than the fold changes of the selected genes. Even so, we assume that this genes do not increase the classification power by adding them to the small gene sets selected by GA/SVM500. Our hypothesis is that those genes are particularly redundant to other selected genes or combinations of genes. In about 70% of the small lists selected by the GA/SVM500 we observe at least one of the established pluripotency genes [168–170]. Furthermore we assume that each of the small sets includes at least one general indicator for pluripotency of which some are not yet included into the popular regulatory networks.

For the AD data set the explanation for not finding the general known biomarkers is much easier. The genes *APOE*, *PSEN1* and *PSEN2* are not differentially expressed in our samples of Alzheimer's disease affected brain and the healthy control group. Before feature selection, we filter the data set in a preprocessing step (see Section 3.2.2) and we use only the 1,000 most promising genes for our analysis. The best ranked of these genes, based on t-test and fold change, is *PSEN2* at position 2,807. The other genes perform even worse. As the genes listed above are manly related to the familiar form of Alzheimer the low fold change is expectably.

5.5.2 Comparison of our data sets

In the first part of the data set comparison in Section 4.4 we focused on the classification performance of different feature selection and classification methods. In this section we compare different aspects of the biological relevance of the results obtained from the two data sets.

Gene set enrichment analysis and examination of single genes

The most obvious differences between the data sets are the results of the gene set enrichment analysis. In both data sets information gain and random forest find larger enrichments in known gene lists than our wrapper of genetic algorithm and support vector machine (GA/SVM500). Although, the results for the PLURI data set is much better than for the AD data set. Additionally, in literature we find more evidences for a participation of our selected genes in pluripotency processes than in Alzheimer's disease.

There are two possible hypotheses for this observation. First, genes responsible for the mechanisms in Alzheimer's disease are more difficult to identify resulting in gene lists containing several unknown genes. The fact that the classification accuracies obtained on the AD data set are much lower than the accuracies observed on the PLURI data set supports this hypothesis (discussed in Section 4.4). But even if the classification performance on the AD data set is not as good as on the PLURI data set, the best possible genes of the data set are selected by the feature selection algorithms. For this reason, a second hypothesis explaining the results of the gene set enrichment analysis as well as the number of evidences found in recent work is the amount of already verified knowledge. Meanwhile we understand much more about the mechanisms of pluripotency than about Alzheimer. So the number of genes with a completely unknown relationship to pluripotency is much lower than the number of genes with an unknown role in Alzheimer. That the gene regulatory networks for pluripotency [168–170] are much more

expanded than Alzheimer nets [205] supports the second hypothesis. As a consequence, the GA/SVM500 that is supposed to find more peripheral genes than the other two methods show worse results in gene set enrichment analysis and find fewer links to other publications.

Overlap

A second point is the overlap between the obtained gene lists. Concerning the gene lists selected by our GA/SVM500 the overlap with the lists of information gain and random forest is nearly identical on both data sets. Looking at the overlap between the top 100 genes of random forest and information gain we observe an overlap of 68% on the PLURI data set and only 33% on the AD data set. The difference is even larger when we refer to the top 20 genes. There we observe an overlap of 80% on the PLURI data set and only 45% on the AD data set. There are several possible reasons for the variations in the number of common genes. As one possibility we consider the better learnability of the PLURI data set that is supported by the observed classification accuracies, discussed in Section 4.4. Another possible reason is different complexity of Alzheimer and pluripotency networks. We also consider the size of the two data sets as a reason for the varying overlap between the two data sets.

To clear up this question future work should concentrate on additional data sets, including new data sets for pluripotency and Alzheimer.

Chapter 6

Conclusion

The automated identification of biomarkers for all kinds of biological and medical problems is an important issue in the field of bioinformatics. Biomarkers are able to reflect normal biological or pathological processes during different cell states. They allow the classification of different tissues and extend our knowledge about underlying molecular mechanisms.

Gene expression is a fundamental part of the regulatory processes inside a cell. So, using the expression level of genes as biomarkers for cellular states is a promising approach, especially, considering the large quantity of freely available microarray data.

We apply three promising feature selection methods to two microarray data sets. This way, we identify a number of genes that are suitable biomarker candidates for pluripotency and Alzheimer's disease. Many of the identified genes are already known to play an important role in the two biological states. Our results also support the relevance of some biomarker candidates, that are part of current research, for example *Dub1* for pluripotency or *SST* for Alzheimer's disease. Beside the genes that are under investigation as potential biomarkers we find completely new candidates. We expect that biological experiments will validate some of the new identified genes as biomarkers.

The most interesting biomarker candidate for pluripotency is the gene *Fam134b*. From all what has become known about the function of *Fam134b* we assume a close relation to pluripotency. So, we suggest a further investigation of this gene and its products in biological experiments.

For Alzheimer's disease, we observe a range of pseudogenes as the most promising candidates. These relatives of genes that lose their protein coding function may play a role in regulating protein-coding transcripts and other regulatory processes. Our results support a more exact investigation of various pseudogenes regarding their role in Alzheimer's disease.

Biomarker detection from microarray data is not limited to pluripotency

and Alzheimer's disease, for example *PRKXP1* or *RPL13P5*. For this reason, we are also interested in the evaluation of the used feature selection methods. We compared three different feature selection algorithms with each other. Information gain and random forest are very popular and widely used techniques for identifying interesting genes from microarray data. The wrapper of genetic algorithm and support vector machine (GA/SVM) is less popular but also showed good results in biomarker selection for cancer data as well as in other applications. For comparison we focus on the classification capability of the selected gene subsets. Even if all three methods show reasonable results on both data sets our GA/SVM exceed information gain and random forest.

To further improve the quality of the selected biomarkers multiple parameters can be trained. Especially for the GA/SVM we suggest an inclusion of prior knowledge of gene interactions. This can be realized by implementing the internal solution representation of the GA/SVM in a tree structure based on the gene distances in regulatory networks or gene ontology similarities.

Supported by our results the GA/SVM takes a special place among the used feature selection methods. Besides a ranked list of potential biomarker the algorithm selects a number of small sets of genes that are together well suited for classification. These small sets are the real strength of the GA/SVM. We expect that genes selected together in a small set fulfill different essential functions for the specific biological state. For this reason analyzing the small gene sets is a promising approach for understanding the complex mechanisms of different biological states. Because of the random nature of the GA/SVM the small gene sets differ a lot.

As a validation of the small sets selected by the GA/SVM in biological experiments is challenging we concentrate on gene pairs occurring in the small sets more often or less often than expected. The classification capability of the genes often selected together by the GA/SVM is much higher than the classification capability of genes rarely selected together. So we show that the GA/SVM prefers gene combinations that increase the accuracy of a classifier when used for training. Validating the role of those gene pairs in biological experiments is challenging as additional interactions cannot be excluded. Even so, our results give useful indications for further biological research.

Besides the investigation of gene pairs we expect a deeper insight into the processes of pluripotency and Alzheimer's disease from analyzing combinations of three or more genes. We expect the information gained from a large number of varying small gene sets is useful for a future modeling of regulatory networks.

Besides the identification of new biomarker candidates for pluripotency and Alzheimer's disease we show the strength of an algorithm which has so

far attracted less attention.

Appendix A

Appendix

A.1 All GEO data series forming the PLURI data set

record	description	label	
		+	-
GSE3653	Inducible Ngn3 Embryonic Stem Cells [354]	2	2
GSE4189	The Oct4 and Nanog transcription network that regulates pluripotency in mouse embryonic stem cells [355]	5	0
GSE4309	An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis [356]	36	0
GSE6933	Unique Molecular Signature of Multipotent Adult Progenitor Cells [357]	3	3
GSE7688	Genome-wide mapping and analysis of active promoters in mouse ES cells and adult organs [358]	1	4
GSE8024	Murine ES cells, neural precursor cells and embryonic fibroblasts [359]	3	5
GSE9244	WT vs Klf5 KO ES [360]	2	0
GSE9563	Translation state array analysis of Mouse embryonic stem cells and embryoid bodies [361]	15	15
GSE9954	Large-scale analysis of the mouse transcriptome [362]	3	67
GSE10573	Superseries Endoh2008 PcG Pou5f1 [363]	4	0

Continued on next page

record	description	label	
		+	-
GSE10610	Expression data from mouse germline stem (GS), multipotent germline stem (mGS), and embryonic stem (ES) cells. [364]	1	1
GSE10776	Expression profiles of Embryonic stem cells derived from normal fertilization and parthenogenesis	6	0
GSE10806	Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors [365]	9	2
GSE10871	Differentiated, partially- and fully-reprogrammed MEFs/B-cells [366]	4	16
GSE10970	Efficient Array-based Identification of Novel Cardiac Genes through Differentiation of Mouse ESCs [367]	3	9
GSE11274	Induction of Pluripotency in Adult Unipotent Germline Stem Cells [368]	8	9
GSE12499	Oct4-Induced Pluripotency in Adult Neural Stem Cells [369]	3	7
GSE13408	Cell cycle exit and terminal differentiation independent of the Rb gene family during embryonic development	3	0
GSE13805	Expression data from wild type and calreticulin deficient murine embryonic stem cells. [370]	2	0
GSE15173	Dppa4 is dispensable for embryonic stem cell identity and germ cell development, but essential for embryogenesis [371]	3	0
GSE15267	Expression data of induced pluripotent stem cell	6	0
GSE17879	Activin/Nodal signaling in mouse embryonic stem cells [372]	2	0
GSE19076	Expression data from wild type, Ring1B-/-, Eed-/-, and Ring1B/Eed double deficient mouse ES cells [373]	3	0
GSE20527	Effect of BMP4 and noggin on gene expression in murine R1 ES cells [374]	1	0

Continued on next page

record	description	label	
		+	-
GSE21515	Expression data from mouse ES and iPS cells [375]	18	0

Table A.1: PLURI data set: Selection of GEO data series and the label of the containing data sets (+: pluripotent samples, -: non-pluripotent samples).

A.2 Partitioning of the PLURI data set

Fold 1	Fold 2	Fold 3
GSM144634	GSM144622	GSM144635
GSM185509	GSM144623	GSM198065
GSM185511	GSM144624	GSM234773
GSM185513	GSM144636	GSM241851
GSM198062	GSM185510	GSM241852
GSM198064	GSM185512	GSM241856
GSM198066	GSM198063	GSM241858
GSM198070	GSM198067	GSM241863
GSM241853	GSM198072	GSM241865
GSM241857	GSM234772	GSM241869
GSM241859	GSM241847	GSM241873
GSM241860	GSM241848	GSM241876
GSM241861	GSM241849	GSM252064
GSM241862	GSM241850	GSM252068
GSM241864	GSM241854	GSM252069
GSM241868	GSM241855	GSM252072
GSM241870	GSM241866	GSM252073
GSM241872	GSM241867	GSM252077
GSM241874	GSM241871	GSM252082
GSM241875	GSM252066	GSM252092
GSM252065	GSM252067	GSM252093
GSM252070	GSM252075	GSM252094
GSM252071	GSM252076	GSM252095
GSM252074	GSM252078	GSM252104
GSM252080	GSM252079	GSM252107
GSM252086	GSM252081	GSM252109
GSM252087	GSM252083	GSM252112
GSM252090	GSM252084	GSM252115
GSM252096	GSM252085	GSM252116
GSM252098	GSM252088	GSM252125
GSM252103	GSM252089	GSM252129
GSM252106	GSM252091	GSM252133
GSM252110	GSM252097	GSM265042
GSM252114	GSM252099	GSM272035
GSM252117	GSM252100	GSM272036
GSM252118	GSM252101	GSM272753

Continued on next page

Fold 1	Fold 2	Fold 3
GSM252120	GSM252102	GSM272839
GSM252121	GSM252105	GSM272847
GSM252122	GSM252108	GSM272890
GSM252127	GSM252111	GSM275554
GSM252130	GSM252113	GSM275555
GSM265040	GSM252119	GSM275556
GSM267415	GSM252123	GSM275557
GSM272037	GSM252124	GSM275564
GSM272052	GSM252126	GSM275567
GSM272053	GSM252128	GSM277761
GSM272054	GSM252131	GSM277765
GSM272846	GSM252132	GSM277768
GSM275544	GSM266065	GSM279200
GSM275546	GSM266837	GSM284790
GSM275551	GSM267413	GSM284793
GSM275552	GSM272836	GSM284794
GSM275558	GSM272837	GSM284797
GSM275560	GSM272848	GSM284806
GSM275562	GSM275547	GSM314040
GSM275563	GSM275548	GSM314042
GSM275581	GSM275561	GSM314043
GSM277759	GSM275566	GSM314044
GSM277760	GSM275580	GSM314046
GSM277763	GSM277757	GSM314048
GSM277764	GSM277758	GSM338371
GSM277766	GSM277762	GSM347150
GSM284788	GSM277767	GSM381301
GSM284791	GSM279201	GSM381304
GSM284795	GSM279202	GSM381305
GSM284796	GSM284789	GSM472235
GSM284799	GSM284792	GSM515594
GSM284801	GSM284798	GSM537474
GSM284805	GSM284800	GSM537475
GSM314039	GSM284807	GSM537477
GSM314045	GSM314038	GSM537478
GSM314047	GSM338369	GSM537481
GSM378798	GSM338373	GSM537488
GSM381308	GSM347151	GSM537489
GSM446646	GSM378796	GSM537490

Continued on next page

Fold 1	Fold 2	Fold 3
GSM472236	GSM378797	GSM85006
GSM537476	GSM381306	GSM85008
GSM537479	GSM381307	GSM85009
GSM537482	GSM446645	GSM94859
GSM537484	GSM472237	GSM98548
GSM537485	GSM537473	GSM98550
GSM537486	GSM537480	GSM98553
GSM94856	GSM537483	GSM98554
GSM94857	GSM537487	GSM98555
GSM94858	GSM85007	GSM98556
GSM98549	GSM94860	GSM98558
GSM98551	GSM98557	GSM98561
GSM98552	GSM98559	GSM98564
GSM98560	GSM98563	GSM98569
GSM98562	GSM98566	GSM98570
GSM98565	GSM98567	GSM98573
GSM98568	GSM98571	GSM98575
GSM98572	GSM98578	GSM98576
GSM98574	GSM98579	GSM98577
GSM98582	GSM98580	GSM98581
GSM98583		

Table A.2: PLURI data set: Partitioning of the GEO data sets into three subsets (folds).

A.3 Partitioning of the AD data set

Fold 1	Fold 2	Fold 3
GSM119617	GSM119615	GSM119619
GSM119618	GSM119616	GSM119620
GSM119621	GSM119622	GSM119623
GSM119626	GSM119624	GSM119625
GSM119627	GSM119628	GSM119631
GSM119629	GSM119632	GSM119633
GSM119630	GSM119635	GSM119636
GSM119634	GSM119637	GSM119638
GSM119639	GSM119641	GSM119643
GSM119640	GSM119644	GSM119645
GSM119642	GSM119646	GSM119647
GSM119651	GSM119648	GSM119649
GSM119652	GSM119653	GSM119650
GSM119654	GSM119656	GSM119655
GSM119658	GSM119659	GSM119657
GSM119664	GSM119661	GSM119660
GSM119665	GSM119667	GSM119662
GSM119673	GSM119671	GSM119663
GSM119675	GSM119674	GSM119666
GSM119676	GSM119677	GSM119668
GSM119678	GSM119679	GSM119669
GSM119681	GSM119682	GSM119670
GSM119687	GSM119684	GSM119672
GSM119688	GSM238790	GSM119680
GSM238791	GSM238797	GSM119683
GSM238794	GSM238798	GSM119685
GSM238795	GSM238801	GSM119686
GSM238799	GSM238803	GSM238763
GSM238802	GSM238804	GSM238792
GSM238807	GSM238806	GSM238793
GSM238808	GSM238812	GSM238796
GSM238810	GSM238815	GSM238800
GSM238811	GSM238817	GSM238805
GSM238813	GSM238820	GSM238809
GSM238816	GSM238823	GSM238821
GSM238818	GSM238835	GSM238822

Continued on next page

Fold 1	Fold 2	Fold 3
GSM238819	GSM238839	GSM238825
GSM238824	GSM238841	GSM238826
GSM238827	GSM238845	GSM238838
GSM238834	GSM238847	GSM238840
GSM238837	GSM238851	GSM238842
GSM238843	GSM238855	GSM238856
GSM238844	GSM238857	GSM238858
GSM238846	GSM238860	GSM238861
GSM238848	GSM238865	GSM238863
GSM238854	GSM238867	GSM238864
GSM238862	GSM238874	GSM238868
GSM238870	GSM238877	GSM238871
GSM238873	GSM238942	GSM238872
GSM238875	GSM238944	GSM238948
GSM238941	GSM238945	GSM238949
GSM238943	GSM238947	GSM238951
GSM238946	GSM238952	GSM238953
GSM238955		GSM238963

Table A.3: AD data set: Partitioning of the GEO data sets into three subsets (folds).

Bibliography

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, December 2003.

[2] S. K. Singh, C. Hawkins, I. D. Clarke, J. A. Squire, J. Bayani, T. Hide, R. M. Henkelman, M. D. Cusimano, and P. B. Dirks. Identification of human brain tumour initiating cells. *Nature*, 432:396–401, Nov 2004.

[3] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.*, 69:89–95, Mar 2001.

[4] B. Weigelt, F. L. Baehner, and J. S. Reis-Filho. The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J. Pathol.*, 220:263–280, Jan 2010.

[5] Ian Jeffery, Desmond Higgins, and Aedin Culhane. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7(1):359, 2006.

[6] Iffat A Gheyas and Leslie S Smith. Feature subset selection in large dimensionality domains. *Pattern Recognition*, 43(1):5–13, 2010.

[7] O. Lindvall and Z. Kokaia. Stem cells for the treatment of neurological disorders. *Nature*, 441:1094–1096, Jun 2006.

[8] B. E. Strauer, C. M. Schannwell, and M. Brehm. Therapeutic potentials of stem cells in cardiac diseases. *Minerva Cardioangiol*, 57:249–267, Apr 2009.

[9] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi. Forecasting the global burden of Alzheimer's disease. *Alzheimers Dement*, 3:186–191, Jul 2007.

[10] C. P. Ferri, M. Prince, C. Brayne, H. Brodaty, L. Fratiglioni, M. Ganguli, K. Hall, K. Hasegawa, H. Hendrie, Y. Huang, A. Jorm, C. Mathers, P. R. Menezes, E. Rimmer, and M. Scazufca. Global prevalence of dementia: a Delphi consensus study. *Lancet*, 366:2112–2117, Dec 2005.

[11] L. Scheubert, R. Schmidt, D. Repsilber, M. Lustrek, and G. Fuellen. Learning biomarkers of pluripotent stem cells in mouse. *DNA Res.*, 18:233–251, 2011.

[12] X. Cui and G. A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, 4:210, 2003.

[13] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18:546–554, Apr 2002.

[14] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011.

[15] *Matlab: Version 7.10.0*. The MathWorks Inc., Natick, Massachusetts, 2010.

[16] *Mathematica: Version 8.0*. Wolfram Research, Inc., Champaign, IL, USA, 2010.

[17] Aidong Zhang. *Advanced analysis of gene expression microarray data*. World Scientific Publishing Co. Pte. Ltd., 2006.

[18] Lei Guo, Edward K. Lobenhofer, Charles Wang, Richard Shippy, Stephen C. Harris, Lu Zhang, Nan Mei, Tao Chen, Damir Herman, Federico M. Goodsaid, Patrick Hurban, Kenneth L. Phillips, Jun Xu, Xutao Deng, Yongming Andrew A. Sun, Weida Tong, Yvonne P. Dragan, and Leming Shi. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature biotechnology*, 24(9):1162–1169, September 2006.

[19] Sung Choe, Michael Boutros, Alan Michelson, George Church, and Marc Halfon. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6(2):R16+, 2005.

[20] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.*, 98:5116–5121, Apr 2001.

[21] Rainer Opgen-Rhein, Verena Zuber, and Korbinian Strimmer. *Package 'st' - Shrinkage t Statistic and CAT Score*, 2011.

[22] Joachim Hartung, Baerbel Elpelt, and Karl-Heinz Kloesener. *Statistik*. R. Oldenburg Verlag, 15th edition, 2002.

[23] Bernhard Rüger. *Test-und Schätztheorie*, volume 2. R. Oldenburg Verlag, 2002.

[24] Karl Bosch. *Statistik-Taschenbuch*. R. Oldenburg Verlag, 3rd edition, 1998.

[25] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3rd edition, 2007.

[26] B. L. Welch. The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1/2):28–35, 1947.

[27] D. M. Smith W. N. Venables and the R Development Core Team. *An Introduction to R*, 2011.

[28] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

[29] Robert Tibshirani. A comparison of fold-change and the t-statistic for microarray data analysis. *Analysis*, pages 2–3, 2007.

[30] Alan Dabney and John Storey. *Bioconductor's qvalue package*, 2011.

[31] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.

[32] Claude E. Shannon, Warren Weaver, and Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, September 1998.

[33] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition*. Wiley-Interscience, 2nd edition, July 2006.

[34] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, August 2005.

[35] Gabriele Sales and Chiara Romualdi. parmigene - a parallel R package for mutual information estimation and gene network reconstruction. *Bioinformatics*, 27(13):1876–1877, July 2011.

[36] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.

[37] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, 31(4):e15+, February 2003.

[38] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, April 2003.

[39] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, January 2003.

[40] Cheng Li and Wing H. Wong. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, 2(8):research0032.1–research0032.11, 2001.

[41] Frederick Mosteller and John W. Tukey. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley Series in Behavioral Science. Addison Wesley, 1st edition, January 1977.

[42] *Affymetrix Power Tools (APT) Software Package*, 2010.

[43] C. Backes, A. Keller, J. Kuentzer, B. Kneissl, N. Comtesse, Y. A. Elnakady, R. Müller, E. Meese, and H. P. Lenhof. GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Research*, 35(Web Server issue), July 2007.

[44] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: an update. *Special Interest Group on Knowledge Discovery and Data Mining Explorer Newsletter*, 11(1):10–18, November 2009.

[45] G. J. McLachlan, K. Do, and C Ambroise. *Analyzing Microarray Gene Expression Data*. Wiley-Interscience, 2004.

[46] J. Ross Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[47] David W. Aha and Dennis Kibler. Instance-based learning algorithms. In *Machine Learning*, pages 37–66, 1991.

[48] Michel M. Deza and Elena Deza. *Encyclopedia of Distances*. Springer, 1st edition, August 2009.

[49] R. De Maesschalck. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, January 2000.

[50] Gang Qian, Shamik Sural, Yuelong Gu, and Sakti Pramanik. Similarity between euclidean and cosine angle distance for nearest neighbor queries. In *Proceedings of the 2004 ACM symposium on Applied computing*, SAC '04, pages 1232–1237. ACM, 2004.

[51] George John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann, 1995.

[52] Nello Cristianini and John Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, New York, NY, USA, 2000.

[53] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Mach. Learn.*, 20(3):273–297, September 1995.

[54] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[55] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[56] David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, 1st edition, January 1989.

[57] Cheng-Lung Huang and Chieh-Jen Wang. A GA-based feature selection and parameters optimization for support vector machines. *Expert Systems with Applications*, 31(2):231 – 240, 2006.

[58] Gilbert Sywerda. Uniform crossover in genetic algorithms. In *Proceedings of the third international conference on Genetic algorithms*, pages 2–9, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[59] T. M. Therneau and K. V. Ballman. What does PLIER really do? *Cancer Inform*, 6:423–431, 2008.

[60] Affymetrix. *Guide to Probe Logarithmic Intensity Error (PLIER) Estimation*, 2005.

[61] Earl Hubbell, Wei-Min Liu, and Rui Mei. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–1592, December 2002.

[62] Frederick Livingston. Implementation of Breiman's Random Forest Machine Learning Algorithm. In *ECE591Q Machine Learning Journal Paper*, 2005.

[63] V. Fernandez, O. Salamero, B. Espinet, F. Sole, C. Royo, A. Navarro, F. Camacho, S. Bea, E. Hartmann, V. Amador, L. Hernandez, C. Agostinelli, R. L. Sargent, M. Rozman, M. Aymerich, D. Colomer, N. Villamor, S. H. Swerdlow, S. A. Pileri, F. Bosch, M. A. Piris, E. Montserrat, G. Ott, A. Rosenwald, A. Lopez-Guillermo, P. Jares, S. Serrano, and E. Campo. Genomic and gene expression profiling defines indolent forms of mantle cell lymphoma. *Cancer Res.*, 70:1408–1418, Feb 2010.

[64] D. Gresham, M. J. Dunham, and D. Botstein. Comparing whole genomes using DNA microarrays. *Nat. Rev. Genet.*, 9:291–302, Apr 2008.

[65] G. Jonsson, J. Staaf, J. Vallon-Christersson, M. Ringner, K. Holm, C. Hegardt, H. Gunnarsson, R. Fagerholm, C. Strand, B. A. Agnarsson, O. Kilpivaara, L. Luts, P. Heikkila, K. Aittomaki, C. Blomqvist, N. Loman, P. Malmstrom, H. Olsson, O. T. Johannsson, A. Arason, H. Nevanlinna, R. B. Barkardottir, and A. Borg. Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res.*, 12:R42, 2010.

[66] G. F. Hatfull, D. Jacobs-Sera, J. G. Lawrence, W. H. Pope, D. A. Russell, C. C. Ko, R. J. Weber, M. C. Patel, K. L. Germane, R. H. Edgar, N. N. Hoyte, C. A. Bowman, A. T. Tantoco, E. C. Paladin, M. S. Myers, A. L. Smith, M. S. Grace, T. T. Pham, M. B. O'Brien, A. M. Vogelsberger, A. J. Hryckowian, J. L. Wynalek, H. Donis-Keller, M. W. Bogel, C. L. Peebles, S. G. Cresawn, and R. W. Hendrix. Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J. Mol. Biol.*, 397:119–143, Mar 2010.

[67] S. Q. Le and R. Durbin. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.*, 21:952–960, Jun 2011.

[68] N. E. van Bers, K. van Oers, H. H. Kerstens, B. W. Dibbits, R. P. Crooijmans, M. E. Visser, and M. A. Groenen. Genome-wide SNP detection in the great tit Parus major using high throughput sequencing. *Mol. Ecol.*, 19 Suppl 1:89–99, Mar 2010.

[69] X. Q. Xia, Z. Jia, S. Porwollik, F. Long, C. Hoemme, K. Ye, C. Muller-Tidow, M. McClelland, and Y. Wang. Evaluating oligonucleotide properties for DNA microarray probe design. *Nucleic Acids Res.*, 38:e121, Jun 2010.

[70] Y. Deng, Z. He, J. D. Van Nostrand, and J. Zhou. Design and analysis of mismatch probes for long oligonucleotide microarrays. *BMC Genomics*, 9:491, 2008.

[71] S. Lemoine, F. Combes, and S. Le Crom. An evaluation of custom microarray applications: the oligonucleotide design challenge. *Nucleic Acids Res.*, 37:1726–1739, Apr 2009.

[72] X. Li, Z. He, and J. Zhou. Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. *Nucleic Acids Res.*, 33:6114–6123, 2005.

[73] G. Golfier, S. Lemoine, A. van Miltenberg, A. Bendjoudi, J. Rossier, S. Le Crom, and M. C. Potier. Selection of oligonucleotides for whole-genome microarrays with semi-automatic update. *Bioinformatics*, 25:128–129, Jan 2009.

[74] D. M. Smith W. N. Venables and the R Development Core Team. *An Introduction to R*, 2011.

[75] E. Nagele, M. Han, C. Demarshall, B. Belinka, and R. Nagele. Diagnosis of Alzheimer's disease based on disease-specific autoantibody profiles in human sera. *PLoS ONE*, 6:e23112, 2011.

[76] J. A. Webster, J. R. Gibbs, J. Clarke, M. Ray, W. Zhang, P. Holmans, K. Rohrer, A. Zhao, L. Marlowe, M. Kaleem, D. S. McCorquodale, C. Cuello, D. Leung, L. Bryden, P. Nath, V. L. Zismann, K. Joshipura, M. J. Huentelman, D. Hu-Lince, K. D. Coon, D. W. Craig, J. V. Pearson, C. B. Heward, E. M. Reiman, D. Stephan, J. Hardy, and A. J. Myers. Genetic control of human brain transcript expression in Alzheimer disease. *Am. J. Hum. Genet.*, 84:445–458, Apr 2009.

[77] W. S. Liang, T. Dunckley, T. G. Beach, A. Grover, D. Mastroeni, D. G. Walker, R. J. Caselli, W. A. Kukull, D. McKeel, J. C. Morris, C. Hulette, D. Schmechel, G. E. Alexander, E. M. Reiman, J. Rogers, and D. A. Stephan. Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiol. Genomics*, 28:311–322, Feb 2007.

[78] W. S. Liang, E. M. Reiman, J. Valla, T. Dunckley, T. G. Beach, A. Grover, T. L. Niedzielko, L. E. Schneider, D. Mastroeni, R. Caselli, W. Kukull, J. C. Morris, C. M. Hulette, D. Schmechel, J. Rogers, and D. A. Stephan. Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proc. Natl. Acad. Sci. U.S.A.*, 105:4441–4446, Mar 2008.

[79] E. L. Heinzen, W. Yoon, M. E. Weale, A. Sen, N. W. Wood, J. R. Burke, K. A. Welsh-Bohmer, C. M. Hulette, S. M. Sisodiya, and D. B. Goldstein. Alternative ion channel splicing in mesial temporal lobe epilepsy and Alzheimer's disease. *Genome Biol.*, 8:R32, 2007.

[80] Affymetrix. *GeneChip Human Genome U133 Arrays*, 2011.

[81] H. Göhlmann and W. Talloen. *Gene expression studies using affymetrix microarrays*. Chapman and Hall/CRC mathematical & computational biology series. CRC Press, 2009.

[82] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13):1675–1680, December 1996.

[83] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Res.*, 33:D34–38, Jan 2005.

[84] M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev. dbEST–database for expressed sequence tags. *Nat. Genet.*, 4:332–333, Aug 1993.

[85] K. Shakya, H. J. Ruskin, G. Kerr, M. Crane, and J. Becker. Comparison of microarray preprocessing methods. *Adv. Exp. Med. Biol.*, 680:139–147, 2010.

[86] Bettina Harr and Christian Schlötterer. Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Research*, 34(2):e8, 2006.

[87] Roel Verhaak, Frank Staal, Peter Valk, Bob Lowenberg, Marcel Reinders, and Dick de Ridder. The effect of oligonucleotide microarray data preprocessing on the analysis of patient-cohort studies. *BMC Bioinformatics*, 7(1):105, 2006.

[88] David L Wheeler, Deanna M Church, Scott Federhen, Alex E Lash, Thomas L Madden, Joan U Pontius, Gregory D Schuler, Lynn M Schriml, Edwin Sequeira, Tatiana A Tatusova, and Lukas Wagner. Database resources of the National Center for Biotechnology. *Nucleic Acids Res*, 31:28–33, 2003.

[89] I. B. Jeffery, D. G. Higgins, and A. C. Culhane. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7:359, 2006.

[90] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, 7:55–65, Jan 2006.

[91] Youngchao Ge, Sandrine Dudoit, and Terence Speed. Resampling-based multiple testing for microarray data analysis. Technical report, University of California, Berkeley, 2003.

[92] Hardeo Sahai and M.I. Ageel. *The analysis of variance: fixed, random, and mixed models*. Birkhäuser, 2000.

[93] S. Dudoit, M. J. van der Laan, and K. S. Pollard. Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Stat Appl Genet Mol Biol*, 3:Article13, 2004.

[94] A. Statnikov, L. Wang, and C. F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9:319, 2008.

[95] R. Diaz-Uriarte and S. Alvarez de Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3, 2006.

[96] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 1:24–45, 2004.

[97] N. R. Garge, G. P. Page, A. P. Sprague, B. S. Gorman, and D. B. Allison. Reproducible clusters from microarray research: whither? *BMC Bioinformatics*, 6 Suppl 2:S10, Jul 2005.

[98] W. Kong, X. Mou, Q. Liu, Z. Chen, C. R. Vanderburg, J. T. Rogers, and X. Huang. Independent component analysis of Alzheimer's DNA microarray gene expression data. *Mol Neurodegener*, 4:5, 2009.

[99] E. Y. Kim, S. Y. Kim, D. Ashlock, and D. Nam. MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinformatics*, 10:260, 2009.

[100] L. Zhang, Y. Zheng, D. Li, and Y. Zhong. Self-organizing map of gene regulatory networks for cell phenotypes during reprogramming. *Comput. Biol. Chem.*, 35:211–217, Aug 2011.

[101] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, January 2007.

[102] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16:906–914, Oct 2000.

[103] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U.S.A.*, 97:262–267, Jan 2000.

[104] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1-3):389–422, 2002.

[105] Roy Varshavsky, Assaf Gottlieb, Michal Linial, and David Horn. Novel unsupervised feature filtering of biological data. *Bioinformatics*, 22:e507–e513, July 2006.

[106] Jennifer G. Dy and Carla E. Brodley. Feature Selection for Unsupervised Learning. *J. Mach. Learn. Res.*, 5:845–889, 2004.

[107] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 333–342. ACM, 2010.

[108] M. Pirooznia, J. Y. Yang, M. Q. Yang, and Y. Deng. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9 Suppl 1:S13, 2008.

[109] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, October 2007.

[110] H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform*, 13:51–60, 2002.

[111] A. Vergara and E. Llobet. Feature selection versus feature compression in the building of calibration models from FTIR-spectrophotometry datasets. *Talanta*, 88:95–103, Jan 2012.

[112] Yifeng Zeng, Jian Luo, and Shuyuan Lin. Classification using markov blanket for feature selection. In *GrC*, pages 743–747. IEEE, 2009.

[113] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *in ICML*, pages 856–863, 2003.

[114] Chen Liao, Shutao Li, and Zhiyuan Luo. Gene selection using wilcoxon rank sum test and support vector machine for cancer classification. In Yuping Wang, Yiu ming Cheung, and Hailin Liu, editors, *Computational Intelligence and Security, International Conference, CIS 2006, Guangzhou, China, November 3-6, 2006, Revised Selected Papers*, volume 4456 of *Lecture Notes in Computer Science*, pages 57–66. Springer, 2006.

[115] Xin Jin, Anbang Xu, Rongfang Bie, and Ping Guo. Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles. In *Proceedings of the 2006 international conference on Data Mining for Biomedical Applications*, BioDM'06, pages 106–115. Springer-Verlag, 2006.

[116] S. F. Cotter, K. Kreutz-Delgado, and B. D. Rao. Backward sequential elimination for sparse vector subset selection. *Signal Processing*, 81(9):1849–1864, September 2001.

[117] S. Colak and C. Isik. Feature subset selection for blood pressure classification using orthogonal forward selection. In *Bioengineering Conference, 2003 IEEE 29th Annual*, pages 122 – 123. IEEE, 2003.

[118] David Gelbart, Nelson Morgan, and Alexey Tsymbal. Hill-climbing feature selection for multi-stream asr. In *INTERSPEECH*, pages 2967–2970. ISCA, 2009.

[119] Zhong Yan and Chunwei Yuan. Ant colony optimization for feature selection in face recognition. In David Zhang and Anil K. Jain, editors, *Biometric Authentication, First International Conference, ICBA 2004, Hong Kong, China, July 15-17, 2004, Proceedings*, volume 3072 of *Lecture Notes in Computer Science*, pages 221–226. Springer, 2004.

[120] Ronen Meiri and Jacob Zahavi. Using simulated annealing to optimize the feature selection problem in marketing applications. *European Journal of Operational Research*, 171(3):842–858, June 2006.

[121] S. Gerd. *Biomarker*. Schattauer, 2008.

[122] E. J. Moler, M. L. Chow, and I. S. Mian. Analysis of molecular profile data using generative and discriminative methods. *Physiol. Genomics*, 4:109–126, Dec 2000.

[123] M. L. Chow, E. J. Moler, and I. S. Mian. Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol. Genomics*, 5:99–111, Mar 2001.

[124] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, J. Trent, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, S. Gruvberger, N. Loman, O. Johannsson, H. Olsson, and G. Sauter. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, 344:539–548, Feb 2001.

[125] R. Etzioni, C. Kooperberg, M. Pepe, R. Smith, and P. H. Gann. Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics*, 4:523–538, Oct 2003.

[126] W. K. Han, G. Wagener, Y. Zhu, S. Wang, and H. T. Lee. Urinary biomarkers in the early detection of acute kidney injury after cardiac surgery. *Clin. J. Am. Soc. Nephrol.*, 4:873–882, May 2009.

[127] R. M. Pfeiffer and E. Bur. A model free approach to combining biomarkers. *Biom J*, 50:558–570, Aug 2008.

[128] I. J. Kullo and L. T. Cooper. Early identification of cardiovascular risk using genomics and proteomics. *Nat Rev Cardiol*, 7:309–317, Jun 2010.

[129] P. Walsh, M. Elsabbagh, P. Bolton, and I. Singh. In search of biomarkers for autism: scientific, social and ethical challenges. *Nat. Rev. Neurosci.*, 12:603–612, Oct 2011.

[130] M. Xiong, X. Fang, and J. Zhao. Biomarker identification by feature wrappers. *Genome Res.*, 11:1878–1887, Nov 2001.

[131] Thanyaluk Jirapech-Umpai and Stuart Aitken. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, Dec 2004.

[132] Leping Li, Clarice R. Weinberg, Thomas A. Darden, and Lee G. Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method. *Bioinformatics*, 17(12):1131–1142, 2001.

[133] Michal Karzynski, Álvaro Mateos, Javier Herrero, and Joaquín Dopazo. Using a genetic algorithm and a perceptron forfeature selection and supervised class learningin dna microarray data. *Artif. Intell. Rev.*, 20(1-2):39–51, October 2003.

[134] C. H. Ooi and Patrick Tan. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19(1), 2003.

[135] Béatrice Duval and Jin-Kao Hao. Advances in metaheuristics for gene selection and classification of microarray data. *Briefings in Bioinformatics*, 11(1):127–141, January 2010.

[136] Nicoletta Dessì and Barbara Pes. An evolutionary method for combining different feature selection criteria in microarray data classification. *J. Artif. Evol. App.*, pages 3:1–3:10, January 2009.

[137] Shutao Li, Xixian Wu, and Xiaoyan Hu. Gene selection using genetic algorithm and support vectors machines. *Soft Comput.*, 12:693–698, February 2008.

[138] Feng Tan, Xuezheng Fu, Yanqing Zhang, and Anu G. Bourgeois. A genetic algorithm-based method for feature subset selection. *Soft Comput.*, 12:111–120, September 2007.

[139] Edmundo Bonilla Huerta, Beatrice Duval, and Jin kao Hao. A hybrid GA/SVM approach for gene selection and classification of microarray data. In *EvoWorkshops 2006, LNCS 3907*, pages 34–44. Springer, 2006.

[140] J. J. Liu, G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, and X. B. Ling. Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics*, 21(11):2691–2697, June 2005.

[141] Sihua Peng, Qianghua Xu, Xuefeng B. Ling, Xiaoning Peng, Wei Du, and Liangbiao Chen. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Letters*, 555(2):358–362, December 2003.

[142] Sung-Bae Cho and Hong-Hee Won. Machine learning in dna microarray analysis for cancer classification. *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics*, 19:189–198, 2003.

[143] L. Li, W. Jiang, X. Li, K. L. Moser, Z. Guo, L. Du, Q. Wang, E. J. Topol, Q. Wang, and S. Rao. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics*, 85(1):16–23, January 2005.

[144] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U.S.A.*, 98:15149–15154, Dec 2001.

[145] Zhanchao Li, Xuan Zhou, Zong Dai, and Xiaoyong Zou. Classification of g-protein coupled receptors based on support vector machine with maximum relevance minimum redundancy and genetic algorithm. *BMC Bioinformatics*, 11:325, 2010.

[146] E. Pourbasheer, S. Riahi, M. R. Ganjali, and P. Norouzi. Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity. *Eur. J. Med. Chem.*, 44:5023–5028, Dec 2009.

[147] M. S. Mohamad, S. Omatu, S. Deris, M. F. Misman, and M. Yoshioka. A multi-objective strategy in genetic algorithms for gene selection of gene expression data. *Artificial Life and Robotics*, 13(2):410–413, March 2009.

[148] Justin Healey, editor. *Stem Cell Research*, volume 178 of *Issues in Society*. Spinneypress, 2003.

[149] K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126:663–676, Aug 2006.

[150] M. Nakagawa, M. Koyanagi, K. Tanabe, K. Takahashi, T. Ichisaka, T. Aoi, K. Okita, Y. Mochiduki, N. Takizawa, and S. Yamanaka. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat. Biotechnol.*, 26:101–106, Jan 2008.

[151] J. B. Kim, V. Sebastiano, G. Wu, M. J. Arauzo-Bravo, P. Sasse, L. Gentile, K. Ko, D. Ruau, M. Ehrich, D. van den Boom, J. Meyer, K. Hubner, C. Bernemann, C. Ortmeier, M. Zenke, B. K. Fleischmann, H. Zaehres, and H. R. Scholer. Oct4-induced pluripotency in adult neural stem cells. *Cell*, 136:411–419, Feb 2009.

[152] S. P. Medvedev, A. I. Shevchenko, N. A. Mazurok, and S. M. Zakiian. OCT4 and NANOG are the key genes in the system of pluripotency maintenance in mammalian cells. *Genetika*, 44:1589–1608, Dec 2008.

[153] I. Chambers, D. Colby, M. Robertson, J. Nichols, S. Lee, S. Tweedie, and A. Smith. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell*, 113:643–655, May 2003.

[154] D. C. Hay, L. Sutherland, J. Clark, and T. Burdon. Oct-4 knockdown induces similar patterns of endoderm and trophoblast differentiation markers in human and mouse embryonic stem cells. *Stem Cells*, 22:225–235, 2004.

[155] M. M. Matin, J. R. Walsh, P. J. Gokhale, J. S. Draper, A. R. Bahrami, I. Morton, H. D. Moore, and P. W. Andrews. Specific knockdown of Oct4 and beta2-microglobulin expression by RNA interference in human embryonic stem cells and embryonic carcinoma cells. *Stem Cells*, 22:659–668, 2004.

[156] K. Mitsui, Y. Tokuzawa, H. Itoh, K. Segawa, M. Murakami, K. Takahashi, M. Maruyama, M. Maeda, and S. Yamanaka. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell*, 113:631–642, May 2003.

[157] L. A. Boyer, T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122:947–956, Sep 2005.

[158] M. Boiani, S. Eckardt, H. R. Scholer, and K. J. McLaughlin. Oct4 distribution and level in mouse clones: consequences for pluripotency. *Genes Dev.*, 16:1209–1219, May 2002.

[159] H. Niwa, J. Miyazaki, and A. G. Smith. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat. Genet.*, 24:372–376, Apr 2000.

[160] L. H. Looijenga, H. Stoop, H. P. de Leeuw, C. A. de Gouveia Brazao, A. J. Gillis, K. E. van Roozendaal, E. J. van Zoelen, R. F. Weber, K. P. Wolffenbuttel, H. van Dekken, F. Honecker, C. Bokemeyer, E. J. Perlman, D. T. Schneider, J. Kononen, G. Sauter, and J. W. Oosterhuis. POU5F1 (OCT3/4) identifies cells with pluripotent potential in human germ cell tumors. *Cancer Res.*, 63:2244–2250, May 2003.

[161] S. Masui, Y. Nakatake, Y. Toyooka, D. Shimosato, R. Yagi, K. Takahashi, H. Okochi, A. Okuda, R. Matoba, A. A. Sharov, M. S. Ko, and H. Niwa. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nat. Cell Biol.*, 9:625–635, Jun 2007.

[162] Y. Tani, Y. Akiyama, H. Fukamachi, K. Yanagihara, and Y. Yuasa. Transcription factor SOX2 up-regulates stomach-specific pepsinogen A gene expression. *J. Cancer Res. Clin. Oncol.*, 133:263–269, Apr 2007.

[163] Y. H. Loh, Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K. Y. Wong, K. W. Sung, C. W. Lee, X. D. Zhao, K. P. Chiu, L. Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C. L. Wei, Y. Ruan, B. Lim, and H. H. Ng. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, 38:431–440, Apr 2006.

[164] D. J. Rodda, J. L. Chew, L. H. Lim, Y. H. Loh, B. Wang, H. H. Ng, and P. Robson. Transcriptional regulation of nanog by OCT4 and SOX2. *J. Biol. Chem.*, 280:24731–24737, Jul 2005.

[165] G. Pan, J. Li, Y. Zhou, H. Zheng, and D. Pei. A negative feedback loop of transcription factors that controls stem cell pluripotency and self-renewal. *FASEB J.*, 20:1730–1732, Aug 2006.

[166] I. Glauche, M. Herberg, and I. Roeder. Nanog variability and pluripotency regulation of embryonic stem cells–insights from a mathematical model analysis. *PLoS ONE*, 5:e11238, 2010.

[167] I. Chambers. The molecular basis of pluripotency in mouse embryonic stem cells. *Cloning Stem Cells*, 6:386–391, 2004.

[168] A. Som, C. Harder, B. Greber, M. Siatkowski, Y. Paudel, G. Warsow, C. Cap, H. Scholer, and G. Fuellen. The PluriNetWork: an electronic representation of the network underlying pluripotency in mouse, and its applications. *PLoS ONE*, 5:e15165, 2010.

[169] F. J. Muller, L. C. Laurent, D. Kostka, I. Ulitsky, R. Williams, C. Lu, I. H. Park, M. S. Rao, R. Shamir, P. H. Schwartz, N. O. Schmidt, and J. F. Loring. Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, 455:401–405, Sep 2008.

[170] B. D. Macarthur, A. Ma'ayan, and I. R. Lemischka. Systems biology of stem cell fate and cellular reprogramming. *Nat. Rev. Mol. Cell Biol.*, 10:672–681, Oct 2009.

[171] J. B. Kim, H. Zaehres, G. Wu, L. Gentile, K. Ko, V. Sebastiano, M. J. Arauzo-Bravo, D. Ruau, D. W. Han, M. Zenke, and H. R. Scholer. Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors. *Nature*, 454:646–650, Jul 2008.

[172] N. Liu, M. Lu, X. Tian, and Z. Han. Molecular mechanisms involved in self-renewal and pluripotency of embryonic stem cells. *J. Cell. Physiol.*, 211:279–286, May 2007.

[173] O. Dreesen and A. H. Brivanlou. Signaling pathways in cancer and embryonic stem cells. *Stem Cell Rev*, 3:7–17, Jan 2007.

[174] K. Okita and S. Yamanaka. Intracellular signaling pathways regulating pluripotency of embryonic stem cells. *Curr Stem Cell Res Ther*, 1:103–111, Jan 2006.

[175] M. Boiani and H. R. Scholer. Regulatory networks in embryo-derived pluripotent stem cells. *Nat. Rev. Mol. Cell Biol.*, 6:872–884, Nov 2005.

[176] J. Nichols, I. Chambers, T. Taga, and A. Smith. Physiological rationale for responsiveness of mouse embryonic stem cells to gp130 cytokines. *Development*, 128:2333–2339, Jun 2001.

[177] S. A. Noggle, D. James, and A. H. Brivanlou. A molecular basis for human embryonic stem cell pluripotency. *Stem Cell Rev*, 1:111–118, 2005.

[178] Z. Wu, W. Zhang, G. Chen, L. Cheng, J. Liao, N. Jia, Y. Gao, H. Dai, J. Yuan, L. Cheng, and L. Xiao. Combinatorial signals of activin/nodal and bone morphogenic protein regulate the early lineage segregation of human embryonic stem cells. *J. Biol. Chem.*, 283:24991–25002, Sep 2008.

[179] T. Vogel, S. Ahrens, N. Buttner, and K. Krieglstein. Transforming growth factor beta promotes neuronal cell fate of mouse cortical and hippocampal progenitors in vitro and in vivo: identification of Nedd9 as an essential signaling component. *Cereb. Cortex*, 20:661–671, Mar 2010.

[180] R. Chen, G. Halder, Z. Zhang, and G. Mardon. Signaling by the TGF-beta homolog decapentaplegic functions reiteratively within the network of genes controlling retinal cell fate determination in Drosophila. *Development*, 126:935–943, Feb 1999.

[181] A. Bilitou and S. Ohnuma. The role of cell cycle in retinal development: cyclin-dependent kinase inhibitors co-ordinate cell-cycle inhibition, cell-fate determination and differentiation in the developing retina. *Dev. Dyn.*, 239:727–736, Mar 2010.

[182] N. Sato, I. M. Sanjuan, M. Heke, M. Uchida, F. Naef, and A. H. Brivanlou. Molecular signature of human embryonic stem cells and its comparison with the mouse. *Dev. Biol.*, 260:404–413, Aug 2003.

[183] D. James, A. J. Levine, D. Besser, and A. Hemmati-Brivanlou. TGF-beta/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development*, 132:1273–1282, Mar 2005.

[184] D. Besser. Expression of nodal, lefty-a, and lefty-B in undifferentiated human embryonic stem cells requires activation of Smad2/3. *J. Biol. Chem.*, 279:45076–45084, Oct 2004.

[185] R. P. Sharma and V. L. Chopra. Effect of the Wingless (wg1) mutation on wing and haltere development in Drosophila melanogaster. *Dev. Biol.*, 48:461–465, Feb 1976.

[186] Alexandra Klaus and Walter Birchmeier. Wnt signalling and its impact on development and cancer. *Nature Reviews Cancer*, 8(5):387–398, May 2008.

[187] N. Sato, L. Meijer, L. Skaltsounis, P. Greengard, and A. H. Brivanlou. Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. *Nat. Med.*, 10:55–63, Jan 2004.

[188] T. Burdon, C. Stracey, I. Chambers, J. Nichols, and A. Smith. Suppression of SHP-2 and ERK signalling promotes self-renewal of mouse embryonic stem cells. *Dev. Biol.*, 210:30–43, Jun 1999.

[189] L. Armstrong, O. Hughes, S. Yung, L. Hyslop, R. Stewart, I. Wappler, H. Peters, T. Walter, P. Stojkovic, J. Evans, M. Stojkovic, and M. Lako. The role of PI3K/AKT, MAPK/ERK and NFkappabeta signalling in the maintenance of human embryonic stem cell pluripotency and viability highlighted by transcriptional profiling and functional analysis. *Hum. Mol. Genet.*, 15:1894–1913, Jun 2006.

[190] D. Morgensztern and H. L. McLeod. PI3K/Akt/mTOR pathway as a target for cancer therapy. *Anticancer Drugs*, 16:797–803, Sep 2005.

[191] T. A. Yap, M. D. Garrett, M. I. Walton, F. Raynaud, J. S. de Bono, and P. Workman. Targeting the PI3K-AKT-mTOR pathway: progress, pitfalls, and promises. *Curr Opin Pharmacol*, 8:393–412, Aug 2008.

[192] M. Alvarez, E. Roman, E. S. Santos, and L. E. Raez. New targets for non-small-cell lung cancer therapy. *Expert Rev Anticancer Ther*, 7:1423–1437, Oct 2007.

[193] K. Takahashi, K. Mitsui, and S. Yamanaka. Role of ERas in promoting tumour-like properties in mouse embryonic stem cells. *Nature*, 423:541–545, May 2003.

[194] K. Takahashi, M. Murakami, and S. Yamanaka. Role of the phosphoinositide 3-kinase pathway in mouse embryonic stem (ES) cells. *Biochem. Soc. Trans.*, 33:1522–1525, Dec 2005.

[195] M. Stadtfeld, M. Nagaya, J. Utikal, G. Weir, and K. Hochedlinger. Induced pluripotent stem cells generated without viral integration. *Science*, 322:945–949, Nov 2008.

[196] S. Z. Pavletic, I. F. Khouri, M. Haagenson, R. J. King, P. J. Bierman, M. R. Bishop, M. Carston, S. Giralt, A. Molina, E. A. Copelan, O. Ringden, V. Roy, K. Ballen, D. R. Adkins, P. McCarthy, D. Weisdorf, E. Montserrat, and C. Anasetti. Unrelated donor marrow transplantation for B-cell chronic lymphocytic leukemia after using myeloablative conditioning: results from the Center for International Blood and Marrow Transplant research. *J. Clin. Oncol.*, 23:5788–5794, Aug 2005.

[197] E. Gyan, C. Foussard, P. Bertrand, P. Michenet, S. Le Gouill, C. Berthou, H. Maisonneuve, V. Delwail, R. Gressin, P. Quittet, J. P. Vilque, B. Desablens, J. Jaubert, J. F. Ramee, N. Arakelyan, A. Thyss, C. Molucon-Chabrot, R. Delepine, N. Milpied, P. Colombat, and E. Deconinck. High-dose therapy followed by autologous purged stem cell transplantation and doxorubicin-based chemotherapy in patients with advanced follicular lymphoma: a randomized multicenter study by the GOELAMS with final results after a median follow-up of 9 years. *Blood*, 113:995–1001, Jan 2009.

[198] Edward A. Copelan. Hematopoietic stem-cell transplantation. *New England Journal of Medicine*, 354(17):1813–1826, 2006.

[199] T Gorba, S Harper, and P J Mee. Prospects for neural stem cell therapy of alzheimer disease. *Stem Cells Regenerative Medicine*, pages 337–348, 2011.

[200] S. U. Kim and J. de Vellis. Stem cell-based cell therapy in neurological diseases: a review. *J. Neurosci. Res.*, 87:2183–2200, Aug 2009.

[201] O. Lindvall and Z. Kokaia. Stem cells in human neurodegenerative disorders–time for clinical translation? *J. Clin. Invest.*, 120:29–40, Jan 2010.

[202] J. L. DeSousa, S. Daya, and R. Malhotra. Adnexal surgery in patients undergoing ocular surface stem cell transplantation. *Ophthalmology*, 116:235–242, Feb 2009.

[203] A. M. Newman and J. B. Cooper. AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. *BMC Bioinformatics*, 11:117, 2010.

[204] P. Gellert, K. Jenniches, T. Braun, and S. Uchida. C-It: a knowledge database for tissue-enriched genes. *Bioinformatics*, 26:2328–2333, Sep 2010.

[205] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28:27–30, Jan 2000.

[206] Z. Lichner, E. Pall, A. Kerekes, E. Pallinger, P. Maraghechi, Z. Bosze, and E. Gocza. The miR-290-295 cluster promotes pluripotency maintenance by regulating cell cycle phase distribution in mouse embryonic stem cells. *Differentiation*, 81:11–24, Jan 2011.

[207] J. Wang, P. Alexander, L. Wu, R. Hammer, O. Cleaver, and S. L. McKnight. Dependence of mouse embryonic stem cells on threonine catabolism. *Science*, 325:435–439, Jul 2009.

[208] Q. L. Ying, J. Nichols, I. Chambers, and A. Smith. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell*, 115:281–292, Oct 2003.

[209] J. Zhang and L. Li. BMP signaling and stem cell regulation. *Dev. Biol.*, 284:1–11, Aug 2005.

[210] H. Huang, T. J. Song, X. Li, L. Hu, Q. He, M. Liu, M. D. Lane, and Q. Q. Tang. BMP signaling pathway is required for commitment of C3H10T1/2 pluripotent stem cells to the adipocyte lineage. *Proc. Natl. Acad. Sci. U.S.A.*, 106:12670–12675, Aug 2009.

[211] T. Sumi, N. Tsuneyoshi, N. Nakatsuji, and H. Suemori. Defining early lineage specification of human embryonic stem cells by the orchestrated balance of canonical Wnt/beta-catenin, Activin/Nodal and BMP signaling. *Development*, 135:2969–2979, Sep 2008.

[212] G. Roma, G. Cobellis, P. Claudiani, F. Maione, P. Cruz, G. Tripoli, M. Sardiello, I. Peluso, and E. Stupka. A novel view of the transcriptome revealed from gene trapping in mouse embryonic stem cells. *Genome Res.*, 17:1051–1060, Jul 2007.

[213] D. Zhang, W. Jiang, M. Liu, X. Sui, X. Yin, S. Chen, Y. Shi, and H. Deng. Highly efficient differentiation of human ES cells and iPS cells into mature pancreatic insulin-producing cells. *Cell Res.*, 19:429–438, Apr 2009.

[214] T. Miura, M. P. Mattson, and M. S. Rao. Cellular lifespan and senescence signaling in embryonic stem cells. *Aging Cell*, 3:333–343, Dec 2004.

[215] J. Peltier, A. O'Neill, and D. V. Schaffer. PI3K/Akt and CREB regulate adult neural hippocampal progenitor proliferation and differentiation. *Dev Neurobiol*, 67:1348–1361, Sep 2007.

[216] Y. J. Zhou, K. S. Magnuson, T. P. Cheng, M. Gadina, D. M. Frucht, J. Galon, F. Candotti, R. L. Geahlen, P. S. Changelian, and J. J. O'Shea. Hierarchy of protein tyrosine kinases in interleukin-2 (IL-2) signaling: activation of syk depends on Jak3; however, neither Syk nor Lck is required for IL-2-mediated STAT activation. *Mol. Cell. Biol.*, 20:4371–4380, Jun 2000.

[217] X. Hu, J. Chen, L. Wang, and L. B. Ivashkiv. Crosstalk among Jak-STAT, Toll-like receptor, and ITAM-dependent pathways in macrophage activation. *J. Leukoc. Biol.*, 82:237–243, Aug 2007.

[218] M. Lu, C. H. Glover, A. H. Tien, R. K. Humphries, J. M. Piret, and C. D. Helgason. Involvement of tyrosine kinase signaling in maintaining murine embryonic stem cell functionality. *Exp. Hematol.*, 35:1293–1302, Aug 2007.

[219] L. Aghajanova, S. Shen, A. M. Rojas, S. J. Fisher, J. C. Irwin, and L. C. Giudice. Comparative transcriptome analysis of human trophectoderm and embryonic stem cell-derived trophoblasts reveal key participants in early implantation. *Biol. Reprod.*, 86:1–21, Jan 2012.

[220] W. K. Tang, C. H. Chui, S. Fatima, S. H. Kok, K. C. Pak, T. M. Ou, K. S. Hui, M. M. Wong, J. Wong, S. Law, S. W. Tsao, K. Y. Lam, P. S. Beh, G. Srivastava, A. S. Chan, K. P. Ho, and J. C. Tang. Oncogenic properties of a novel gene JK-1 located in chromosome 5p and its overexpression in human esophageal squamous cell carcinoma. *Int. J. Mol. Med.*, 19:915–923, Jun 2007.

[221] S. M. Murphy, G. L. Davidson, S. Brandner, H. Houlden, and M. M. Reilly. Mutation in FAM134B causing severe hereditary sensory neuropathy. *J. Neurol. Neurosurg. Psychiatr.*, Nov 2010.

[222] I. Kurth, T. Pamminger, J. C. Hennings, D. Soehendra, A. K. Huebner, A. Rotthier, J. Baets, J. Senderek, H. Topaloglu, S. A. Farrell, G. Nurnberg, P. Nurnberg, P. De Jonghe, A. Gal, C. Kaether, V. Timmerman, and C. A. Hubner. Mutations in FAM134B, encoding a newly identified Golgi protein, cause severe sensory and autonomic neuropathy. *Nat. Genet.*, 41:1179–1181, Nov 2009.

[223] B. Fredricsson. Alkaline phosphatase in the Golgi substance of intestinal epithelium. *Exp. Cell Res.*, 10:63–65, Feb 1956.

[224] J. A. Thomson, J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones. Embryonic stem cell lines derived from human blastocysts. *Science*, 282:1145–1147, Nov 1998.

[225] B. A. Eipper, S. L. Milgram, E. J. Husten, H. Y. Yun, and R. E. Mains. Peptidylglycine alpha-amidating monooxygenase: a multifunctional protein with catalytic, processing, and routing domains. *Protein Sci.*, 2:489–497, Apr 1993.

[226] Y. Zhu, M. Carroll, F. R. Papa, M. Hochstrasser, and A. D. D'Andrea. DUB-1, a deubiquitinating enzyme with growth-suppressing activity. *Proc. Natl. Acad. Sci. U.S.A.*, 93:3275–3279, Apr 1996.

[227] M. Leeb and A. Wutz. Ring1B is crucial for the regulation of developmental control genes and PRC1 proteins but not X inactivation in embryonic cells. *J. Cell Biol.*, 178:219–229, Jul 2007.

[228] Y. L. Lee, K. F. Lee, J. S. Xu, K. L. Kwok, J. M. Luk, W. M. Lee, and W. S. Yeung. Embryotrophic factor-3 from human oviductal cells affects the messenger RNA expression of mouse blastocyst. *Biol. Reprod.*, 68:375–382, Feb 2003.

[229] K. A. D'Amour and F. H. Gage. Genetic and functional differences between multipotent neural and pluripotent embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 100 Suppl 1:11866–11872, Sep 2003.

[230] S. Boue, I. Paramonov, M. J. Barrero, and J. C. Izpisua Belmonte. Analysis of human and mouse reprogramming of somatic cells to induced pluripotent stem cells. What is in the plate? *PLoS ONE*, 5, 2010.

[231] V. Botquin, H. Hess, G. Fuhrmann, C. Anastassiadis, M. K. Gross, G. Vriend, and H. R. Scholer. New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. *Genes Dev.*, 12:2073–2090, Jul 1998.

[232] K. M. Haston, J. Y. Tung, and R. A. Reijo Pera. Dazl functions in maintenance of pluripotency and genetic and epigenetic programs of differentiation in mouse primordial germ cells in vivo and in vitro. *PLoS ONE*, 4:e5654, 2009.

[233] P. Sampath, D. K. Pritchard, L. Pabon, H. Reinecke, S. M. Schwartz, D. R. Morris, and C. E. Murry. A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. *Cell Stem Cell*, 2:448–460, May 2008.

[234] N. Salomonis, C. R. Schlieve, L. Pereira, C. Wahlquist, A. Colas, A. C. Zambon, K. Vranizan, M. J. Spindler, A. R. Pico, M. S. Cline, T. A. Clark, A. Williams, J. E. Blume, E. Samal, M. Mercola, B. J. Merrill, and B. R. Conklin. Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation. *Proc. Natl. Acad. Sci. U.S.A.*, 107:10514–10519, Jun 2010.

[235] A. J. Levine and A. H. Brivanlou. GDF3, a BMP inhibitor, regulates cell fate in stem cells and early embryos. *Development*, 133:209–216, Jan 2006.

[236] A. T. Clark, R. T. Rodriguez, M. S. Bodnar, M. J. Abeyta, M. I. Cedars, P. J. Turek, M. T. Firpo, and R. A. Reijo Pera. Human STELLAR, NANOG, and GDF3 genes are expressed in pluripotent cells and map to chromosome 12p13, a hotspot for teratocarcinoma. *Stem Cells*, 22:169–179, 2004.

[237] K. V. Tarasov, Y. S. Tarasova, W. L. Tam, D. R. Riordon, S. T. Elliott, G. Kania, J. Li, S. Yamanaka, D. G. Crider, G. Testa, R. A. Li, B. Lim, C. L. Stewart, Y. Liu, J. E. Van Eyk, R. P. Wersto, A. M. Wobus, and K. R. Boheler. B-MYB is essential for normal cell cycle progression and chromosomal stability of embryonic stem cells. *PLoS ONE*, 3:e2478, 2008.

[238] X. Zhang, J. Zhang, T. Wang, M. A. Esteban, and D. Pei. Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells. *J. Biol. Chem.*, 283:35825–35833, Dec 2008.

[239] W. Shi, H. Wang, G. Pan, Y. Geng, Y. Guo, and D. Pei. Regulation of the pluripotency marker Rex-1 by Nanog and Sox2. *J. Biol. Chem.*, 281:23319–23325, Aug 2006.

[240] H. Wu, A. C. D'Alessio, S. Ito, K. Xia, Z. Wang, K. Cui, K. Zhao, Y. E. Sun, and Y. Zhang. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*, 473:389–393, May 2011.

[241] G. Ficz, M. R. Branco, S. Seisenberger, F. Santos, F. Krueger, T. A. Hore, C. J. Marques, S. Andrews, and W. Reik. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, 473:398–402, May 2011.

[242] S. Ito, A. C. D'Alessio, O. V. Taranova, K. Hong, L. C. Sowers, and Y. Zhang. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, 466:1129–1133, Aug 2010.

[243] C. Priller, T. Bauer, G. Mitteregger, B. Krebs, H. A. Kretzschmar, and J. Herms. Synapse formation and function is modulated by the amyloid precursor protein. *J. Neurosci.*, 26:7212–7221, Jul 2006.

[244] P. R. Turner, K. O'Connor, W. P. Tate, and W. C. Abraham. Roles of amyloid precursor protein and its fragments in regulating neural activity, plasticity and memory. *Prog. Neurobiol.*, 70:1–32, May 2003.

[245] T. Matsui, M. Ingelsson, H. Fukumoto, K. Ramasamy, H. Kowa, M. P. Frosch, M. C. Irizarry, and B. T. Hyman. Expression of APP pathway mRNAs and proteins in Alzheimer's disease. *Brain Res.*, 1161:116–123, Aug 2007.

[246] C. L. Masters, G. Simms, N. A. Weinman, G. Multhaup, B. L. McDonald, and K. Beyreuther. Amyloid plaque core protein in Alzheimer disease and Down syndrome. *Proc. Natl. Acad. Sci. U.S.A.*, 82:4245–4249, Jun 1985.

[247] E. Abramov, I. Dolev, H. Fogel, G. D. Ciccotosto, E. Ruff, and I. Slutsky. Amyloid-beta as a positive endogenous regulator of release probability at hippocampal synapses. *Nat. Neurosci.*, 12:1567–1576, Dec 2009.

[248] A. Watanabe, M. Hasegawa, M. Suzuki, K. Takio, M. Morishima-Kawashima, K. Titani, T. Arai, K. S. Kosik, and Y. Ihara. In vivo phosphorylation sites in fetal and adult rat tau. *J. Biol. Chem.*, 268:25712–25717, Dec 1993.

[249] J. Lewis, D. W. Dickson, W. L. Lin, L. Chisholm, A. Corral, G. Jones, S. H. Yen, N. Sahara, L. Skipper, D. Yager, C. Eckman, J. Hardy, M. Hutton, and E. McGowan. Enhanced neurofibrillary degeneration in transgenic mice expressing mutant tau and APP. *Science*, 293:1487–1491, Aug 2001.

[250] J. Gotz, F. Chen, J. van Dorpe, and R. M. Nitsch. Formation of neurofibrillary tangles in P301l tau transgenic mice induced by Abeta 42 fibrils. *Science*, 293:1491–1495, Aug 2001.

[251] E. D. Roberson, K. Scearce-Levie, J. J. Palop, F. Yan, I. H. Cheng, T. Wu, H. Gerstein, G. Q. Yu, and L. Mucke. Reducing endogenous tau ameliorates amyloid beta-induced deficits in an Alzheimer's disease mouse model. *Science*, 316:750–754, May 2007.

[252] R. Dixit, J. L. Ross, Y. E. Goldman, and E. L. Holzbaur. Differential regulation of dynein and kinesin motor proteins by tau. *Science*, 319:1086–1089, Feb 2008.

[253] D. N. Drechsel, A. A. Hyman, M. H. Cobb, and M. W. Kirschner. Modulation of the dynamic instability of tubulin assembly by the microtubule-associated protein tau. *Mol. Biol. Cell*, 3:1141–1154, Oct 1992.

[254] G. Lee, N. Cowan, and M. Kirschner. The primary structure and heterogeneity of tau protein from mouse brain. *Science*, 239:285–288, Jan 1988.

[255] L. M. Ittner, Y. D. Ke, F. Delerue, M. Bi, A. Gladbach, J. van Eersel, H. Wolfing, B. C. Chieng, M. J. Christie, I. A. Napier, A. Eckert, M. Staufenbiel, E. Hardeman, and J. Gotz. Dendritic function of tau mediates amyloid-beta toxicity in Alzheimer's disease mouse models. *Cell*, 142:387–397, Aug 2010.

[256] C. Ballatore, V. M. Lee, and J. Q. Trojanowski. Tau-mediated neurodegeneration in Alzheimer's disease and related disorders. *Nat. Rev. Neurosci.*, 8:663–672, Sep 2007.

[257] D. J. Selkoe. Alzheimer's disease is a synaptic failure. *Science*, 298:789–791, Oct 2002.

[258] G. M. Shankar, S. Li, T. H. Mehta, A. Garcia-Munoz, N. E. Shepardson, I. Smith, F. M. Brett, M. A. Farrell, M. J. Rowan, C. A. Lemere, C. M. Regan, D. M. Walsh, B. L. Sabatini, and D. J. Selkoe. Amyloid-beta protein dimers isolated directly from Alzheimer's brains impair synaptic plasticity and memory. *Nat. Med.*, 14:837–842, Aug 2008.

[259] L. Zhao, Q. L. Ma, F. Calon, M. E. Harris-White, F. Yang, G. P. Lim, T. Morihara, O. J. Ubeda, S. Ambegaokar, J. E. Hansen, R. H. Weisbart, B. Teter, S. A. Frautschy, and G. M. Cole. Role of p21-activated kinase pathway defects in the cognitive deficits of Alzheimer disease. *Nat. Neurosci.*, 9:234–242, Feb 2006.

[260] A. Goate, M. C. Chartier-Harlin, M. Mullan, J. Brown, F. Crawford, L. Fidani, L. Giuffra, A. Haynes, N. Irving, and L. James. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature*, 349:704–706, Feb 1991.

[261] R. Ghidoni, V. Albertini, R. Squitti, A. Paterlini, A. Bruno, S. Bernardini, E. Cassetta, P. M. Rossini, F. Squitieri, L. Benussi, and G. Binetti. Novel T719P AbetaPP mutation unbalances the relative proportion of amyloid-beta peptides. *J. Alzheimers Dis.*, 18:295–303, 2009.

[262] C. De Jonghe, C. Esselens, S. Kumar-Singh, K. Craessaerts, S. Serneels, F. Checler, W. Annaert, C. Van Broeckhoven, and B. De Strooper. Pathogenic APP mutations near the gamma-secretase cleavage site differentially affect Abeta secretion and APP C-terminal fragment stability. *Hum. Mol. Genet.*, 10:1665–1671, Aug 2001.

[263] M. Cruts, B. Dermaut, R. Rademakers, M. Van den Broeck, F. Stogbauer, and C. Van Broeckhoven. Novel APP mutation V715A associated with presenile Alzheimer's disease in a German family. *J. Neurol.*, 250:1374–1375, Nov 2003.

[264] R. Sherrington, E. I. Rogaev, Y. Liang, E. A. Rogaeva, G. Levesque, M. Ikeda, H. Chi, C. Lin, G. Li, K. Holman, T. Tsuda, L. Mar, J. F. Foncin, A. C. Bruni, M. P. Montesi, S. Sorbi, I. Rainero, L. Pinessi, L. Nee, I. Chumakov, D. Pollen, A. Brookes, P. Sanseau, R. J. Polinsky, W. Wasco, H. A. Da Silva, J. L. Haines, M. A. Perkicak-Vance, R. E. Tanzi, A. D. Roses, P. E. Fraser, J. M. Rommens, and P. H. St George-Hyslop. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature*, 375:754–760, Jun 1995.

[265] D. Campion, J. M. Flaman, A. Brice, D. Hannequin, B. Dubois, C. Martin, V. Moreau, F. Charbonnier, O. Didierjean, and S. Tardieu. Mutations of the presenilin I gene in families with early-onset Alzheimer's disease. *Hum. Mol. Genet.*, 4:2373–2377, Dec 1995.

[266] C. M. van Duijn, M. Cruts, J. Theuns, G. Van Gassen, H. Backhovens, M. van den Broeck, A. Wehnert, S. Serneels, A. Hofman, and C. Van Broeckhoven. Genetic association of the presenilin-1 regulatory region with early-onset Alzheimer's disease in a population-based sample. *Eur. J. Hum. Genet.*, 7:801–806, 1999.

[267] E. Levy-Lahad, W. Wasco, P. Poorkaj, D. M. Romano, J. Oshima, W. H. Pettingell, C. E. Yu, P. D. Jondro, S. D. Schmidt, and K. Wang. Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science*, 269:973–977, Aug 1995.

[268] E. I. Rogaev, R. Sherrington, E. A. Rogaeva, G. Levesque, M. Ikeda, Y. Liang, H. Chi, C. Lin, K. Holman, and T. Tsuda. Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature*, 376:775–778, Aug 1995.

[269] R. W. Mahley and S. C. Rall. Apolipoprotein E: far more than a lipid transport protein. *Annu Rev Genomics Hum Genet*, 1:507–537, 2000.

[270] K. H. Weisgraber and R. W. Mahley. Human apolipoprotein E: the Alzheimer's disease connection. *FASEB J.*, 10:1485–1494, Nov 1996.

[271] D. M. Holtzman, K. R. Bales, T. Tenkova, A. M. Fagan, M. Parsadanian, L. J. Sartorius, B. Mackey, J. Olney, D. McKeel, D. Wozniak, and S. M. Paul. Apolipoprotein E isoform-dependent amyloid deposition and neuritic degeneration in a mouse model of Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.*, 97:2892–2897, Mar 2000.

[272] J. C. Dodart, R. A. Marr, M. Koistinaho, B. M. Gregersen, S. Malkani, I. M. Verma, and S. M. Paul. Gene delivery of human apolipoprotein E alters brain Abeta burden in a mouse model of Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.*, 102:1211–1216, Jan 2005.

[273] W. J. Strittmatter, K. H. Weisgraber, D. Y. Huang, L. M. Dong, G. S. Salvesen, M. Pericak-Vance, D. Schmechel, A. M. Saunders, D. Goldgaber, and A. D. Roses. Binding of human apolipoprotein E to synthetic amyloid beta peptide: isoform-specific effects and implications for late-onset Alzheimer disease. *Proc. Natl. Acad. Sci. U.S.A.*, 90:8098–8102, Sep 1993.

[274] C. Klein, E. M. Kramer, A. M. Cardine, B. Schraven, R. Brandt, and J. Trotter. Process outgrowth of oligodendrocytes is promoted by interaction of fyn kinase with the cytoskeletal protein tau. *J. Neurosci.*, 22:698–707, Feb 2002.

[275] G. J. Ho, M. Hashimoto, A. Adame, M. Izu, M. F. Alford, L. J. Thal, L. A. Hansen, and E. Masliah. Altered p59Fyn kinase expression accompanies disease progression in Alzheimer's disease: implications for its functional role. *Neurobiol. Aging*, 26:625–635, May 2005.

[276] J. Chin, J. J. Palop, J. Puolivali, C. Massaro, N. Bien-Ly, H. Gerstein, K. Scearce-Levie, E. Masliah, and L. Mucke. Fyn kinase induces synaptic and cognitive impairments in a transgenic mouse model of Alzheimer's disease. *J. Neurosci.*, 25:9694–9703, Oct 2005.

[277] J. C. Cruz, D. Kim, L. Y. Moy, M. M. Dobbin, X. Sun, R. T. Bronson, and L. H. Tsai. p25/cyclin-dependent kinase 5 induces production and intraneuronal accumulation of amyloid beta in vivo. *J. Neurosci.*, 26:10536–10541, Oct 2006.

[278] J. C. Cruz and L. H. Tsai. A Jekyll and Hyde kinase: roles for Cdk5 in brain development and disease. *Curr. Opin. Neurobiol.*, 14:390–394, Jun 2004.

[279] E. Rockenstein, M. Torrance, A. Adame, M. Mante, P. Bar-on, J. B. Rose, L. Crews, and E. Masliah. Neuroprotective effects of regulators of the glycogen synthase kinase-3beta signaling pathway in a transgenic model of Alzheimer's disease are associated with reduced amyloid precursor protein phosphorylation. *J. Neurosci.*, 27:1981–1991, Feb 2007.

[280] E. Rockenstein, M. Torrance, M. Mante, A. Adame, A. Paulino, J. B. Rose, L. Crews, H. Moessler, and E. Masliah. Cerebrolysin decreases amyloid-beta production by regulating amyloid protein precursor maturation in a transgenic model of Alzheimer's disease. *J. Neurosci. Res.*, 83:1252–1261, May 2006.

[281] V. W. Tsai, H. L. Scott, R. J. Lewis, and P. R. Dodd. The role of group I metabotropic glutamate receptors in neuronal excitotoxicity in Alzheimer's disease. *Neurotox Res*, 7:125–141, 2005.

[282] G. M. Shankar, B. L. Bloodgood, M. Townsend, D. M. Walsh, D. J. Selkoe, and B. L. Sabatini. Natural oligomers of the Alzheimer amyloid-beta protein induce reversible synapse loss by modulating an NMDA-type glutamate receptor-dependent signaling pathway. *J. Neurosci.*, 27:2866–2875, Mar 2007.

[283] S. Li, S. Hong, N. E. Shepardson, D. M. Walsh, G. M. Shankar, and D. Selkoe. Soluble oligomers of amyloid Beta protein facilitate hippocampal long-term depression by disrupting neuronal glutamate uptake. *Neuron*, 62:788–801, Jun 2009.

[284] T. Nakamura and S. A. Lipton. Redox regulation of mitochondrial fission, protein misfolding, synaptic damage, and neuronal cell death: potential implications for Alzheimer's and Parkinson's diseases. *Apoptosis*, 15:1354–1363, Nov 2010.

[285] M. T. Lin and M. F. Beal. Alzheimer's APP mangles mitochondria. *Nat. Med.*, 12:1241–1243, Nov 2006.

[286] R. A. Nixon and A. M. Cataldo. Lysosomal system pathways: genes to neurodegeneration in Alzheimer's disease. *J. Alzheimers Dis.*, 9:277–289, 2006.

[287] S. M. Greenberg, W. Q. Qiu, D. J. Selkoe, A. Ben-Itzhak, and K. S. Kosik. Amino-terminal region of the beta-amyloid precursor protein activates mitogen-activated protein kinase. *Neurosci. Lett.*, 198:52–56, Sep 1995.

[288] C. K. Combs, D. E. Johnson, S. B. Cannady, T. M. Lehman, and G. E. Landreth. Identification of microglial signal transduction pathways mediating a neurotoxic response to amyloidogenic fragments of beta-amyloid and prion proteins. *J. Neurosci.*, 19:928–939, Feb 1999.

[289] B. Webster, L. Hansen, A. Adame, L. Crews, M. Torrance, L. Thal, and E. Masliah. Astroglial activation of extracellular-regulated kinase in early stages of Alzheimer disease. *J. Neuropathol. Exp. Neurol.*, 65:142–151, Feb 2006.

[290] X. Zhu, A. K. Raina, C. A. Rottkamp, G. Aliev, G. Perry, H. Boux, and M. A. Smith. Activation and redistribution of c-jun N-terminal kinase/stress activated protein kinase in degenerating neurons in Alzheimer's disease. *J. Neurochem.*, 76:435–441, Jan 2001.

[291] Q. L. Ma, M. E. Harris-White, O. J. Ubeda, M. Simmons, W. Beech, G. P. Lim, B. Teter, S. A. Frautschy, and G. M. Cole. Evidence of Abeta- and transgene-dependent defects in ERK-CREB signaling in Alzheimer's models. *J. Neurochem.*, 103:1594–1607, Nov 2007.

[292] Q. L. Ma, F. Yang, F. Calon, O. J. Ubeda, J. E. Hansen, R. H. Weisbart, W. Beech, S. A. Frautschy, and G. M. Cole. p21-activated kinase-aberrant activation and translocation in Alzheimer disease pathogenesis. *J. Biol. Chem.*, 283:14132–14143, May 2008.

[293] C. R. Jack, M. S. Albert, D. S. Knopman, G. M. McKhann, R. A. Sperling, M. C. Carrillo, B. Thies, and C. H. Phelps. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7:257–262, May 2011.

[294] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack, J. Kaye, T. J. Montine, D. C. Park, E. M. Reiman, C. C. Rowe, E. Siemers, Y. Stern, K. Yaffe, M. C. Carrillo, B. Thies, M. Morrison-Bogorad, M. V. Wagster, and C. H. Phelps. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7:280–292, May 2011.

[295] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, P. J. Snyder,

M. C. Carrillo, B. Thies, and C. H. Phelps. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7:270–279, May 2011.

[296] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub, and C. H. Phelps. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7:263–269, May 2011.

[297] H. Barthel, H. J. Gertz, S. Dresel, O. Peters, P. Bartenstein, K. Buerger, F. Hiemeyer, S. M. Wittemer-Rump, J. Seibyl, C. Reininger, O. Sabri, R. Ehret, A. Drzezga, B. Krause, A. Kurz, P. Bartenstein, D. Rujescu, K. Buerger, J. Kotzerke, M. Bauer, T. Kuwert, J. Kornhuber, A. Bockisch, H. C. Diener, J. Wiltfang, S. Dresel, I. Heuser, O. Peters, O. Schober, T. Fey, A. Bauer, W. Maier, A. Buck, C. Hock, C. van Dyck, R. E. Carson, D. Jennings, J. Seibyl, S. De Santi, K. Friedman, J. R. Steinerman, D. Blaufox, O. Sabri, C. Rowe, M. Woodward, B. Chatterton, R. J. Prowse, G. Larcos, R. Purcell, O. Sabri, H. Barthel, and H. J. Gertz. Cerebral amyloid-beta PET with florbetaben (18F) in patients with Alzheimer's disease and healthy controls: a multicentre phase 2 diagnostic study. *Lancet Neurol*, 10:424–435, May 2011.

[298] M. F. Folstein, S. E. Folstein, and P. R. McHugh. 'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*, 12:189–198, Nov 1975.

[299] L. Mucke. Neuroscience: Alzheimer's disease. *Nature*, 461:895–897, Oct 2009.

[300] N. Scarmeas, Y. Stern, R. Mayeux, and J. A. Luchsinger. Mediterranean diet, Alzheimer disease, and vascular mediation. *Arch. Neurol.*, 63:1709–1717, Dec 2006.

[301] S. Kalmijn, L. J. Launer, A. Ott, J. C. Witteman, A. Hofman, and M. M. Breteler. Dietary fat intake and the risk of incident dementia in the Rotterdam Study. *Ann. Neurol.*, 42:776–782, Nov 1997.

[302] R. Stewart, Q. L. Xue, K. Masaki, H. Petrovitch, G. W. Ross, L. R. White, and L. J. Launer. Change in blood pressure and incident dementia: a 32-year prospective study. *Hypertension*, 54:233–240, Aug 2009.

[303] M. Michikawa. The role of cholesterol in pathogenesis of Alzheimer's disease: dual metabolic interaction between amyloid beta-protein and cholesterol. *Mol. Neurobiol.*, 27:1–12, Feb 2003.

[304] M. Kivipelto, E. L. Helkala, M. P. Laakso, T. Hanninen, M. Hallikainen, K. Alhainen, S. Iivonen, A. Mannermaa, J. Tuomilehto, A. Nissinen, and H. Soininen. Apolipoprotein E epsilon4 allele, elevated midlife total cholesterol level, and high midlife systolic blood pressure are independent risk factors for late-life Alzheimer disease. *Ann. Intern. Med.*, 137:149–155, Aug 2002.

[305] A. Ciobica, M. Padurariu, W. Bild, and C. Stefanescu. Cardiovascular risk factors as potential markers for mild cognitive impairment and Alzheimer's disease. *Psychiatr Danub*, 23:340–346, Dec 2011.

[306] D. A. Valenti. Alzheimer's disease: visual system review. *Optometry*, 81:12–21, Jan 2010.

[307] S. Frost, R. N. Martins, and Y. Kanagasingam. Ocular biomarkers for early detection of Alzheimer's disease. *J. Alzheimers Dis.*, 22:1–16, 2010.

[308] S. Frost, Martins R.N., and Kanagasingam. Retinal vascular parameters as biomarkers for alzheimer's disease. *Alzheimers Dement.*, 7:S136, 2011.

[309] L. Bertram, M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.*, 39:17–23, Jan 2007.

[310] Dennis P. Wall, Rimma Pivovarov, Mark Tong, Jae-Yoon Y. Jung, Vincent A. Fusaro, Todd F. DeLuca, and Peter J. Tonellato. Genotator: a disease-agnostic tool for genetic annotation of disease. *BMC medical genomics*, 3(1):50+, 2010.

[311] M. Soler-Lopez, A. Zanzoni, R. Lluis, U. Stelzl, and P. Aloy. Interactome mapping suggests new mechanistic details underlying Alzheimer's disease. *Genome Res.*, 21:364–376, Mar 2011.

[312] J. Goni, F. J. Esteban, N. V. de Mendizabal, J. Sepulcre, S. Ardanza-Trevijano, I. Agirrezabal, and P. Villoslada. A computational analysis of protein-protein interaction networks in neurodegenerative diseases. *BMC Syst Biol*, 2:52, 2008.

[313] G. H. Fisher, A. D'Aniello, A. Vetere, L. Padula, G. P. Cusano, and E. H. Man. Free D-aspartate and D-alanine in normal and Alzheimer brain. *Brain Res. Bull.*, 26:983–985, Jun 1991.

[314] S. C. Papasozomenos. The heat shock-induced hyperphosphorylation of tau is estrogen-independent and prevented by androgens: implications for Alzheimer disease. *Proc. Natl. Acad. Sci. U.S.A.*, 94:6612–6617, Jun 1997.

[315] I. I. Kruman, R. P. Wersto, F. Cardozo-Pelaez, L. Smilenov, S. L. Chan, F. J. Chrest, R. Emokpae, M. Gorospe, and M. P. Mattson. Cell cycle activation linked to neuronal cell death initiated by DNA damage. *Neuron*, 41:549–561, Feb 2004.

[316] M. Sastre, T. Klockgether, and M. T. Heneka. Contribution of inflammatory processes to Alzheimer's disease: molecular mechanisms. *Int. J. Dev. Neurosci.*, 24:167–176, 2006.

[317] S. Oddo, A. Caccamo, J. D. Shepherd, M. P. Murphy, T. E. Golde, R. Kayed, R. Metherate, M. P. Mattson, Y. Akbari, and F. M. LaFerla. Triple-transgenic model of Alzheimer's disease with plaques and tangles: intracellular Abeta and synaptic dysfunction. *Neuron*, 39:409–421, Jul 2003.

[318] P. Das and T. Golde. Dysfunction of TGF-beta signaling in Alzheimer's disease. *J. Clin. Invest.*, 116:2855–2857, Nov 2006.

[319] J. Adachi, Y. Mori, S. Matsui, and T. Matsuda. Comparison of gene expression patterns between 2,3,7,8-tetrachlorodibenzo-p-dioxin and a natural arylhydrocarbon receptor ligand, indirubin. *Toxicol. Sci.*, 80:161–169, Jul 2004.

[320] J. McLaurin and P. E. Fraser. Effect of amino-acid substitutions on Alzheimer's amyloid-beta peptide-glycosaminoglycan interactions. *Eur. J. Biochem.*, 267:6353–6361, Nov 2000.

[321] G. S. h. Burbaeva, I. S. Boksha, E. B. Tereshkina, O. K. Savushkina, L. I. Starodubtseva, and M. S. Turishcheva. Glutamate metabolizing enzymes in prefrontal cortex of Alzheimer's disease patients. *Neurochem. Res.*, 30:1443–1451, Nov 2005.

[322] M. A. Kennedy, N. Kabbani, J. P. Lambert, L. A. Swayne, F. Ahmed, D. Figeys, S. A. Bennett, J. Bryan, and K. Baetz. Srf1 is a novel regulator of phospholipase D activity and is essential to buffer the toxic effects of C16:0 platelet activating factor. *PLoS Genet.*, 7:e1001299, 2011.

[323] A. A. Farooqui, L. A. Horrocks, and T. Farooqui. Interactions between neural membrane glycerophospholipid and sphingolipid mediators: a recipe for neural cell survival or suicide. *J. Neurosci. Res.*, 85:1834–1850, Jul 2007.

[324] J. Pi, Y. Bai, J. M. Reece, J. Williams, D. Liu, M. L. Freeman, W. E. Fahl, D. Shugar, J. Liu, W. Qu, S. Collins, and M. P. Waalkes. Molecular mechanism of human Nrf2 activation and degradation: role of sequential

phosphorylation by protein kinase CK2. *Free Radic. Biol. Med.*, 42:1797–1806, Jun 2007.

[325] D. Boyd-Kimball, H. Mohmmad Abdul, T. Reed, R. Sultana, and D. A. Butterfield. Role of phenylalanine 20 in Alzheimer's amyloid beta-peptide (1-42)-induced oxidative stress and neurotoxicity. *Chem. Res. Toxicol.*, 17:1743–1749, Dec 2004.

[326] T. C. Birdsall. Therapeutic applications of taurine. *Altern Med Rev*, 3:128–136, Apr 1998.

[327] P. R. Louzada, A. C. Paula Lima, D. L. Mendonca-Silva, F. Noel, F. G. De Mello, and S. T. Ferreira. Taurine prevents the neurotoxicity of beta-amyloid and glutamate receptor agonists: activation of GABA receptors and possible implications for Alzheimer's disease and other neurological disorders. *FASEB J.*, 18:511–518, Mar 2004.

[328] M. Chiappelli, B. Borroni, S. Archetti, E. Calabrese, M. M. Corsi, M. Franceschi, A. Padovani, and F. Licastro. VEGF gene and phenotype relation with Alzheimer's disease and mild cognitive impairment. *Rejuvenation Res*, 9:485–493, 2006.

[329] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 35:D61–65, Jan 2007.

[330] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res.*, 12:996–1006, Jun 2002.

[331] R. J. Blaschke, A. P. Monaghan, D. Bock, and G. A. Rappold. A novel murine PKA-related protein kinase involved in neuronal differentiation. *Genomics*, 64:187–194, Mar 2000.

[332] A. Bettencourt da Cruz, J. Wentzell, and D. Kretzschmar. Swiss Cheese, a protein involved in progressive neurodegeneration, acts as a noncanonical regulatory subunit for PKA-C3. *J. Neurosci.*, 28:10885–10892, Oct 2008.

[333] W. Li, Z. X. Yu, and R. M. Kotin. Profiles of PrKX expression in developmental mouse embryo and human tissues. *J. Histochem. Cytochem.*, 53:1003–1009, Aug 2005.

[334] Z. Zhang, P. Harrison, and M. Gerstein. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.*, 12:1466–1482, Oct 2002.

[335] F. M. Gebhardt, H. A. Scott, and P. R. Dodd. Housekeepers for accurate transcript expression analysis in Alzheimer's disease autopsy brain tissue. *Alzheimers Dement*, 6:465–474, Nov 2010.

[336] M. Squillario and A. Barla. A computational procedure for functional characterization of potential marker genes from molecular data: Alzheimer's as a case study. *BMC Med Genomics*, 4:55, 2011.

[337] E. F. Vanin. Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.*, 19:253–272, 1985.

[338] R. C. Pink, K. Wicks, D. P. Caley, E. K. Punch, L. Jacobs, and D. R. Carter. Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA*, 17:792–798, May 2011.

[339] Y. Yano, R. Saito, N. Yoshida, A. Yoshiki, A. Wynshaw-Boris, M. Tomita, and S. Hirotsune. A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *J. Mol. Med.*, 82:414–422, Jul 2004.

[340] P. M. Harrison, D. Zheng, Z. Zhang, N. Carriero, and M. Gerstein. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res.*, 33:2374–2383, 2005.

[341] D. Zheng, Z. Zhang, P. M. Harrison, J. Karro, N. Carriero, and M. Gerstein. Integrated pseudogene annotation for human chromosome 22: evidence for transcription. *J. Mol. Biol.*, 349:27–45, May 2005.

[342] D. Zheng, A. Frankish, R. Baertsch, P. Kapranov, A. Reymond, S. W. Choo, Y. Lu, F. Denoeud, S. E. Antonarakis, M. Snyder, Y. Ruan, C. L. Wei, T. R. Gingeras, R. Guigo, J. Harrow, and M. B. Gerstein. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res.*, 17:839–851, Jun 2007.

[343] E. Chiefari, S. Iiritano, F. Paonessa, I. Le Pera, B. Arcidiacono, M. Filocamo, D. Foti, S. A. Liebhaber, and A. Brunetti. Pseudogene-mediated posttranscriptional silencing of HMGA1 can result in insulin resistance and type 2 diabetes. *Nat Commun*, 1:40, 2010.

[344] G. Suo, J. Han, X. Wang, J. Zhang, Y. Zhao, Y. Zhao, and J. Dai. Oct4 pseudogenes are transcribed in cancers. *Biochem. Biophys. Res. Commun.*, 337:1047–1051, Dec 2005.

[345] M. Zou, E. Y. Baitei, A. S. Alzahrani, F. Al-Mohanna, N. R. Farid, B. Meyer, and Y. Shi. Oncogenic activation of MAP kinase by BRAF pseudogene in thyroid tumors. *Neoplasia*, 11:57–65, Jan 2009.

[346] L. Poliseno, L. Salmena, J. Zhang, B. Carver, W. J. Haveman, and P. P. Pandolfi. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465:1033–1038, Jun 2010.

[347] J. S. Mattick. Long noncoding RNAs in cell and developmental biology. *Semin. Cell Dev. Biol.*, 22:327, Jun 2011.

[348] D. P. Caley, R. C. Pink, D. Trujillano, and D. R. Carter. Long noncoding RNAs, chromatin, and development. *ScientificWorldJournal*, 10:90–102, 2010.

[349] M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458:223–227, Mar 2009.

[350] X. Gong, R. Wu, Y. Zhang, W. Zhao, L. Cheng, Y. Gu, L. Zhang, J. Wang, J. Zhu, and Z. Guo. Extracting consistent knowledge from highly inconsistent cancer gene data sources. *BMC Bioinformatics*, 11:76, 2010.

[351] M. Zhang, C. Yao, Z. Guo, J. Zou, L. Zhang, H. Xiao, D. Wang, D. Yang, X. Gong, J. Zhu, Y. Li, and X. Li. Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics*, 24:2057–2063, Sep 2008.

[352] F. C. Mansergh, C. S. Daly, A. L. Hurley, M. A. Wride, S. M. Hunter, and M. J. Evans. Gene expression profiles during early differentiation of mouse embryonic stem cells. *BMC Dev. Biol.*, 9:5, 2009.

[353] S. J. Bruce, B. B. Gardiner, L. J. Burke, M. M. Gongora, S. M. Grimmond, and A. C. Perkins. Dynamic transcription programs during ES cell differentiation towards mesoderm in serum versus serum-freeBMP4 culture. *BMC Genomics*, 8:365, 2007.

[354] N. R. Treff, R. K. Vincent, M. L. Budde, V. L. Browning, J. F. Magliocca, V. Kapur, and J. S. Odorico. Differentiation of embryonic stem cells conditionally expressing neurogenin 3. *Stem Cells*, 24:2529–2537, Nov 2006.

[355] Y. H. Loh, Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K. Y. Wong, K. W. Sung, C. W. Lee, X. D. Zhao, K. P. Chiu, L. Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C. L. Wei, Y. Ruan, B. Lim, and H. H. Ng. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, 38:431–440, Apr 2006.

[356] K. Kurimoto, Y. Yabuta, Y. Ohinata, Y. Ono, K. D. Uno, R. G. Yamada, H. R. Ueda, and M. Saitou. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res.*, 34:e42, 2006.

[357] F. Ulloa-Montoya, B. L. Kidder, K. A. Pauwelyn, L. G. Chase, A. Luttun, A. Crabbe, M. Geraerts, A. A. Sharov, Y. Piao, M. S. Ko, W. S. Hu, and C. M. Verfaillie. Comparative transcriptome analysis of embryonic and adult stem cells with extended and limited differentiation capacity. *Genome Biol.*, 8:R163, 2007.

[358] L. O. Barrera, Z. Li, A. D. Smith, K. C. Arden, W. K. Cavenee, M. Q. Zhang, R. D. Green, and B. Ren. Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. *Genome Res.*, 18:46–59, Jan 2008.

[359] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448:553–560, Aug 2007.

[360] M. Ema, D. Mori, H. Niwa, Y. Hasegawa, Y. Yamanaka, S. Hitoshi, J. Mimura, Y. Kawabe, T. Hosoya, M. Morita, D. Shimosato, K. Uchida, N. Suzuki, J. Yanagisawa, K. Sogawa, J. Rossant, M. Yamamoto, S. Takahashi, and Y. Fujii-Kuriyama. Krüppel-like factor 5 is essential for blastocyst development and the normal self-renewal of mouse ESCs. *Cell Stem Cell*, 3:555–567, Nov 2008.

[361] P. Sampath, D. K. Pritchard, L. Pabon, H. Reinecke, S. M. Schwartz, D. R. Morris, and C. E. Murry. A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. *Cell Stem Cell*, 2:448–460, May 2008.

[362] L. Thorrez, K. Van Deun, L. C. Tranchevent, L. Van Lommel, K. Engelen, K. Marchal, Y. Moreau, I. Van Mechelen, and F. Schuit. Using ribosomal protein genes as reference: a tale of caution. *PLoS ONE*, 3:e1854, 2008.

[363] M. Endoh, T. A. Endo, T. Endoh, Y. Fujimura, O. Ohara, T. Toyoda, A. P. Otte, M. Okano, N. Brockdorff, M. Vidal, and H. Koseki. Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity. *Development*, 135:1513–1524, Apr 2008.

[364] M. Kanatsu-Shinohara, J. Lee, K. Inoue, N. Ogonuki, H. Miki, S. Toyokuni, M. Ikawa, T. Nakamura, A. Ogura, and T. Shinohara. Pluripotency of a

single spermatogonial stem cell in mice. *Biol. Reprod.*, 78:681–687, Apr 2008.

[365] J. B. Kim, H. Zaehres, G. Wu, L. Gentile, K. Ko, V. Sebastiano, M. J. Arauzo-Bravo, D. Ruau, D. W. Han, M. Zenke, and H. R. Scholer. Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors. *Nature*, 454:646–650, Jul 2008.

[366] T. S. Mikkelsen, J. Hanna, X. Zhang, M. Ku, M. Wernig, P. Schorderet, B. E. Bernstein, R. Jaenisch, E. S. Lander, and A. Meissner. Dissecting direct reprogramming through integrative genomic analysis. *Nature*, 454:49–55, Jul 2008.

[367] R. A. Miller, N. Christoforou, J. Pevsner, A. S. McCallion, and J. D. Gearhart. Efficient array-based identification of novel cardiac genes through differentiation of mouse ESCs. *PLoS ONE*, 3:e2176, 2008.

[368] K. Ko, N. Tapia, G. Wu, J. B. Kim, M. J. Bravo, P. Sasse, T. Glaser, D. Ruau, D. W. Han, B. Greber, K. Hausdorfer, V. Sebastiano, M. Stehling, B. K. Fleischmann, O. Brustle, M. Zenke, and H. R. Scholer. Induction of pluripotency in adult unipotent germline stem cells. *Cell Stem Cell*, 5:87–96, Jul 2009.

[369] J. B. Kim, V. Sebastiano, G. Wu, M. J. Arauzo-Bravo, P. Sasse, L. Gentile, K. Ko, D. Ruau, M. Ehrich, D. van den Boom, J. Meyer, K. Hubner, C. Bernemann, C. Ortmeier, M. Zenke, B. K. Fleischmann, H. Zaehres, and H. R. Scholer. Oct4-induced pluripotency in adult neural stem cells. *Cell*, 136:411–419, Feb 2009.

[370] R. S. Faustino, A. Chiriac, N. J. Niederlander, T. J. Nelson, A. Behfar, P. K. Mishra, S. Macura, M. Michalak, A. Terzic, and C. Perez-Terzic. Decoded Calreticulin Deficient Embryonic Stem Cell Transcriptome Resolves Latent Cardiophenotype. *Stem Cells*, May 2010.

[371] B. Madan, V. Madan, O. Weber, P. Tropel, C. Blum, E. Kieffer, S. Viville, and H. J. Fehling. The pluripotency-associated gene Dppa4 is dispensable for embryonic stem cell identity and germ cell development but essential for embryogenesis. *Mol. Cell. Biol.*, 29:3186–3203, Jun 2009.

[372] K. E. Galvin, E. D. Travis, D. Yee, T. Magnuson, and J. L. Vivian. Nodal signaling regulates the bone morphogenic protein pluripotency pathway in mouse embryonic stem cells. *J. Biol. Chem.*, 285:19747–19756, Jun 2010.

[373] M. Leeb, D. Pasini, M. Novatchkova, M. Jaritz, K. Helin, and A. Wutz. Polycomb complexes act redundantly to repress genomic repeats and genes. *Genes Dev.*, 24:265–276, Feb 2010.

[374] T. Fei, K. Xia, Z. Li, B. Zhou, S. Zhu, H. Chen, J. Zhang, Z. Chen, H. Xiao, J. D. Han, and Y. G. Chen. Genome-wide mapping of SMAD target genes reveals the role of BMP signaling in embryonic stem cell fate determination. *Genome Res.*, 20:36–44, Jan 2010.

[375] L. Liu, G. Z. Luo, W. Yang, X. Zhao, Q. Zheng, Z. Lv, W. Li, H. J. Wu, L. Wang, X. J. Wang, and Q. Zhou. Activation of the imprinted Dlk1-Dio3 region correlates with pluripotency levels of mouse stem cells. *J. Biol. Chem.*, 285:19483–19490, Jun 2010.