Machine Learning Methods for Life Sciences: Intelligent Data Analysis in Bio- and Chemoinformatics

vorgelegt von
Diplom-Physiker

Johannes Mohr

aus Berlin

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
Dr. rer. nat.

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr.-Ing. Jörg Raisch
Berichter:	Prof. Dr. rer. nat. Klaus Obermayer
Berichter:	Prof. Dr. rer. nat. Sepp Hochreiter

Tag der wissenschaftlichen Aussprache: 19.12.2008

Berlin 2009

D 83

This thesis is dedicated with love
to my wife Sambu
and my daughter Yumi.

Zusammenfassung

In den letzten Jahren haben die experimentellen Techniken innerhalb der Lebenswissenschaften rapide Fortschritte gemacht. Zusätzlich hat die Integration von Methoden verschiedener Disziplinen zur Bildung neuer Forschungsgebiete geführt, wie genetische Bildgebung, molekulare Medizin und biologische Psychologie. Der experimentelle Fortschritt wurde von einem wachsenden Bedarf an intelligenter Datenanalyse begleitet, deren Ziel es ist, einen gegebenen Datensatz unter Einbeziehung von Domänenwissen auf die meistversprechende Art und Weise zu analysieren. Dies schließt die Repräsentation der Daten, die Auswahl der Variablen, die Vorverarbeitung, die Modellannahmen, die Wahl der Methoden für Prädiktion, Modellselektion und Regularisierung ebenso ein wie die Interpretation der Ergebnisse. Das Thema der vorliegenden Arbeit ist die intelligente Datenanalyse in den Bereichen Bioinformatik und Chemoinformatik mit Hilfe von Methoden des maschinellen Lernens.

Das Ziel der genetischen Bildgebung ist es, durch Assoziationsstudien zwischen potentiell relevanten genetischen Variablen und Endophänotypen einen Einblick in genetisch beeinflusste psychiatrische Erkrankungen zu erlangen. Im Rahmen dieser Arbeit werden zwei verschiedene Methoden zur explorativen Analyse entwickelt: Das erste Verfahren basiert auf P-SVM Merkmalselektion für multiple Regression und modelliert additive und multiplikative Geneffekte auf einen Endophänotypen mittels eines spärlichen Regressionsmodells. Die zweite Methode führt ein neues Lernparadigma namens Target Selection ein, um eine Assoziation zwischen einer einzelnen genetischen Variablen und einem multidimensionalen Endophänotypen zu modellieren. Oftmals sind in der Literatur mehrere verschiedene genetische Assoziationsmodelle vertreten, und die Frage ist, wieviel Evidenz ein gemessener Datensatz für jedes dieser Modelle bietet. Zu diesem Zweck wird in der vorliegenden Arbeit eine auf Informationskriterien basierende Modellvergleichsmethode für die genetische Bildgebung vorgeschlagen.

Das Ziel der Analyse quantitativer Struktur-Wirkungs-Beziehungen (QSAR) ist es, die biologische Aktivität einer Substanz anhand ihrer Molekularstruktur vorherzusagen. Traditionell basieren QSAR Methoden auf einer Menge von molekularen Deskriptoren, die zur Bildung eines Prädiktionsmodells benutzt werden. In dieser Arbeit wird eine Deskriptor-freie Methode zur 3D QSAR Analyse vorgeschlagen, welche das Konzept von Molekül-Kerneln einführt, um die Ähnlichkeit zwischen den 3D-Strukturen zweier Moleküle zu erfassen. Die Molekül-Kernel können zusammen mit der P-SVM, einer kürzlich eingeführten Support-Vektor Maschine für dyadische Daten, dazu verwendet werden, explanatorische QSAR Modelle zu bauen, die keine explizite Konstruktion von Deskriptoren mehr benötigen. Die resultierenden Modelle verwenden direkt die struk-

turelle Ähnlichkeit zwischen den vorherzusagenden Substanzen und einer Menge von Support-Molekülen. Die vorgeschlagene Methode wird auf QSAR- und Genotoxizitätsdatensätze angewandt.

Abstract

In the past few years, experimental techniques in the life sciences have undergone a rapid progress. Moreover, the integration of methods from different disciplines has led to the formation of new fields of research, like imaging genetics, molecular medicine and biological psychology. The experimental progress has come along with an increasing need for intelligent data analysis, which aims at analyzing a given dataset in the most promising way taking domain knowledge into account. This includes the representation of the data, the choice of variables, the preprocessing, the handling of missing values, the model assumptions, the choice of methods for prediction, model selection and regularization, as well as the interpretation of the results. The topic of this thesis is intelligent data analysis in the fields of bioinformatics and chemoinformatics using machine learning techniques.

The goal of imaging genetics is to gain insight into genetically determined psychiatric diseases by association studies between potentially relevant genetic variants and endophenotypes. In this thesis, two different methods for an exploratory analysis are developed: The first method is based on P-SVM feature selection for multiple regression and models additive and multiplicative gene effects on an endophenotype using a sparse regression model. The second method introduces a new learning paradigm called target selection to model the association between a single genetic variable and a multidimensional endophenotype. Often, several different models for genetic association are suggested in the literature, and the question is how much evidence a measured dataset provides for each of them. For this purpose, a method for model comparison in imaging genetics is suggested in this thesis, which is based on the use of information criteria.

The aim of quantitative structure activity relationship (QSAR) analysis is to predict the biological activity of compounds from their molecular structure. Traditionally, QSAR methods are based on extracting a set of molecular descriptors and using them to build a predictive model. In this thesis, a descriptor-free method for 3D QSAR analysis is proposed, which introduces the concept of molecule kernels to measure the similarity between the 3D structures of a pair of molecules. The molecule kernels can be used together with the P-SVM, a recently proposed support vector machine for dyadic data, to build explanatory QSAR models which do not require an explicit descriptor construction. The resulting models make direct use of the structural similarities between the compounds which are to be predicted and a set of support molecules. The proposed method is applied to QSAR- and genotoxicity datasets.

Contents

1 Introduction

1.1 Machine Learning for Life Sciences

Life sciences encompass all research on structures of living beings and processes in which living beings are involved. Apart from general biology, this includes also disciplines like medicine, genetics, pharmacy, toxicology, nutrition science, biochemistry, biophysics, psychology and neurobiology. In the past few years, life science related technologies have made a huge progress. For example, the development of functional magnetic resonance imaging in the early nineties has been a breakthrough in the field of brain imaging, since it enabled researchers to noninvasively record brain signals at a resolution of a few millimeters. An example from molecular biology is the gene expression microarray, a high-throughput technique which allows to assess the expression levels of thousands of genes in parallel. Moreover, experimental techniques from different fields are now often combined, allowing the integration of different measurements in order to uncover the mechanisms of life and the natural environment. Examples are the recently evolved field of genomic imaging, in which brain imaging techniques like (functional) magnetic resonance imaging are combined with genotyping (Mohr et al., 2006), or the combined analysis of psychological scores and genetic markers (Lohoff et al., 2008). As a consequence of this synergy among different subjects, the field of life sciences has become increasingly interdisciplinary.

These advances in experimental techniques have come along with the need for intelligent data analysis. In most studies, multiple measurements and other attributes are combined, resulting in heterogeneous datasets with a large number of variables, which might be of different type, have different data range, or contain missing values. Frequently, the available sample size is relatively small, because some of the measurements are either costly or difficult to obtain, as it is often the case in genetics (Mohr et al.,

2008d) and medical research (Mohr et al., 2008b). The final goal of data analysis is either to illicit possible relations between the measured variables or to build predictive models. Usually, only a subset of the variables is relevant for the given problem, and both model selection and model regularization are required. Given a specific research question, the task of intelligent data analysis is to apply methods which are most useful and revealing for a given problem domain and dataset. Specifically, it needs to be decided which variables are included in the analysis, what preprocessing is applied, what is done with missing values, what model assumptions are made a priori, which data analysis methods and statistical tests are applied, what model selection and regularization techniques are employed and, finally, how the results are interpreted. All this should be done based on the available domain knowledge and the data characteristics.

If the analysis is confirmatory, i.e. the goal is to test a given hypothesis, only the variables involved in this hypothesis need to be measured, and standard statistical tests are sufficient. However, if previous studies have led to ambiguous results and several candidate hypotheses are available, then all these different models need to be compared on the current dataset. In general, this requires adequate procedures for model comparison. A third case is the fully exploratory analysis, where a large number of measurements and attributes are collected, which according to the experience and theories of the domain expert are of potential importance to the research question. Since there is a large number of possible hypotheses (or models), much thought has to be put into the choice of analysis techniques and model assumptions, in order to avoid data dredging and multiple testing issues.

In life sciences, many objects of research come naturally as structured data, i.e. they can be expressed as a sequence, tree or graph. For example, DNA can be viewed as sequence of nucleotides, and chemical molecules can be represented as graphs. However, most statistical or pattern analysis methods are based on a representation of the data as vectors in an Euclidean space. For this reason, structured data is often transformed into Euclidean vectors by describing the objects through a series of characteristics. However, this description only captures certain aspects of the original objects, and some of the information which was contained in the structure is usually lost. Another problem is that it is usually not clear a priori which characteristics or features of the objects are relevant for the task. Moreover, the descriptor variables are usually quite heterogeneous, since they capture very different aspects of the original objects. Thus it is doubtful, whether it is justified to embed these variables into a Euclidean space and use a Euclidean distance measure. Therefore, it is desirable to derive methods which can directly work with structured data (Bakhir et al., 2007).

This thesis focuses on intelligent data analysis in different areas of the life sciences. Four machine learning methods are proposed and applied to problems in bioinformatics and chemoinformatics. Similar to life sciences, the still very young research area of machine leaning has undergone a rapid advancement over the last three decades. This development started at the end of the eighties, when data-driven approaches were applied to problems that the artificial intelligence field of computer science previously attempted to solve by rule-based methods, such as expert systems. The principles of

neural information processing in the brain inspired artificial neural networks (ANNs), where a parallel architecture of a large number of simple, interconnected units is used to obtain a distributed representation of information, and where the adjustment of the connection weights is data-driven and follows local learning rules. The theoretical foundations of neural networks were already laid in the forties, fifties and sixties of the last century (McCulloch and Pitts, 1943; Rosenblatt, 1958), but a major breakthrough came in the eighties, when principles from physics (Hopfield, 1982; Kirkpatrick et al., 1983; Ackley et al., 1985) and information theory (Linsker, 1988) were applied to ANNs. Moreover, new architectures, like the RBF nets (Broomhead and Lowe, 1988), and new learning algorithms based on cost functions, such as the backpropagation algorithm for multilayer perceptrons (Rumelhart et al., 1986), were proposed. Another important milestone of the 1980s was the development of self-organizing maps (Kohonen, 1982b,a), using competitive learning. In the nineties, concepts from statistical learning theory (Vapnik and Chervonenkis, 1991) lead to large-margin methods (Vapnik, 1998), such as support-vector machines, and other kernel methods (Schölkopf and Smola, 2002). Several mathematical areas, for example graph theory, probability theory, statistics and estimation theory, gave rise to new sub-areas of machine learning, such as independent component analysis (Hyvärinen et al., 2001), Gaussian processes (Williams and Rasmussen, 1996) and probabilistic graphical models (Jordan et al., 1998; Edwards, 2000; Cowell et al., 1999). While originally, most machine learning algorithms were designed for vectorial data, an increasing number of methods for structured data, like sequences, trees, and graphs have been developed in recent years (Bakhir et al., 2007).

Like life sciences, machine learning has thus evolved to be a very diverse and interdisciplinary field of research. Similar to the effect the new technological developments had on experimental life sciences, the increasing availability of computational power facilitates the application of computationally intensive machine learning techniques for intelligent data analysis. Further details on the history of machine learning can for example be found in (Haykin, 1994; Duda et al., 2001; Ripley, 1996). In the following, a brief overview on some principal concepts of machine learning will be given.

Machine learning is an area of research which is concerned with the development of mathematical principles and algorithms to find a suitable mathematical description of some process or phenomenon based on collected or measured data. Commonly, the observations or measurements are subject to noise or stochasticity. In the context of machine learning, they are interpreted to be outcomes of random experiments mapped by random variables to a set of numerical values. The purpose for building a model can be to learn something about the relations between the random variables, to understand the underlying mechanisms or to make predictions on new data. While in the first two cases the model has to be explicit enough to allow interpretation, in the third case a 'black box'-model would also be sufficient. The process of building models can be based on either inductive or deductive principles. In deductive reasoning, general principles are used to infer specific conclusions. Inductive reasoning, on the other hand, uses collected information to derive a general principle. The process of data-driven modeling

using computational statistical techniques is an example of inductive learning.

Broadly, one differentiates between *supervised learning*, where a *teacher* provides target values or a reinforcement signal, and *unsupervised learning* where the goal is to find structure in the data, while no external feedback is available. Examples for supervised learning methods are neural networks and support vector machines, examples for unsupervised learning methods are clustering algorithms, self-organizing maps and component analysis. These two types of learning are combined in *semi-supervised learning*, where both labeled and unlabeled examples are employed. In supervised learning an explicit model is constructed from labeled *training data*. Once the model is constructed, the training data is no longer needed, since prediction on the *test data* is done using the model alone. If the goal is only prediction, also *transductional learning* can be used, where inference is based directly on the knowledge of labeled and unlabeled data at prediction time.

In every modeling process, some assumptions have to be made, either explicitly or implicitly. This usually involves assumptions about the dataset used to build the model, e.g. the assumption that the data points are independent and identically distributed (i.i.d.), and assumptions about the underlying process which generated the data. In principle, the whole scientific process relies on a series of decisions and made assumptions, starting with the phrasing of a research question, followed by the design and conduction of an experimental or observational study and the final analysis of the data. Among other things, a scientist choses which variables need to be measured or collected to investigate a certain phenomenon. At the point of data analysis the researcher has to decide which of these variables should be actually included in the analysis and in what fashion, what kind of models will be considered and which statistical methods should be employed for model building and testing. While for unsupervised learning the quality of a model is usually judged by subjective criteria (depending on the specific purpose), in supervised learning an objective criterion for model comparison is provided by the prediction performance on yet unseen data from the same distribution as the training data. An important point is that a model needs to be falsifiable. The scientific modeling process as a whole consists of an iterative procedure: Based on some observations a model is build. The model is able to make predictions which can be compared with the actual observations. The result is then used to validate and, if necessary, to improve the model.

No matter how good the modeling approach, one has to bear in mind that statistics cannot be used to prove causality ('*correlation does not imply causation*'). This is due to the fact that in real-world situations there is always an infinite number of factors which could influence the random variable of interest. While a thoroughly working researcher will always try to consider all factors or covariates which might influence the target, it still remains unknown whether all relevant variables have been measured and included in the analysis. A statistical dependency between two variables could either indicate a direct causal effect or result from a common influence of an unmeasured 'hidden' variable. Here it is important to note that standard statistical procedures, like analysis of variance and regression analysis, as well as the machine learning methods proposed

in this thesis always presume a specific effect model, specifying *input* variables (also called *independent* or *predictive* variables) and an *target* variable (also called *dependent* or *output* variable). The assumed direction from input to target does not necessarily correspond to the direction from cause to effect.

1.2 Organization of this Thesis

This thesis is divided into four chapters. In the first three chapters novel approaches for bioinformatics are suggested. Three machine learning methods for genotype-phenotype analysis are introduced and applied to problems in genomic imaging. The final chapter focuses on learning with structured data in the area of chemoinformatics. A kernel method for molecules is proposed and applied to quantitative structure activity relationship (QSAR) analysis and genotoxicity prediction.

The goal of genotype-phenotype association studies is to model potential genetic influences on certain observed characteristics of a living being. In this thesis, the special case of population association studies that involve a set of candidate polymorphisms and some quantitative phenotypes is considered. Conventional statistical methods can either not model additive effects, or cannot properly account for multiplicative interactions if the sample size is small. In addition, they require corrections for multiple testing which leads to a reduction in power. Moreover, conventional methods are not suitable to analyze multidimensional phenotypic patterns associated with the same genetic variable. In order to overcome these limitations, two novel statistical tests based on machine learning techniques are proposed in chapters 2 and 3 of this thesis.

Chapter 2 introduces a novel method for detecting both additive and multiplicative (epistatic) effects of several polymorphisms on a real-valued phenotype (Mohr et al., 2008b). The method, called machine learning genotype-phenotype analysis (MLGPA), assumes a multiple regression model in which a phenotype is modeled as a linear combination of genetic variables under the assumption of Gaussian noise. The prediction function is chosen by a learning machine which combines the Potential Support Vector Machine (P-SVM) for variable selection with ordinary least-squares regression. The model complexity is adjusted using an information criterion. A leave-one-out cross-validation estimate of the generalization error is then used as test statistic, and the hypothesis of independence between predictors and target is tested via a label permutation test. The method is applied to study synergistic effects of the dopaminergic and glutamatergic system on hippocampal volume in alcohol-dependent patients.

In chapter 3, an alternative model assumption is made, in which a single genetic variable influences a multidimensional phenotype. For this purpose, a new learning paradigm called target selection is introduced (Mohr et al., 2008e). The target selection paradigm is constituted by a generative model, in which there is a set of discrete random variables $\mathbf{Y}$, one of which (Y_g) influences the distribution of a second set of variables $\mathbf{X}$. This gives rise to an optimal classification model, which specifies the class probabilities of the relevant target variable Y_g given the input $\mathbf{X}$. The learning task in target selection is to recover this model as good as possible from a given dataset

by selecting one of the target variables and using a probabilistic classification method to fit a parametric model. An objective function is proposed for this setting, which is derived from the concept of mutual information. Moreover, a learning method for target selection is suggested based on a probabilistic C-SVM as classification method. The new paradigm and the proposed algorithm for target selection are applied to genotype-phenotype association analysis. For this purpose, the vector of phenotypes is identified with the input variables of a learning machine, while the binary genetic variables are identified with the targets. A label permutation test is used to assess whether an established dependency between the multidimensional phenotype and one of the genetic variables is significant. The proposed method is evaluated on the dataset from the study in chapter 2, which is re-analyzed with the target selection method under the above described alternative model assumptions.

The methods in chapters 2 and 3 limit the use of prior knowledge to the choice of candidate markers and phenotypes and some assumptions about the general nature of the models. Often, however, there is already a limited set of certain candidate hypotheses (or models) available, either from previous studies or the literature. Then the goal is to compare the set of a priori models based on a given dataset. In the framework of intelligent data analysis, such a study can be seen as part of a larger modeling process, in which multiple hypotheses are entertained. Whenever relevant experimental or observational data is collected, it will lend more support to some of the hypotheses and less support to others. Repeated collection of data (also by other research groups) will then over time lead to a refinement of the 'working set' of model hypotheses. In chapter 4 of this thesis, information criteria (ICs) are employed to compare a given set of candidate regression models on a genomic imaging dataset (Mohr et al., 2008c). Two theoretically well-founded information criteria, the *Akaike Information Criterion (AIC)* and the *Bayesian Information Criterion (BIC)*, are used to compare the different candidate hypotheses. In contrast to a model comparison based on likelihood-ratio tests, the model comparison based on ICs does not require the models to be nested. Moreover, all models can be compared at once, there is no need for pairwise comparisons.

The aim of quantitative structure activity relationship (QSAR) analysis is to predict the biological activity of the toxicity of a chemical compound based on its molecular structure. Traditionally, a set of molecular descriptors is extracted from the structure, embedded into a Euclidean vector space and used for building a predictive model. However, for structured data like molecules it is usually advantageous to employ algorithms which can directly work with the structural representation itself. Therefore, in chapter 5 a novel QSAR approach is proposed that can directly work on molecular structures and is independent of the spatial pre-alignment of the compounds (Mohr et al., 2008a). It introduces the concept of a molecule kernel to measure the similarity between the 3D structures of two molecules. Predictors can be build using the molecule kernel in conjunction with the potential support vector machine (P-SVM), a recently proposed machine learning method for dyadic data. The resulting models make direct use of the structural similarities between the compounds in the test set

and a subset of the training set and do not require an explicit descriptor construction. A problem of the descriptor-based methods is that they work in the quite abstract descriptor space, so that the resulting models can be used for prediction, but do not allow an insight into the underlying mechanisms and are hard to interpret. In drug development, however, there is a high interest in explanatory models, as they can be used by the expert to suggest specific modifications which influence the activity or toxicity of a compound. In this work, it is shown how the molecule kernel method can be used to obtain such explanatory models. Furthermore, a method for visualizing these models is suggested. The molecule kernel method is applied to QSAR and genotoxicity (chromosome aberration test) prediction, and the generalization performance is compared to state-of-the-art techniques. This evaluation shows that the proposed method outperforms its competitors in both application areas.

1.3 A Short Primer on Genetics and MRI

In the following, some concepts from genetics which are used in chapters 2, 3 and 4 will be shortly reviewed. Since the phenotypes analyzed in these chapters are based on magnetic resonance imaging (MRI) measurements, also a quick explanation of the basic physical principles of (functional) MRI is given. This is not meant to be an exhaustive introduction to these fields. Instead, the purpose of this primer is to familiarize readers not working in these areas with some basic concepts and technical terms used in this thesis.

1.3.1 Genetics

The total genetic information of a living being is referred to as the genome. For humans, the genome consists of 23 pairs of chromosomes, which are contained in every cell. The main chromosomal component is desoxyribonucleic acid (DNA), which is the carrier of genetic information. The molecular structure of DNA consists of two intertwined chains held together by hydrogen bounds. The main elements of DNA are the purine bases guanine (G) and adenine (A) and the pyrimidine bases cytosine (C) and thymine (T), where G pairs exclusively with C, and A with T.

During the process of transcription, the DNA is transformed (by an enzyme called RNA polymerase) into messenger ribonucleic acid (mRNA), which is a single-stranded polynucleotide that is complementary to DNA, but uses uracil (U) instead of thymine (T). In addition to the sequences of DNA which are coding for proteins, the exons, there are also regulatory sequences which direct and regulate protein synthesis. They lie either before the coding sequence (5' untranslated region, or 5'UTR) or behind the coding sequence (3' untranslated region, or 3'UTR). In addition, there are also non-coding intragenic regions, or introns, which are removed during the transcription of DNA into mRNA by a process called splicing.

The mRNA is then translated by ribosomes into proteins that have biochemical or structural functions, a process called translation. The sequences of nucleotides code for

20 different amino acids which form the building blocks for proteins. The amino acids combine in sequence to form a polymer (polypeptide chain), which then folds into a specific 3D shape. Proteins are made up of one or more chains, and act as catalysts (enzymes).

The process from DNA to proteins is summarized in Figure 1.1. The field of genomics is concerned with the study of the genome. In transcriptomics, the expression level of mRNAs in a given cell population is studied. For example, DNA microarrays are used as a high throughput technique to reveal the relative amounts of mRNA in a cell. However, the number of protein molecules which are synthesized based on a given mRNA molecule not only depends on the mRNA expression level but also on the translation-initiation features of the mRNA sequence. The set of proteins expressed at a given time under defined conditions is studied in the area of proteomics.

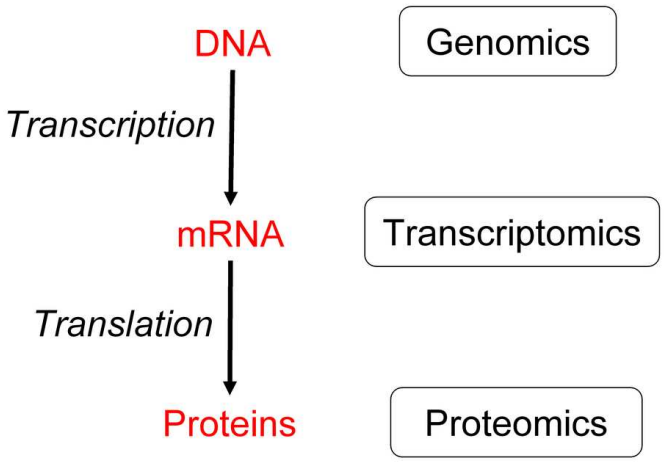


Figure 1.1: Genomics, Transcriptomics and Proteomics

A single nucleotide polymorphism (SNP) represents an exchange of one single nucleotide at a specific locus within the genome sequence (see Fig. 1.2) that is transmitted to all offspring. While in principle, there could be four different alleles (forms), corresponding to the four nucleotides, almost all common SNPs have a low mutation rate and therefore only two alleles. For such a diallelic SNP, one distinguishes the more frequent form (the major allele) from the less frequent form (the minor allele) within a population. For a variation to be considered a SNP, the minor allele must occur in at least 1% of the population. SNPs can be found at approximately every 300-1000 base pairs and form the main basis of inter-individual differences in humans. More than 10 million common SNPs exist in the human genome.

Loci close to each other on a chromosome are more frequently inherited together (linkage), because they are less likely to be separated by chromosomal crossover. Crossover occurs during meiosis, when matching regions on matching chromosomes break and then reconnect to the other chromosome. This breakage and rejoining of parental

chromosomes results in an exchange of genes and is called genetic recombination. A non-random association of the alleles at two loci (not necessarily on the same chromosome) within a population sample is called linkage disequilibrium (LD).

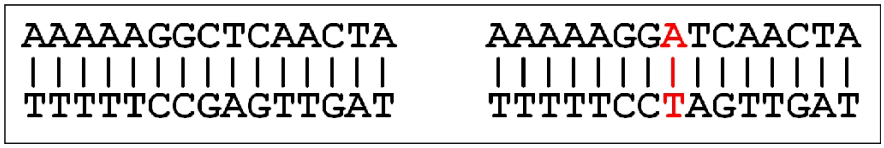


Figure 1.2: Single Nucleotide Polymorphism
On the left the original double-stranded DNA is shown, on the right the mutated DNA with the introduction of a SNP, in this case a substitution from C to A. The corresponding base on the second chain is altered accordingly.

A haplotype is a combination of alleles at different loci on the same chromosome. Since each individual has two chromosomes, each person has two haplotypes, the set of which is often called the diplotype. In the measurement of a SNP (genotyping) usually the phase information cannot be obtained, which means that it cannot be determined which allele came from which chromosome. Thus the measurement of a SNP results in three different genotypes: Two homozygous (AA, aa) and one heterozygous (Aa), where A refers to the major and a to the minor allele. Other genetic markers, like microsatellites, can have more than two alleles, and have more then three different genotypes. A haplotype block is a set of closely located loci which are inherited together. These concepts are illustrated in Figure 1.3. Much of the human genome is organized in a block-like structure, with no or little recombination occurring within the blocks, only at their boundaries.

Since experimental techniques to directly obtain haplotype information are too costly, the haplotype configuration for each individual needs to be inferred by statistical methods. Various methods for haplotype reconstruction (phasing) have been proposed, which, depending on the used optimization criterion and algorithms, yield slightly different results. Moreover, the assignment of haplotypes to individuals is probabilistic. While knowing the true diplotype would be more informative than knowing only the genotype, this is usually no longer the case for the inferred diplotype, due to the phase uncertainty (Balding, 2006). However, for a high linkage disequilibrium between markers, the uncertainty in the individual haplotype assignments is quite low, and it is valid to use the most likely haplotype configuration in the further analysis.

In this thesis, only data from population association studies is analyzed. In population association studies a number of candidate polymorphisms are genotyped for a collection of unrelated individuals. Unrelated means that the relationships between the subjects are assumed to be only distant. The data used in a population study should always be tested for Hardy-Weinberg equilibrium (HWE), which holds at a locus in a population when the two alleles within an individual are statistically independent. A deviation from HWE could for example be be caused by inbreeding, or by population

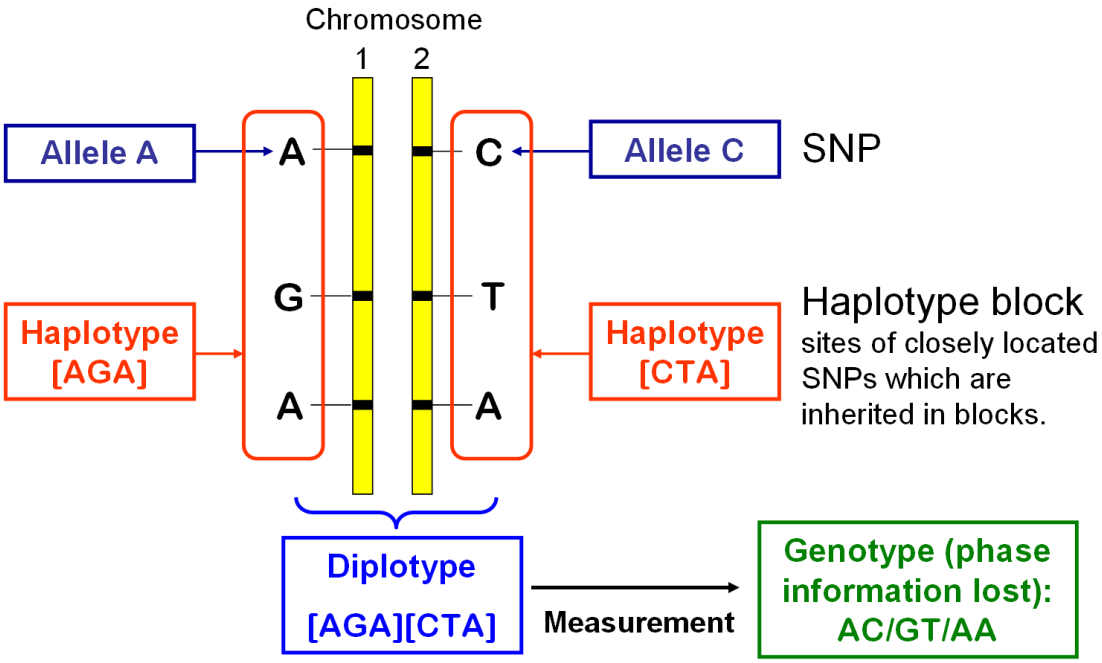


Figure 1.3: Haplotype, Diplotype and Genotype

stratification, where the sample contains subgroups of individuals which are on average more closely related to each other than the rest of the population (Balding, 2006). Loci for which deviations from HWE are found using the χ^2 test or Fisher's exact test are usually discarded.

A phenotype is an observable characteristic of an organism. Genotype-phenotype association studies aim at identifying genetic patterns that influence certain phenotypes. In general this could be genetically inherited traits of person, e.g. the eye color. In a medical context, however, the phenotypes usually consist of a status variable, indicating whether a person is suffering from a certain disease, or a so-called endophenotype, a quantitative trait associated with the disease, e.g. blood pressure. In the first case, the phenotype is encoded as a nominal variable, in the second case it is represented by a quantitative variable. Diseases with a genetic component are divided into two categories: While in simple disease a single genetic factor is involved, in complex diseases an influence of several genetic factors is assumed, interacting with each other and the environment.

In general, a genotype-phenotype analysis can either be based on SNPs or on haplotypes. For SNPs, there are two possible models of allele interaction, the dominant and the linear model. In the dominant model, the effect depends on whether the dominant allele is present or not. In an exploratory analysis, it is usually not clear which allele is dominant, therefore a dominant model needs to be considered for each of the two alleles. In the linear model, it is assumed that the heterozygous case lies directly between the two homozygous ones. Even if the SNP that actually causes the effect

on a phenotype is not genotyped, the SNP might be in LD with one of the genotyped SNPs, in which case a genotype-phenotype association can be found, though it is not directly causal.

Analysis based on haplotype makes use of the block-like structure of the genome, and leads to a dimension reduction by accounting for the high correlations of polymorphisms within such a block. The reason for this is that some of the many possible allele combinations at the loci within a haplotype block will not or very rarely occur. Because haplotypes include frequently occurring combinations of alleles at different loci, they can implicitly account for some epistatic effects among the SNPs. However, haplotype-based models are subject to the uncertainty in the phasing. It is also not quite clear how to deal with rare haplotypes: Leaving them out, leads to a loss of information and possibly a loss in power, while including them increases the dimensionality of the problem. Moreover, the block boundaries will depend on the phasing algorithm used, its parameters, the spread of the SNPs across the gene, and the sample population. Also the block structure is usually not hard, but soft, i.e. there can be correlations between SNPs from different blocks, and some recombinations could have occurred within a block. Therefore, the hard block structure is usually an idealized view of reality which is only valid in approximation.

1.3.2 Magnetic Resonance Imaging

This thesis focuses on the study of complex psychiatric diseases, like alcohol dependence or schizophrenia. In complex psychiatric diseases, associations with the disease status are often obscured, since the disease risk is probably influenced by a large variety of genetic and environmental factors. Therefore, current psychiatric research aims at analyzing genetic influences on intermediate phenotypes more directly related to neurobiology. One way to obtain such endophenotypes is the use of brain imaging techniques, such as magnetic resonance imaging (MRI) and functional magnetic resonance imaging (fMRI).

Magnetic resonance imaging (MRI) measurements are based on the fact that protons have a quantum mechanical property called spin, which leads to a magnetic moment. In the strong static magnetic field of an MRI scanner, the spins align either parallel or antiparallel to the field vector $\mathbf{B}_0$. The spin magnetic moments precess around an axis along the direction of the field at the Larmor frequency ω_0, which is proportional to the strength of the external magnetic field and is given by

$$\omega_0 = \gamma \left\| \mathbf{B}_0 \right\| , \tag{1.1}$$

where γ is the gyromagnetic ratio.

Since the parallel alignment corresponds to a lower energy level and is preferred, there is a small net magnetic moment. An electromagnetic radio-frequency (rf) pulse at the Larmor frequency is able to transfer some of its energy to the protons, lifting them in the higher energy state. This leads to a reduction in the magnetization in the direction of $\mathbf{B}_0$ (longitudinal magnetization). At the same time, the rf pulse synchronizes the

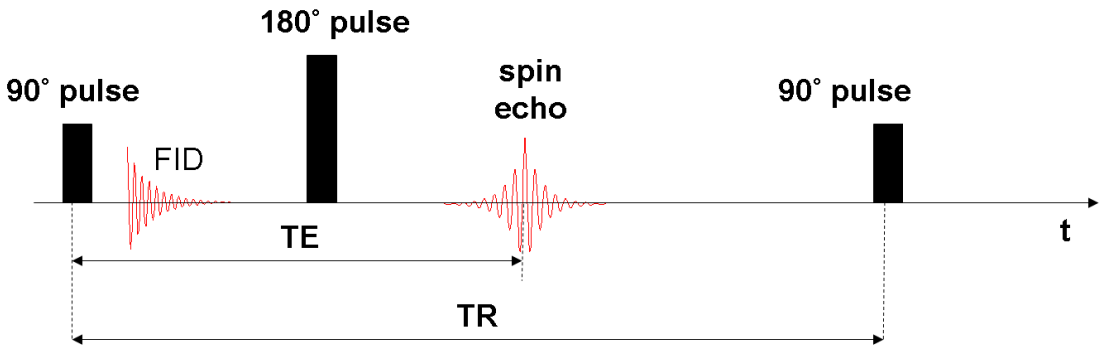


Figure 1.4: Spin-Echo Sequence

phase of the precession, leading to a magnetic moment in the plane orthogonal to the external field (transversal magnetization). A pulse of exactly the right duration to tilt the net magnetic moment orthogonal to the direction of $\mathbf{B}_0$ is called a 90° pulse. The transversal component of the magnetization can be measured as an oscillatory signal induced in the receiver coil, which is called free induction decay (FID).

The moment the radio frequency pulse is switched off, the longitudinal magnetization increases again, a process described by the longitudinal relaxation time constant T_1 (ranging from 300 to 2000 ms). The reason for this lies in the so-called spin-lattice relaxation: the spins precessing at the Larmor frequency want to transfer energy to the magnetic fields of their surroundings (the lattice). This energy transfer can be done most effectively if the precession of the magnetic fields in the lattice is close to the Larmor frequency, and if the lattice molecules are medium-sized and slow-moving. Therefore, the composition of the surroundings influences T_1: The small and fast-moving water molecules lead to a long T_1, whereas fat molecules result in a short T_1.

Independently from the longitudinal relaxation, perturbations by magnetic fields of other spins and the macromolecules in the tissue cause a random change of the precession frequency of each proton, leading to a dephasing of proton spins. This decay process is characterized by the transversal relaxation time T_2, which lies in the range of 30-150 ms. T_2 in surroundings consisting of water is longer then in less pure liquids containing larger molecules. The reason for this is that larger molecules move less quickly, so their local magnetic fields do not cancel out that well. In addition to this so-called spin-spin relaxation, local inhomogeneities of the macroscopic field and susceptibility variations of the tissue (e.g. at air/tissue interfaces, component changes) cause the total magnetization in the volume element to decrease. This effect is characterized by the much shorter relaxation time $T_2^* \approx \frac{1}{3}T_2$, which is observed as FID signal.

In a so-called spin-echo sequence, a 90° rf pulse is followed by a 180° rf pulse (Figure 1.4). After the 90° rf pulse, the spins start to dephase. The 180° rf pulse inverts the spin magnetic moments, so that the dephasing spins run back into phase, causing the

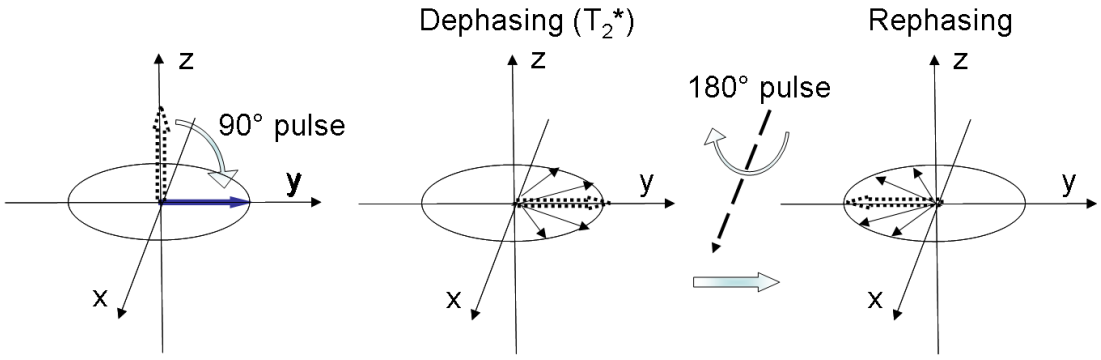


Figure 1.5: Dephasing of the spins due to T_2^* decay and rephasing after the 180° pulse of the spin-echo sequence

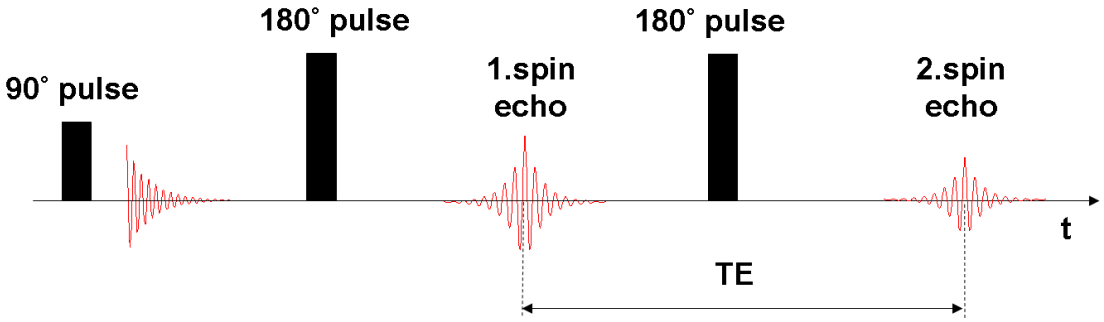


Figure 1.6: Multi-Spin-Echo Sequence

spin-echo signal (Figure 1.5). Important sequence parameters are the time to echo (TE), which corresponds to the time from the 90° pulse to the echo, and the time to repeat (TR), which indicates the time to a repetition of the sequence. If a sequence of several 180° pulses is used to rephase the dephasing protons, several spin echoes are obtained, and one speaks of a multi-spin-echo sequence (Figure 1.6). The amplitudes of the echoes are independent of the inhomogeneous fields which can be considered stationary during the TE, and the signal decrease from one echo to the next is given by pure T_2 decay.

By varying TE and TR, different weightings (e.g. T_1, T_2, proton-density) can be realized, corresponding to different tissue contrasts. In T_1-weighting the white matter in the brain appears white, the gray matter gray and the cerebrospinal fluid (CSF) dark, while in T_2 and T_2^* weighted images, these contrasts are reversed. Since different tissues have different relaxation times and proton densities, specific pulse sequences can be used to obtain different contrasts. By additional application of magnetic field gradients, 3D images coding for the local signal strength can be obtained. Magnetic resonance imaging (MRI) can thus be used to obtain anatomical 3D images of the brain

with a good soft tissue contrast.

Several "fast" sequences have been developed which allow a more rapid acquisition of MR images. These are often based on so-called gradient echoes, which are generated as following: A short application of a gradient field amplifies local field inhomogeneities, speeding up the dephasing of the protons. Application of the same gradient field in the opposite direction leads to a rephasing of the protons, and an echo signal. Since the time-consuming application of the 180° pulse is avoided, this allows for a much shorter TR than the spin-echo sequence. Moreover, instead of a 90° pulse, the fast sequences use smaller flip angles α ($10° \leq \alpha \leq 35°$), so that the longitudinal magnetization is not totally canceled out. This leaves enough longitudinal magnetization for the next pulse even for a short TR. In addition, for a smaller flip angle the rf pulse duration is much shorter.

With the special technique of functional magnetic resonance imaging (fMRI) the blood-oxygenation-level-dependency (BOLD) signal is measured. This signal measures the haemodynamic response and depends on many factors, e.g. the ratio of oxyhemoglobin to deoxyhemoglobin, the blood volume and the blood flow. It is generally assumed that the BOLD signal is an indirect measure of neural activity. The rationale behind this is that an increased neural activity leads to an increased demand for oxygen, which is overcompensated by the vascular system, increasing the amount of oxygenated hemoglobin relative to deoxygenated hemoglobin, which in turn increases the BOLD signal. The rapid technique of echo-planar imaging (EPI) can be used to produce functional MRI images at video rates.

2

MLGPA to Model Epistatic Genetic Effects on a Quantitative Phenotype

2.1 MLGPA

In this chapter, a special case of population association analysis is considered, in which the goal is to test for potential associations between a small number of candidate polymorphisms and a quantitative phenotype while taking epistatic effects (additive and multiplicative interactions between alleles) into account. Whereas epistasis was initially used to depict the suppression of a phenotypic characteristic of one gene by another, here the expression epistatic effects refers to the more recent denotation describing common gene-gene-interactions of both additive and multiplicative nature.

Association analysis for single SNPs can be conducted with either analysis of variance (ANOVA), linear regression or tests for group difference. However, if epistatic interactions between several SNPs need to be considered, analysis becomes more involved. For ANOVA, the inclusion of such effects leads to a large number of parameters which need to be estimated. If the number of parameters exceeds the number of examples, the parameters are not estimable anymore. Moreover, with increasing number of included SNPs and interaction variables, the results become increasingly unreliable and are difficult to interpret. Also, in order to assess which groups are different from which, a large number of follow-up tests need to be carried out, which requires corrections for multiple testing. An alternative to ANOVA, which can also be applied in the case of many genetic variables and few examples, is the conduction of separate tests for group differences, one for each single SNP and each interaction variable. This requires the conduction of separate tests for both models of allele interaction (dominant and linear), since in an exploratory analysis both models are possible. Therefore, a large

number of tests has to be carried out, for which multiple testing corrections need to be applied. Two severe drawbacks of this this approach are its inability to model additive effects between several genetic variables and the necessity for multiple testing corrections, which lead to a reduction in statistical power. These problems often discourage researchers to conduct additional analysis beyond single SNP tests, even though such an analysis might be very revealing Balding (2006).

In order to overcome these issues, a method for finding potential genotype-phenotype associations based on machine learning techniques is introduced in this chapter. The so-called Machine Learning Genotype-Phenotype Analysis (MLGPA) allows to test for associations between a set of candidate genetic markers and a quantitative phenotype, whenever there is no prior hypothesis which and how many of the candidate markers are potentially involved and which model of allele interaction should be assumed. In MLGPA, dependencies are detected by using an unbiased estimate of the generalization error of a learning machine as test statistic. A significantly low estimated generalization error implies a dependency between genes and phenotype which the learning machine was able to learn. In this case, a regression model learned by the learning machine on the dataset can be expected to generalize well to unseen examples.

2.1.1 Generative Model

In MLGPA the following generative model for the data is assumed: For the case where a dependency between the measured genetic variables and phenotype exists, the phenotype is assumed to be generated by a linear combination of a sparse set of genetic regressors (from the larger set of measured genetic variables) and additive noise. The regressors are binary (0/1) variables encoding either the absence/presence of an allele of a single SNP or the simultaneous presence of two specific alleles of two separate SNPs ("interaction variable"). Since each SNP is encoded by two binary regressors indicating the presence of the major or minor allele, which can be combined in a linear combination, this model includes both the dominant and the linear model of allele interaction. The sparseness of the set of regressors accounts for the assumption that from a measured set of genetic variables only very few (if any) are likely to influence a given phenotype. For the case where no genetic dependency between the measured genetic variables and phenotype exists, the phenotype is assumed to be pure noise. Due to this generative model, the functional dependency between a set of independent variables (genetic variables) to a quantitative dependent variable (phenotype) is described in MLGPA by a multiple linear regression ansatz,

$$\hat{y}(\mathbf{x}^{(k)}) = \sum_{i=1}^{d} a_i \cdot x_i^{(k)} + c, \tag{2.1}$$

where $\hat{y}(\mathbf{x}^{(k)})$ denotes the value of the dependent variable predicted for case k, $x_i^{(k)}$ the value of the i^{th} input variable for case k, a_i the corresponding regression coefficient, c the offset and d the number of input variables. Therefore, the regression coefficients

parameterize a set of prediction functions. Due to the sparseness assumption, most regression coefficients are assumed to be zero, motivating the search for parsimonious models via the use of variable selection.

Note that the above generative model does not correspond to an ANOVA model. An ANOVA model conventionally does not include complex interaction terms for factors unless it also includes all simpler terms for those factors, so that usually the full set of parameters is fitted. In contrast to this, MLGPA does not conduct an analysis of variance, but is based on the assumption of a sparse generative model, which should be recovered from the data as well as possible in order to have good generalization performance. Therefore, variable selection is used to exclude unnecessary parameters from the model before fitting the coefficients. Thus the inclusion of an interaction variable in the regression model does not necessary imply the inclusion of the variables of the corresponding single alleles, unless this is warranted by the additional information they provide about the phenotype. Obviously, the regression coefficients of MLGPA do not correspond to main or interaction effects of a traditional ANOVA model.

2.1.2 Missing Values

Sometimes, a dataset contains missing values for SNPs. MLGPA resolves this issue by using a probabilistic approach, where the probabilities of the respective allele configurations are estimated from the frequencies in the dataset. Taking the expectation with respect to the independent variables yields a regression ansatz with probabilistic instead of binary independents. In the presence of missing values, instead of eq. 2.1 the following regression function is used,

$$\hat{y}(E(\mathbf{x}^{(k)})) = \sum_{i=1}^{d} \alpha_i \cdot \Big[P(x_i^{(k)} = 1) \cdot 1 + P(x_i^{(k)} = 0) \cdot 0 \Big] + c \tag{2.2}$$

$$= \sum_{i=1}^{d} \alpha_i \cdot \Big[P(x_i^{(k)} = 1) \Big] + c \tag{2.3}$$

where $E(\cdots)$ denotes the expectation. After the probabilistic input variables have been constructed, redundancies in the dataset are removed; that is input variables which are perfectly correlated or anti-correlated on the dataset cannot be distinguished and are summarized in a single variable.

2.1.3 Variable Selection

In traditional regression analysis the parameters of all regressors are fitted on the whole dataset yielding the maximum likelihood solution in which the mean squared error (MSE) is minimized. Then, statistical tests are conducted whether any of the regression coefficients are significantly different from zero. A critical issue is the selection of variables to be included into the regression equation. Even if a predictive variable does not have an effect on the dependent variable in the population, so that the

'true' regression coefficient is zero, the corresponding maximum likelihood parameter estimate on a finite sample will usually be non-zero, which also influences the regression coefficients of other variables. Therefore, inclusion of many irrelevant variables leads to an increased variance of prediction and imprecise inference. Moreover, if more predictive variables are present than observations, the parameters are not estimable anymore. Both of these issues apply to imaging genetics studies, where a large number of potentially relevant genetic predictive variables need to be evaluated, often exceeding the available sample size, and it is usually assumed that only a small number of these predictors (if any) will actually influence the phenotype in the underlying population. Sometimes, variable selection procedures like forwards selection, backwards elimination and subset selection are used to reduce the number of regressors. A problem with the use of variable selection procedures in traditional regression analysis is that they are likely to overfit on the dataset. This means that the resulting model fits the data better than it fits the underlying population, yielding too optimistic p-values, since the tabulated test statistics do not account for the effects of variable selection (Afifi et al., 2004). Therefore, significance tests for individual regression models are not valid anymore if variable selection has been used. In contrast to this, the test for significantly low generalization error of a learning machine which is employed in MLGPA takes the use of variable selection correctly into account, as will be shown in the following.

In statistical learning theory, the interest lies not in testing the fit of a model to a given dataset, but in the generalization error of learning machines. In machine learning terminology, the independent variables correspond to the *input variables* and the dependent variable to the *output variable* of a learning machine. A learning machine is a procedure which selects a predictor from a family of functions on the basis of a given sample of input-output pairs (training dataset) assumed to be independent and identically distributed (i.i.d.) observations. The goal of learning is to minimize the *generalization error*, which is the expected prediction error on yet unseen examples drawn from the same distribution (test dataset). Results from statistical learning theory (Vapnik, 1998) provide probabilistic bounds on the generalization performance of learning machines on small samples and have led to the development of methods with good generalization ability, like support vector machines (SVMs) (Schölkopf and Smola, 2002). In machine learning, variable selection (Kohavi and John, 1997; Guyon and Elisseeff, 2003) serves two main purposes: First, it improves the generalization ability of a learning machine by removing irrelevant variables. Second, it allows the identification of models which include only the relevant variables which best explain an underlying dependency. Note that applying variable selection to the training set does not induce a bias on the generalization error estimate, since the test set is never used in the variable selection process.

A learning machine aiming to learn the above described generative models for genotype-phenotype relationships needs to provide procedures for variable selection and parameter estimation. For variable selection, MLGPA employs the recently proposed (Hochreiter and Obermayer, 2006) "Potential Support Vector Machine" (P-SVM). In contrast to a conventional support vector machine, which expands the prediction func-

tion into a sparse set of data points (Schölkopf and Smola, 2002), the P-SVM expands it into a sparse set of variables. In previous work, the P-SVM has been shown to be a powerful method to select a compact set of relevant variables (Hochreiter and Obermayer, 2005); it has been successfully applied to problems characterized by large numbers of variables and small sample sizes (Hochreiter and Obermayer, 2004). The mathematical formulation of the P-SVM is described in section 2.1.5. In the P-SVM approach, the number of selected variables is implicitly controlled by a hyperparameter ϵ. In MLGPA, this hyperparameter is selected from a set of candidates using the Bayesian information criterion (Schwarz, 1978), which balances model complexity against model fit. The details of the hyperparameter selection procedure are given in section 2.1.6. For prediction, MLGPA uses ordinary least squares regression, which adjusts the regression parameters for the previously selected variables.

2.1.4 Significance Test

In MLGPA, the estimated generalization error of the learning machine on the given dataset is used as test statistic. An unbiased estimate of the generalization mean squared error is obtained via leave-one-out cross-validation (LOOCV). In this subsampling procedure, each case is used in turn for testing, while the remaining dataset is used for training. Thus for each cross validation fold a separate model will be learned, potentially slightly different with respect to selected variables and coefficients. The LOOCV estimate of the generalization mean-squared error (MSE) is then given by

$$E_{LOOCV}(D) = \frac{1}{n}\sum_{k=1}^{n}(\hat{y}_{D\setminus(x^{(k)},y^{(k)})}(\mathbf{x}^{(k)}) - y^{(k)})^2, \tag{2.4}$$

where $D = \left\{(\mathbf{x}^{(i)}, y^{(i)})\right\}, i \in \{1,\ldots,n\}$ denotes the dataset, and $\hat{y}_{D\setminus(x^{(k)},y^{(k)})}(\mathbf{x}^{(k)})$ is the predicted value on input vector $\mathbf{x}^{(k)}$, where the learning machine has been trained on the reduced dataset $D\setminus(x^{(k)}, y^{(k)})$ in which the k^{th} case has been removed. $E_{LOOCV}(D)$ gives an unbiased estimate on how the learning machine would perform on yet unseen data, and is compared to the error distribution of the H_0-hypothesis of independence using a permutation test. This is done by generating a large number of datasets with permuted output values, and performing on each of them the same analysis as on the original data, including hyperparameter selection, variable selection and regression parameter fitting. The fraction of datasets with lower or equal generalization error compared to the original dataset yields a p-value, which corresponds to the probability that a low or equal generalization error is due to correlations by chance, although input and output are in fact independent (H_0). This p-value is then compared to a significance level α.

MLGPA tests the generalization performance of a learning machine instead of the fit of specific regression models to the data. If a significantly low MSE was found, with high probability the learning machine was able to learn a true underlying dependency

between the set of genetic variables and a phenotype. Then, the learning machine can be used to select a specific model, this time using the whole dataset as training set, which can be expected to generalize well to unseen data from the same distribution. The resulting regression coefficients provide information on the effect direction and strength. In contrast to the conventional method, where severe multiple testing issues are caused by testing for group differences for each genetic configuration, testing the generalization performance of the learning machine on the dataset requires only a single test (per phenotype). However, if more than one phenotype is considered, the false discovery rate (FDR), which is the expected proportion of erroneous rejections among all rejections of H_0, needs to be controlled. In MLGPA this is done using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001), which was shown to be a much more powerful approach to deal with multiple testing issues than procedures controlling the family-wise error rate, e.g. the Bonferroni procedure.

A key feature of MLGPA is the use of variable selection during the training of the learning machine. If a statistical model has too many parameters, overfitting occurs, where the parameters are adjusted to the noise in the data. Then, the model will fit the training data well, but will not generalize to unseen data. Therefore, it is important to use regularization for adjusting the complexity of the model to the right level which allows for good generalization, a principle which is called Occam's razor in the machine learning and statistics literature. In MLGPA, regularization is achieved by terms in the objective function of the P-SVM which enforce the selection of a sparse set of variables, and the adjustment of the model complexity via the Bayesian information criterion. The list of the variables selected during the LOOCV ranked according to selection frequency can be used to judge the robustness of the variable selection by comparing it to the variables selected in the final regression model. If the variables selected most often (e.g. at least half the time) do not coincide with the ones in the final model, it is likely that there are several competing models with almost identical prediction performance, so the existing dependency can be almost equally well modeled using different subsets of variables. For example, if several SNPs are in high linkage disequilibrium (LD), only one of the corresponding allele variables will be selected to provide a sparse solution. Therefore, it is advisable to also look at the SNPs which are in high LD to the ones involved in the final regression model, as they might be the ones actually responsible for an observed effect.

2.1.5 P-SVM Variable Selection for Multiple Linear Regression

Consider we have d genetic variables concatenated into an input vector $\mathbf{x} = (x_1, \ldots, x_d)$. The parametric prediction function of the P-SVM regression is $f(x) = \boldsymbol{x}^T\mathbf{w} + b$, where the parameters are $\mathbf{w}$(d-dimensional column vector) and b. Let X be the $d \times n$ -matrix containing the d-dimensional input vectors for n cases normalized to zero mean and unit variance and y be the n -dimensional vector of the corresponding output variables. The primal optimization problem of the P-SVM is given by (Hochreiter and Obermayer,

2004)

$$\min_{\mathbf{w}} \frac{1}{2}\|X^T\mathbf{w}\|^2 \tag{2.5}$$

$$\text{s.t.} \quad \begin{array}{l} X(X^T\mathbf{w} - y) + \epsilon\mathbf{l} \geq \mathbf{0} \\ X(X^T\mathbf{w} - y) - \epsilon\mathbf{l} \leq \mathbf{0}, \end{array} \tag{2.6}$$

where $\mathbf{l}$ is a $d \times 1$-vector of ones, $\mathbf{0}$ a $d \times 1$ -vector of zeros and ϵ a hyper-parameter of the model. The objective, eq.(2.5), can be derived from bounds on the generalization error using the technique of covering numbers (Hochreiter and Obermayer, 2005), while the constraints, eq.(2.6), ensure that the empirical MSE is minimized. In order to avoid overfitting (adapting the model to noise in the data), a violation of the individual constraints is allowed up to a value of ϵ, in analogy to the concept of ϵ -insensitive loss (Schölkopf and Smola, 2002). The corresponding dual optimization problem is given by

$$\min_{\alpha^+,\alpha^-} \frac{1}{2}(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-)\mathbf{X}\mathbf{X}^T(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-) - \mathbf{y}^T\mathbf{X}^T(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-) + \epsilon^T(\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-) \tag{2.7}$$

$$\text{s.t. } \mathbf{0} \leq \boldsymbol{\alpha}^+, \;\; \mathbf{0} \leq \boldsymbol{\alpha}^-, \tag{2.8}$$

where $\boldsymbol{\alpha}^+$ and $\boldsymbol{\alpha}^-$ are vectors of Lagrange multipliers. This convex optimization problem has a global solution which is found using an efficient sequential minimal optimization (SMO) procedure (Knebel et al., 2008). With $\boldsymbol{\alpha} = \boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-$, the parameters become $\mathbf{w} = \boldsymbol{\alpha}$ and $b = \frac{1}{n}\sum_{i=1}^{n} y_i$. Then the prediction function of the P-SVM is given by

$$f(x) = \sum_{j=1}^{d} \alpha_j x_j + b. \tag{2.9}$$

The first term of the dual objective function in eq.(2.7) corresponds to the empirical covariance matrix $\mathbf{X}\mathbf{X}^T$ of the data and the second term to the correlation $\mathbf{y}^T\mathbf{X}^T$ between input and target. The third term enforces a sparse solution for $\boldsymbol{\alpha}$. All variables x_j for which α_j is non-zero are called the support variables. The objective thus enforces a compact set of support variables, punishes conjoint selection of correlated variables and rewards the selection of variables correlated to the target. P-SVM regression leads to an expansion of the regression function in terms of a sparse set of support variables. Only these variables are needed for the prediction.
The hyperparameter ϵ controls the threshold below which training errors are tolerated. For large values of ϵ, no support variables are chosen, and the prediction function corresponds to the sample mean prediction. As ϵ becomes smaller, more and more variables will become support variables. If ϵ is too small, no training error is tolerated, making the model prone to overfitting. Thus ϵ - regularization corresponds to adjusting the model complexity via the number of support variables as a compromise between high model bias and high model variance.

2.1.6 Hyperparameter Selection

In MLGPA, the hyperparameter ϵ of the P-SVM is selected on each training set of the LOOCV from a set of candidates. For each candidate, the learning machine consisting of P-SVM variable selection and ordinary least-squares regression (OLS) is trained and the Bayesian Information Criterion (BIC) is evaluated, which for regression functions with Gaussian error distribution is given by

$$\text{BIC} \;=\; n\ln(MSE) + k\cdot\ln(n), \tag{2.10}$$

where n is the number of cases, MSE the mean squared error and k the number of regression parameters. The BIC can be used to compare different models (which need not to be nested) by evaluating the fit of a model to the data while penalizing the number of free parameters. A theoretical derivation of the BIC as the Laplace approximation of the likelihood of the model parameters will be given in section 4.2.2.

For each trainings fold i, the candidate ϵ producing the lowest BIC is selected and the corresponding regression model is chosen. Consider we are given a dataset of n cases consisting of input-output pairs, $D = \{(\mathbf{x}^{(i)}, y^{(i)})\},\ i \in \{1,\ldots,n\}$, where the output values y have been standardized to zero mean and unit variance. The outline of the algorithm used for the hyperparameter selection and the calculation of generalization error is given in Algorithm 1.

Algorithm 1 Hyperparameter Selection and Calculation of Generalization Error

BEGIN PROCEDURE
  for $i = \{1,\ldots,n\}$ **do**
    test point: $(\mathbf{x}^{(i)}, y^{(i)})$
    training set: $T_i = D\backslash(\mathbf{x}^{(i)}, y^{(i)})$ of size $n-1$
    for every candidate ϵ **do**
      fit regression model $M(i,\epsilon)$ by
        1. P-SVM variable selection on T_i using ϵ
        2. parameter fit on T_i via OLS regression
      calculate BIC(i,ϵ)
    end for
    select $\epsilon_{opt} = \min_i \text{BIC}(i,\epsilon)$
    use $M(i,\epsilon)$ to predict $(\mathbf{x}^{(i)}, y^{(i)})$
  end for
  calculate $E_{LOOCV}(D)$
END PROCEDURE

2.2 Simulation Study: Sensitivity and Specificity of MLGPA

In order to assess sensitivity and specificity of the MLGPA method, a simulation study was conducted in which MLGPA was applied to synthetic datasets. The analysis was carried out for different numbers of input variables (m=20, 40), different numbers of relevant input variables (k=1, 2) and different sample sizes (n=20, 30, 40).

The specificity (TN /(TN+FP), where TN denotes true negatives and FP denotes false positives) was assessed using Gaussian noise targets. The sensitivity (TP/(TP+FN), where TP denotes true positives and FN denotes false negatives) was assessed using simulated regression targets with additive noise at different noise-to-signal ratios (NSR=0, 0.2, 0.4, 0.6, 0.8, 1), where the NSR is defined as standard deviation of noise divided by standard deviation of signal. For each of these settings, 300 artificial datasets were generated in the following fashion: The input variables were drawn from uniform distribution within the interval (0,1). The noise targets were generated following a standard normal distribution. The simulated regression targets were generated by selecting k ("relevant") variables, multiplying their values with random regression coefficients chosen uniformly from the intervals (-0.6,-0.1) and (0.1,0.6) and adding Gaussian noise at a specified noise-to-signal ratio. The significance level for the tests was set to α=0.05, the number of iterations of the permutation test to 300.

The estimated sensitivities and specificities for the different conditions are listed in Table 2.1. The specificity of the method was always at least 95%, which shows that the Type I error is correctly bounded by the chosen significance level of α=0.05. The sensitivity of the method depended on the noise-to-signal ratio, the sample size, and the total number of input variables. Even for only 20 data points in a 40-dimensional input space, a sensitivity of almost 100% was reached at noise levels up to a NSR of 0.4. The level of noise where a sensitivity of 100% was reached increased with available sample size. This dependency can be explained by the fact that with decreasing noise level the signal will be less and less confounded by the noise, for a fixed sample size. Increasing the sample size increases the robustness with which an underlying dependency can be distinguished from the noise component, leading to an increased sensitivity.

Reducing the number of relevant input variables from k=2 (m=20) to k=1 (m=20) slightly increased sensitivity at a fixed SNR. For small k the contribution of the individual variables to the signal is higher, and less likely to be obscured by noise. The reduction of irrelevant input variables from 38 (m=40, k=2) to 18 (m=20, k=2) also brought an increase in sensitivity. The generalization capability of a learning machine increases when fewer spurious variables are present, because fewer correlations by chance appear in the data. Although MLGPA can deal with datasets containing more variables than examples, it is sensible to keep the number of input variables small in order to increase the sensitivity.

For an evaluation of the variable selection the median numbers of correctly or spuriously selected variables in the final regression model are given in Table 2.2. The analysis shows that usually both relevant variables were found, while with increasing

				Sensitivity					
m	k	n	Specificity	NSR=0	NSR=0.2	NSR=0.4	NSR=0.6	NSR=0.8	NSR=1
20	1	20	96%	100%	100%	100%	97%	75%	50%
		30	96%	100%	100%	100%	100%	94%	76%
		40	95%	100%	100%	100%	100%	99%	93%
20	2	20	95%	100%	100%	100%	92%	65%	45%
		30	95%	100%	100%	100%	99%	89%	75%
		40	97%	100%	100%	100%	100%	98%	90%
40	2	20	96%	100%	100%	99%	76%	45%	27%
		30	96%	100%	100%	100%	97%	76%	59%
		40	97%	100%	100%	100%	100%	97%	79%

Table 2.1: Simulation study: sensitivity and specificity of MLGPA. This table lists the results of the simulation study to assess the sensitivity and specificity of MLGPA for different numbers of total variables (m) and different sample sizes (n) and different numbers of relevant input variables (k=1, 2). For each setting, the analysis was carried out for different noise-to-signal ratios (NSR). For details, see text.

noise-to-signal level also some spurious variables were selected. This effect can be explained by random correlations in the data and is lessened for higher sample size and fewer irrelevant input variables.

2.3 Application to Genomic Imaging Data

Several genes of the dopaminergic and glutamatergic neurotransmitter systems have been found to be associated with alcohol disease and related intermediate phenotypes. Here, genetic variants of the catechol-O-methyltransferase and the metabotropic glutamate receptor 3 gene in alcohol-dependent patients and their association with volumetric measurements of brain structures were evaluated. Conventional statistical methods based on tests for group differences encounter limitations in assessment of epistatic effects. Therefore, the novel MLGPA method is applied to find potential associations between the genetic variants and volumetric measurements obtained from structural MRI. Hippocampal volume was found to be associated with epistatic effects of the COMT-mGluR3 genes in alcohol-dependent patients but not in controls. These data are in line with prior studies supporting a role for dopamine-glutamate-interaction in modulation of alcohol disease.

2.3.1 Medical Background

Alcohol dependence is a widespread disorder, affecting about 10% of the population (Sher et al., 2005). Due to somatic, psychiatric and social consequences, a major

m	k	n	Type of Variables	NSR 0	0.2	0.4	0.6	0.8	1	of a total of
20	1	20	True	1	1	1	1	1	1	1
			Spurious	0	0	0	1	1	2	19
		30	True	1	1	1	1	1	1	1
			Spurious	0	0	0	1	1	1	29
		40	True	1	1	1	1	1	1	1
			Spurious	0	0	0	0.5	1	1	39
20	2	20	True	2	2	2	2	2	2	2
			Spurious	0	0	0	1	1	2	18
		30	True	1	2	2	2	2	2	2
			Spurious	0	0	0	1	1	1	28
		40	True	2	2	2	2	2	2	2
			Spurious	0	0	0	0	1	1	38
40	2	20	True	2	2	2	2	2	2	2
			Spurious	0	0	0	1	2	3	18
		30	True	1	2	2	2	2	2	2
			Spurious	0	0	0	1	2	2	28
		40	True	1	2	2	2	2	2	2
			Spurious	0	0	0	1	2	2	38

Table 2.2: Simulation Study: Evaluating the Variable Selection. This table compares the variables selected in the final regression models of MLGPA to the variables used in generating the targets for the different noise-to-signal ratios (NSRs) analyzed in the simulation study. This analysis was conducted for different numbers of total variables (m), different numbers of relevant variables (k) and different sample sizes (n). For each setting, the median numbers of correctly and spuriously selected variables are listed. The last column provides the maximum possible number of true and spurious variables as reference. For details, see text.

economic impact is related to alcohol dependence. Well designed twin, adoption and family studies have shown that genetic factors play a considerable role for disease risk and symptom characteristics, with 40-60% of the risk variance explained by genetic influences (Goodwin, 1975; Kendler et al., 1994; Bierut et al., 2002; Koehnke, 2008). First degree relatives of alcohol-dependent patients display a 3-4 fold increased risk to develop the disorder, and there is a 55% or higher concordance rate in monozygotic twins compared to a 28% rate for dizygotic twins (Goodwin, 1975). Several linkage studies have been performed that gave evidence for many contributing chromosomal loci, where -among others- glutamatergic and dopaminergic genes are located (Edenberg and Foroud, 2006).

Genetic association studies have shown a relation of alcohol dependence and/or alcohol-related traits with genes of the glutamatergic system, such as N-methyl-D-aspartate receptor channel 1 (NR1), N-methyl-D-aspartate receptor channel 2B (NR2B), glutamate transporter 1 (GLT1), glutamate receptor ionotropic kainate 3 (mGluR7), and clock genes such as Period 2 (PER2) affecting glutamate reuptake (Schumann et al., 2005; Spanagel et al., 2005) as well as genes of the dopaminergic neurotransmitter system, e.g. dopamine receptor D1 (DRD1), dopamine receptor D3 (DRD3), dopamine transporter (DAT) and catechol-O-methyltransferase (COMT) (Enoch et al., 2006; Bowirrat and Oscar-Berman, 2005; Dick and Bierut, 2006). Imaging studies have further supported the role of dopaminergic and glutamatergic alterations in the pathogenesis of alcohol dependence (Heinz et al., 2003; Wong et al., 2003).

However, genetic association studies in complex diseases often give inconsistent results, as it was also shown for the genes mentioned above (Bolos et al., 1990; Schumann et al., 2005; Hines et al., 2005). Reasons for the ambiguity are manifold, including genetic and allelic heterogeneity of the disease, interactive gene-gene effects, contributing external factors and comparatively small sample size (Buckland, 2001). Over the last years, the concept of intermediate phenotypes has been promoted, which tries to dissect complex psychiatric diseases such as alcohol dependence into more basic neurobiological components (Enoch et al., 2003; Hines et al., 2005). Parameters such as electrophysiological measures (Reischies et al., 2005) or structural and functional imaging data (Hariri et al., 2002; Breiter and Gasic, 2004; Heinz et al., 2005; Pezawas et al., 2005) have shown to be reliable tools for genetic association studies in alcoholism and other psychiatric disorders. Several studies have evaluated structural brain alterations associated with alcohol dependence (Hommer, 2003; Pfefferbaum, 2004; Spampanato et al., 2005). Among subcortical and cortical regions found to be different in alcoholics, hippocampal volume reductions have been reported to exceed the general decrease in cerebral atrophy volume found in alcohol-dependent individuals (Agartz et al., 1999; Beresford et al., 2006; Mechtcheriakov et al., 2007). These alterations raise the question whether hippocampal volume may represent a stable or "trait" marker of the disorder that already shows alterations in a presymptomatic stage, or if the hippocampus is more susceptible towards alcohol-toxic influences compared to other brain areas.

In the following study, the proposed MLGPA method was applied to evaluate effects of group II metabotropic glutamate receptor 3 (mGluR3) genetic variants and

their interaction with COMT variations in alcohol-dependent patients and matched controls. COMT was chosen for its potential role in the development of alcohol dependence as indicated by several human studies (Kauhanen et al., 2000; Oroszi and Goldman, 2004). mGluRs play a vital role in synaptic plasticity (Bortolotto et al., 1999), and an essential role for mGluR3 in LTD (long term depression) and a modulatory role for mGluR3 in LTP (long term potentiation) has been found. In addition, it has been suggested that activation of mGluRs modulates excitation and inhibition of dopaminergic mesencephalic neurons (Meltzer et al., 1997; Shen and Johnson, 1997; Campusano et al., 2002; Abarca et al., 2004). Recently, significant statistical epistasis between COMT and mGluR3 has been shown (Nicodemus et al., 2007): several genetic variants of the mGluR3 gene increased the risk of schizophrenia conferred by COMT SNPs, whereas mGluR3 itself did not have an influence on disease risk. Also, epistatic effects of COMT and mGluR3 on working memory function in healthy subjects have been found (Tan et al., 2007). Based on these findings, COMT-mGluR3 interactions were evaluated for their effects on the volume of brain structures measured by magnetic resonance imaging (MRI) in alcohol-dependent patients and controls.

2.3.2 Methods

Patients

This study included 38 patients of Central European descendant (31 male, 7 female, mean age 41 ± 7, range 26-57 years) diagnosed with alcohol dependence according to ICD-10 and DSM-IV. The severity of alcohol dependence was assessed with the Alcohol Dependence Scale (Skinner and Horn, 1984) and the amount of lifetime alcohol intake was measured with the Life Time Drinking History (Skinner and W.J. Sheu, 1982). Patients had no previous substance dependence or current substance abuse other than alcoholism which was confirmed by random breath and urine drug testing. All multi drug abusers were excluded prior to study enrollment. Detoxification was undertaken according to general medical practice using benzodiazepines or chlormethiazole not more than 5 days. Medication was stopped at least one week prior to MRI scan. 41 age matched healthy volunteers of Central European descendant were included as controls (26 male, 15 female; mean age 39 ± 8, range 25-61 years). Standardized clinical assessment with the Structured Clinical Interview I & II (First et al., 1997, 2001) was performed to exclude other axis I psychiatric disorders (and axis II disorders in healthy volunteers). All individuals included into the study were free from any continuous medications and severe somatic disorders, including major neurological and hepatic complications in alcohol-dependent patients. All patients and controls gave fully informed written consent prior to their participation. The study was approved by the local ethics committee and was performed in accordance with the Declaration of Helsinki (1964).

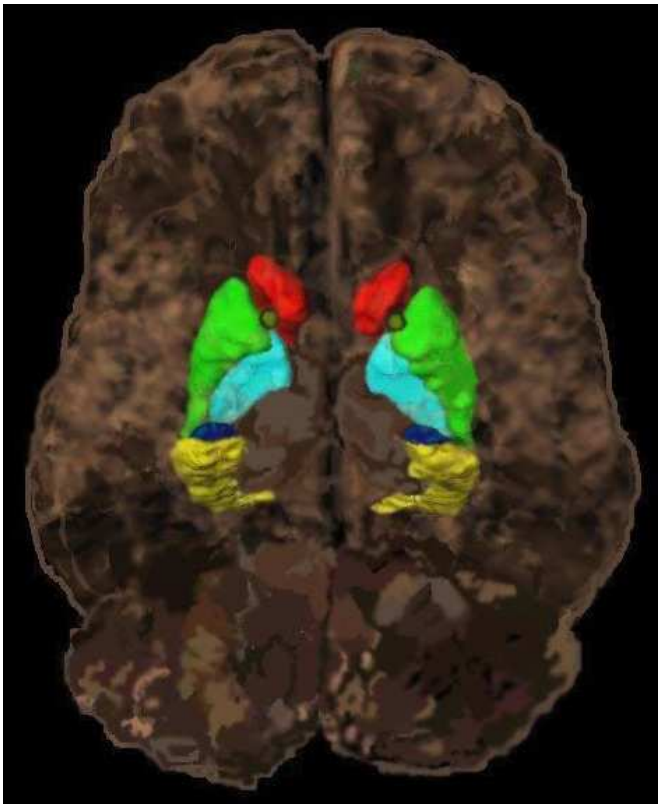


Figure 2.1: Cerebral Regions Included in Segmentation
MRI scan, axial slice, tilted forward, depicting the following regions: red: caudate nucleus; green: putamen; dark green (small area ventral to the putamen): accumbens nucleus; light blue: globus pallidus; dark blue: amygdala; yellow: hippocampus.

Imaging Data

Structural imaging was performed using a 1.5T clinical whole-body MRI (Magnetom VISION; Siemens, Erlangen, Germany) that was equipped with a standard quadrate head coil. A morphological 3D T1-weighted MPRAGE (magnetization prepared rapid gradient echo) image data set (1x1x1 mm voxel size, field of view (FOV) 256 mm, 162 slices, TR = 11.4 ms, TE = 4.4 ms, flip angle $\alpha = 12°$) covering the whole head was acquired. Images were analyzed at the Massachusetts General Hospital as part of the Phenotype Genotype Project in Addiction and Mood Disorder. A brief synopsis of procedures is provided here following those published previously (Makris et al., 2004). Image positions were normalized by imposing a standard three-dimensional coordinate system on each three-dimensional image. Gray-white matter segmentation was performed using a semi-automated intensity contour mapping algorithm for cerebral exterior definition and signal intensity histogram distributions for demarcation of

gray-white matter borders. The cerebral regions investigated in our study are depicted in Figure 2.1. Targeted brain regions were segmented as individual structures using a contour line and manual editing following probabilistically weighted landmarks, cross-referencing of lateral, coronal, and sagittal views, and anatomist supervision. Landmarks and definitions for the targeted brain regions have been well defined elsewhere (Caviness et al., 1996; Makris et al., 2004). Segmentation was performed by two BA level MR technicians blinded to subject diagnosis and group status, and with a randomly ordered sequence of subjects. Intra-rater and inter-rater reliabilities were assessed via percent common voxel assignments (PCVA) (Seidman et al., 2002). Reliabilities were between 85.9 - 90.1. These values are consistent with intra-rater and inter-rater reliabilities reported in previous studies (Goldstein et al., 2002; Seidman et al., 2002; Makris et al., 2004). To rule out gross volumetric effects, total head circumference was calculated for all subjects. Group differences were not significant for this measure ($p > 0.05$).

Genetic Analysis

For genetic analysis, 30 ml of EDTA blood were collected from each individual. Eight single nucleotide polymorphisms (SNPs) distributed over the mGluR3 gene were genotyped (for SNP details see Table 2.3). SNPs were selected for potential functional relevance, location on physical and genetic maps, and minor allele frequency based on the University of California Santa Cruz (UCSC) Human Genome Browser, the National Center of Biotechnology Information (NCBI) database and Applied Biosystems (ABI) SNP Browser software. Two SNPs were removed after genotyping due to limited minor allele frequency (Table 2.3). For assessment of gene-gene-interactions, three SNPs of the COMT gene were chosen (rs2097603, rs4680 [$Val^{158}Met$], rs165599) (Table 2.3) according to prior publications (Meyer-Lindenberg et al., 2006; Nicodemus et al., 2007). Primers were designed for amplification of relevant genomic regions by polymerase chain reaction (PCR); products were cut by allele specific restriction enzymes and visualized after gel electrophoresis. Primer information and specific assay conditions are available on request.

Standard Statistical Methods

Standard statistical analysis was undertaken using Statistical Product and Service Solution (SPSS package 12.0). Group differences in demographic characteristics were assessed by χ^2-test (gender) and Mann-Whitney U-test (age). Influences of gender and age on volumetric hippocampal measures were calculated using Mann-Whitney U-test (gender) and Spearman correlation (age). χ^2-test was used for evaluation of SNP effects on disease status. For effects of each SNP on volumetric data, Mann-Whitney U-test was used for pairwise comparison in a dominant model, assuming a dominant function for one allele. Jonckheere-Terpstra test was chosen for the evaluation of a linear model. For assessment of interactive SNP effects of the two genes, interaction variables were generated and evaluated via Kruskal-Wallis test.

	Chromosomal location	Minor allele frequency	Minor allele frequency/S	Allelic variants	Location
mGluR3					
*rs2073549	80977558	0.008	0.0	A/T	5'UTR
rs1990040	81017056	0.310	0.279	A/G	Intron 1
rs2214653	81049475	0.417	0.353	G/A	Intron 1
rs10238436	81074120	0.267	0.227	C/G	Intron 1
rs6947784	81099550	0.333	0.279	C/G	Intron 2
rs2299219	81121812	0.250	0.198	T/C	Intron 3
rs10266758	81146463	0.258	0.272	G/T	Intron 3
*rs17161026	81172229	0.011	0.013	C/T	Exon 4
COMT					
rs2097603	3779279	0.390	0.448	A/G	Promoter
rs4680	3802473	0.483	0.500	A/G	Exon, AA exchange
rs165599	3807982	0.461	0.317	A/G	3' near gene

Table 2.3: SNPs genotyped within the mGluR3 and COMT gene. Characteristics of SNPs genotyped: Chromosomal location and minor allele frequencies given for Caucasian population is based on the NCBI database. Minor allele frequencies/S refers to allele frequencies in our study. They were not found to be different from the published frequencies. Allelic variants show major/minor allele. Location depicts position within the genome. AA exchange amino acid exchange, A codes for methionine, G codes for valine.
* Due to low minor allele frequencies, these mGluR3 SNPs were excluded from further analysis.

MLGPA

Six of the measured mGluR3 SNPs were selected that showed acceptable allele distribution in our dataset, and three COMT SNPs (rs2097603, rs4680 [Val158Met], rs165599) were incorporated in the analysis Table 2.3). All 18 alleles of the 9 SNPs and all pair-wise multiplicative interactions which involved one allele from the COMT gene and one allele from the mGluR3 gene were included, yielding a total of 90 predictive variables. Since the dataset contained missing values, probabilistic inputs were estimated according to the procedure described above. The candidates for the value for the hyperparameter ϵ were $\{1.0, 1.05, \ldots, 3\}$. The false discovery rate was controlled at $q = 0.05$.

2.3.3 Results

Analysis of demographic data by standard statistical methods did not reveal any significant differences for age (p = 0.403) or gender (p = 0.072) between patients and controls. Evaluation of the effects of disease status on volumetric measurements suggested significant differences for the hippocampus (right: p = 0.006; left: p = 0.044; total: p = 0.014) (Z = -2.767 / -2.012 / -2.460), nucleus pallidus (p = 0.001 / 0.003 / 0.001) (Z = -3.417 / -2.945 / -3.209) and nucleus accumbens (p = 0.006 / 0.021 / 0.008) (Z = -2.764 / -2.312 / -2.652) with smaller volumes in patients compared to controls. These differences could not be accredited to global volume diminishment in the patient group because all images had been normalized for whole brain volume. For further analysis, only these three brain regions were included.

Gender effects on volumetric data were shown in the control group for hippocampus (p = 0.004 / 0.025 / 0.005) (Z = 2.896 / 2.247 / 2.788) and nucleus accumbens (p = 0.002 / 0.011 / 0.002) (Z = -3.059 / -2.544 / -3.140) with smaller volumes for female subjects, but not in the patient group (p>0.05). No gender effects were found in either group for volume of nucleus pallidus (p>0.05). Age did not have any effect on volumetric data in the controls or in the patient group (p>0.05). The life time alcohol amount measured in total amount, amount over the last 5 years and over the last year did not have any effects on hippocampal volume (p > 0.05). Genotypes of all SNPs were in Hardy-Weinberg equilibrium. Allele frequencies of all SNPs were in accordance to published data (NCBI SNP database), (Table 2.3). No association was found for any of these SNPs with the diagnosis of alcohol dependence (p>0.05).

Standard Statistical Methods

Due to low minor allele frequencies, two mGluR3-SNPs were excluded from further statistical analysis (see Table 2.3). Each SNP solely as well as all possible SNP-SNP interactions between mGluR3 and COMT were tested for group differences in volumetric data of hippocampus, nucleus pallidus and nucleus accumbens (right/left/total). Due to the number of SNPs that had to be considered (9), the different models of allelic interaction (dominant and linear) that had to be calculated separately (2), and the

Phenotype	$E_{LOOCV}(D)$	p-value
Right Hippocampus	0.6656240	0.0023*
Left Hippocampus	0.715100	0.0057*
Bilateral Hippocampus	0.555630	0.0004*

Table 2.4: MLGPA: Generalization error estimates and p-values. The estimated generalization mean squared error $E_{LOOCV}(D)$, eq.(2.4), is given for phenotypes where a significant dependency was found on the patient sample. The table also lists the corresponding p-values calculated using 10000 random permutations of the output values. Results significant controlling the FDR at level 0.05 are marked by *.

phenotypic variables that had to be taken into account (9), a large number of tests was necessary (162 tests). Additionally, interactive effects on phenotypic variables further increased the number of tests (324). After Benjamini-Hochberg procedure to correct for multiple testing, all results dropped below significance threshold.

MLGPA

Significant dependencies (controlling the FDR at q=0.05) were found for left, right and bilateral hippocampus volume in the patient group. The generalization errors and p-values for these phenotypes are listed in Table 2.4, the variables selected during the LOOCV in Table 2.5 and the regression models learned on the whole dataset in Table 2.6. For the controls, no significant dependencies were found. For all three phenotypes, the three variables selected more than half of the time in the LOOCV were the same ones used in the regression models learned on the whole dataset, indicating the robustness of the variable selection. Among alcohol-dependent patients, two multiplicative interaction variables between an allele of COMT and an allele of mGluR3, rs6947784(C)•rs165599(G) and rs2214653(G)•rs165599(G), were associated with an increase in volume for the left, right and bilateral hippocampus, while a decrease in volume was associated with rs10266758(G)•rs165599(A) for the left and bilateral hippocampus and with rs10266758(G) for the right hippocampus.

The fact that all three predictive variables included in the final regression model were multiplicative interaction variables between COMT and mGluR3 alleles, although additive interactions between single alleles of both genes were also allowed points towards a multiplicative interaction between these systems. In order to check whether a significant generalization performance could also be reached using purely additive interactions of single alleles, a post-hoc analysis was carried out in which only the 18 single alleles from both genes and no multiplicative interaction variables were included. Due to the used multiple regression ansatz, arbitrary additive interactions between single alleles were allowed. MLGPA was applied in exactly the same way as before. However, in this purely additive setting no significant effects (controlling the FDR at q=0.05) were found, neither for patients nor controls.

Phenotype	Variable	S	$\bar{\alpha}$
Right Hippocampus	rs2214653(G)•rs165599(G)	38	0.712656
	rs6947784(C)•rs165599(G)	38	1.153186
	rs10266758(G)	36	-1.008475
	rs10238436(C)•rs2097603(A)	17	0.442371
	rs10266758(G)•rs165599(A)	2	-0.541244
	rs2214653(A)•rs2097603(A)	2	0..428792
	rs1990040(G)•rs2097603(A)	1	0.364769
Left Hippocampus	rs6947784(C)•rs165599(G)	38	0.911440
	rs10266758(G)•rs165599(A)	37	-0.767511
	rs2214653(G)•rs165599(G)	37	0.577732
	rs2097603(T)	4	0.511084
	rs2214653(A)•rs4680(G)	2	0.503293
	rs2214653(G)•rs2097603(T)	1	0.337316
Bilateral Hippocampus	rs2214653(G)•rs165599(G)	38	0.603048
	rs6947784(C)•rs165599(G)	38	1.090808
	rs10266758(G)•rs165599(A)	37	-0.636309
	rs1990040(G)•rs2097603(A)	1	0.449874

Table 2.5: MLGPA: Variables selected in the LOOCV for the three phenotypes for which a significant dependency was found by MLGPA on the patient sample. The • in the variable names denotes a multiplicative interaction term. Each allele is described by the name of the SNP and the nucleotide (in brackets). S denotes the number of times a variable was selected in the outer leave-one-out cross validation loop, which indicates the robustness of the selection. $\bar{\alpha}$ denotes the average regression coefficient.

Phenotype	Regression Model
Right Hippocampus	$Y = 1.0923 \times \text{rs6947784(C)} \bullet \text{rs165599(G)}$ $+0.75052 \times \text{rs2214653(G)} \bullet \text{rs165599(G)}$ $-1.044 \times \text{rs10266758(G)}$ $+0.31685$
Left Hippocampus	$Y = 0.90765 \times \text{rs6947784(C)} \bullet \text{rs165599(G)}$ $+0.57617 \times \text{rs2214653(G)} \bullet \text{rs165599(G)}$ $-0.76875 \times \text{rs10266758(G)} \bullet \text{rs165599(A)}$ $+0.097414$
Bilateral Hippocampus	$Y = 1.0873 \times \text{rs6947784(C)} \bullet \text{rs165599(G)}$ $+0.6026 \times \text{rs2214653(G))} \bullet \text{rs165599(G)}$ $-0.63139 \times \text{rs10266758(G))} \bullet \text{rs165599(A)}$ -0.063448

Table 2.6: MLGPA: Regression models. This table shows the regression models learned by the learning machine on the whole patient sample in a post-hoc analysis in case a significant dependency was found. The • in the variable names denotes a multiplicative interaction term between two alleles. Each allele is described by the name of the SNP and the nucleotide (in brackets).

2.3.4 Discussion

This study provides evidence that the combined set of genetic variants of the mGluR3 and COMT gene is associated with hippocampal volume in alcohol-dependent patients but not in controls. Our study gave evidence that epistatic effects, describing common gene-gene interactions, between these two systems is likely to be involved. This result is particularly interesting under the light of a recent study from (Nicodemus et al., 2007), who showed an epistatic effect of COMT and mGluR3 concerning the risk for schizophrenia. Also, (Tan et al., 2007) found epistatic effects of COMT and mGluR3 on working memory function in healthy subjects. In our study, the selected variables indicate an additive superposition of multiplicative pair-wise interactions between the three mGluR3-SNPs rs6947784, rs2214653, rs10266758 and the COMT-SNP rs165599 that influence hippocampal volume in the present sample. While additive interaction refers to added effects of certain alleles of single SNPs (each independent of the rest), multiplicative interaction refers to an effect of a certain allele combination involving multiple SNPs. However, for the setting where only the single alleles and not the multiplicative interactions terms were included, no significant effect was found. Therefore, based on our sample, the epistatic effects of mGluR3 and COMT genotypes seem to manifest mainly in a non-additive, multiplicative manner, causing a decreased hippocampal volume in the absence or presence of specific allele combinations involving both COMT and mGluR3 SNPs. At present, information about the functionality of

the relevant mGluR3 SNPs does not exist. Therefore, conclusions concerning glutamate levels cannot be drawn from our study. For rs165599, located in the 3'UTR of the COMT gene, a reduction of COMT expression with the G allele was shown that consecutively leads to increased dopamine levels (Dempster et al., 2006). However, other authors did not confirm the functional relevance of rs165599 (Chen et al., 2004). In our study, the G allele has been associated with an enlarged hippocampal volume. Any conclusions drawn from these results concerning COMT expression and subsequent dopamine levels would be rather speculative.

Both glutamate and dopamine neurotransmitter systems have often been implicated in the pathogenesis of alcohol dependence (Tsai et al., 1998; Heinz et al., 2003; Koob, 2003; Schumann et al., 2005). Glutamatergic neurotransmission, mediated by metabotropic glutamate receptors modulates cue-induced drug-seeking behaviour (Markou, 2007). Metabotropic glutamate receptors are G-protein linked glutamate sensitive receptors that control the activity of membrane enzymes and ion channels (Conn and Pin, 1997). mGluR3s belong to group II mGluRs that are predominantely found presynaptically where they reduce transmitter release as autoreceptors (Shigemoto et al., 1997). Group II metabotropic glutamate receptor agonists that bind to metabotropic glutamate 2/3-receptors decrease alcohol self-administration and cocaine-seeking in rats (Backstrom and Hyytia, 2005; Peters and Kalivas, 2006). Other metabotropic glutamate receptors are also involved in withdrawal symptoms and relapse risk (Dravolina et al., 2007; Kotlinska and Bochenski, 2007).

Coordinated signaling of the dopaminergic and glutamatergic systems at the cellular level is also critical for long-term plasticity and reward-related learning in corticostriatal networks. Activity-dependent changes in synaptic strength are in most cases mediated by glutamate receptors (Bortolotto et al., 1999) and potentiated by activation of dopamine receptors (Jay, 2003). Cells that receive dopaminergic and glutamatergic input act as coincidence detectors in associative learning, and thereby alter the activity of neural ensembles (Kelley, 2004). Activation of mGluRs modulates excitation and inhibition of dopaminergic mesencephalic neurons (Meltzer et al., 1997; Shen and Johnson, 1997; Campusano et al., 2002; Abarca et al., 2004). A recent study suggests an essential role for mGluR3 in LTD, and a modulatory role for mGluR3 in LTP, with effects being mediated by distinct pre- and postsynaptic loci (Altinbilek and Manahan-Vaughan, 2007). It is assumed that mGluR3 agonists have neuroprotective properties mediated by astroglial group II receptors that induce glutamate uptake (Yao et al., 2005; Corti et al., 2007). mGluR3 activation in astrocytes leads to an increased formation and release of TGF , which in turn protects neighbouring neurons against excitotoxic cell death (Bruno et al., 1998). Also, activation of group II mGluRs prevents neurotoxic effects triggered by NMDA in rat hippocampus (Pizzi et al., 1996).

Studies in animals and humans suggest a close functional relationship between memory formation in the hippocampus and dopaminergic neuromodulation (Lisman and Grace, 2005). In aged animals, decreasing dopamine neurotransmission impairs learning (Hersi et al., 1995). In humans, enhancement of dopamine pools with L-DOPA produces a significant improvement in memory (Knecht et al., 2004) and activation of

dopaminergic midbrain structures by reward is associated with enhanced hippocampus-dependent formation of explicit memory (Wittmann et al., 2005). The hippocampus regulates this release of dopamine by disinhibition of the ventral tegmental area (Legault and Wise, 2001) emphasizing its role in reward-related learning. Dopaminergic activity has also been found to influence neuroplastic processes via neurotrophic factors (Ji et al., 2005).

In alcohol-dependent patients, reduced hippocampal volumes exceeding general brain atrophy have been observed (Sullivan et al., 1995; Agartz et al., 1999; Beresford et al., 2006; Mechtcheriakov et al., 2007). This finding may result from increased vulnerability of the hippocampus to the neurotoxic effects of alcohol compared to other brain regions. However, the decrease may at least in part also precede alcohol dependence. Reductions of hippocampal volume have been found in adolescent heavy drinkers (De Bellis et al., 2000; Medina et al., 2007). Adolescents with short-term alcohol use disorder also showed decreased left hippocampal volume, suggesting alterations of hippocampal volume in premorbid stages (Nagel et al., 2005). Taken these data together, an impact of functional genetic variants in dopaminergic and glutamatergic genes may lead to a reduction of hippocampal volume by alterations of different neuroprotective pathways.

An association of COMT/mGluR3 genetic variants with hippocampal volume was not observed in the control group. This finding supports the hypothesis that a gene-gene interaction specific for alcohol-dependent patients is related to hippocampal atrophy. It has been shown in other studies that certain gene effects might only become relevant with distinct exogenous conditions present (Heinz et al., 2000). Chronic alcohol exposure might function as stressor to the brain which leaves certain intracerebral structures such as the hippocampus more susceptible to genetic effects. However, the possibility that the sample size in this study was not large enough to detect similar alterations in the control group, or to dissociate subtypes within the control population that show a similar epistatic effect cannot be excluded. Continued genetic and volumetric studies in adolescents with and without a family history of alcoholism may help to further elucidate the interactions. Also, in our study a gender effect was seen in the control group with female subjects having significantly reduced volumes of the hippocampus compared to male controls. The lack of a gender effect in the patient group may be due to toxic effects of alcohol that interfere with gender effects. Also, alcohol amount in the male population was significantly higher compared with the female group, which might also cause a pronounced reduction of hippocampal volume and thereby intervene with naturally occurring size differences.

Subsequent studies are needed to follow-up questions raised by these findings. First, a replication sample will be collected to examine the reproducibility of this initial analysis. Second, further molecular studies need to be performed to investigate the specific functional consequences the identified genetic variants might have. Third, it should be examined whether individuals with the particular genetic risk constellation for reduced hippocampal volume demonstrate specific clinical features such as alterations in the development of tolerance, withdrawal symptoms, relapse rates, and impaired memory

function since these specific phenomena are modified by dopaminergic and glutamatergic neurotransmission in alcohol-dependent patients (Tsai et al., 1998; Heinz et al., 2003).

2.4 Summary and Conclusions

In this chapter, population association studies with quantitative variables were considered under the assumption that epistatic effects play a role and that the available sample size is small. For this setting, a method for genotype-phenotype analysis based on machine learning techniques (MLGPA) was proposed, which overcomes the limitations of conventional statistical methods. MLGPA does not require prior assumptions about the number of relevant predictive variables, allows for multiplicative as well as additive epistatic effects and uses both models (dominant and linear) of allele interaction. Missing SNP values are dealt with in a probabilistic fashion. Since only a single test needs to be conducted for each phenotype, the proposed method reduces multiple testing issues and can deal with high-dimensional datasets of limited sample size.

MLGPA is conceptually different from approaches based on tests for group differences, where for each group division specified by one of the numerous genetic configurations a separate test needs to be conducted, unless there is a single a priori hypothesis. In MLGPA, the significance of the estimated generalization performance of a learning machine on the given dataset is tested. If the result of the test is not significant, one cannot conclude that there is no genotype-phenotype association, only that the learning machine was not able to find any on the dataset. This is analogous to statistical tests for group differences, where failure to reject the H_0-hypothesis does not imply that H_0 is true. If the result of the test in MLGPA is significant, one can trust the learning machine to learn a model on the given dataset which carries over well to yet unseen examples from the same distribution as the dataset. In this context, it is important that the method uses a cross-validation approach, in which variable selection, hyperparameter selection and parameter fitting make exclusive use of the training set of the respective cross-validation fold, and that the same procedure is also applied to the datasets sampled from the H_0-hypothesis during the label permutation test. Therefore, the significance test employed in MLGPA correctly takes the effect of variable selection into account.

Since the MLGPA method conducts a test for a significantly low generalization error of the learning machine on the given dataset, only a single statistical test is conducted per phenotype. Therefore, multiple testing corrections are not required if a single phenotype is analyzed. Only if several phenotypes are tested, the multiple testing issue arises, because for each phenotype a separate test is conducted. In MLGPA, this problem is handled by controlling the false discovery rate (FDR). The permutation test guarantees that the specificity, and therefore the type I error rate, is independent from the noise level in the data and controlled by the chosen significance level. The sensitivity, however, decreases with increasing noise-to-signal ratio, as the simulation study shows. For higher noise level, a larger sample size is required to reach the same

sensitivity.

The MLGPA method can be modified by employing other algorithms instead of the P-SVM for variable selection on each training set. However, most common variable selection procedures might get stuck in a local optimum, e.g. stepwise regression procedures based on a series of statistical tests, or are computationally too demanding during the conduction of the label permutation test, e.g. an extensive all-subsets search. The P-SVM was chosen because a computationally efficient algorithm exists to find the global optimum of its objective function and because it was found to provide good generalization performance on problems with many variables and limited sample size. This is important, since the sensitivity of MLGPA crucially depends on the generalization capability of the learning machine: Poor learning machines using bad variable selection algorithms will have a low sensitivity and will rarely find significant dependencies in the data, but even then the type I error rate will be controlled.

3

Target Selection to Model a Single Genetic Effect on a Multidimensional Phenotype

3.1 Introduction

In genetic association studies researchers test for significant dependencies between some candidate genetic variants and a set of phenotypic variables. In chapter 2 this problem was approached using feature selection[1]. The binary genetic variables were considered as input and each of the real-valued phenotypes in turn as output of a learning machine. The learning machine conducted ordinary least squares regression on a sparse subset of selected input variables obtained by the P-SVM (Hochreiter and Obermayer, 2006), a recently proposed feature selection approach. Thus for each phenotype a sparse regression model was obtained, whose complexity was adjusted using an information criterion. The generalization performance of the learned model was then assessed using a label permutation test, where the whole learning procedure, including feature selection and regression, was conducted on a large number of datasets with permuted output values. The leave-one-out cross-validation (LOOCV) estimate of the generalization mean squared error was used as test statistic, and a p-value was calculated by assessing the percentage of generalization error estimates on the permuted datasets lower or equal to the one obtained on the original dataset. If this p-value was lower than a chosen significance level, the learned regression model was significant. This means that it is unlikely that such a good generalization performance could have been obtained (due to random correlations in the data) under the H_0 hypothesis of statistical independence

[1]The expression *feature selection* is equivalent to *variable selection*. In this chapter, the expression feature selection is deliberately used to distinguish input variable selection from the novel concept of target variable selection.

between input variables and output variable.

However, even though the learned models are very sparse, the prediction function is a linear combination of a small number of genetic variables. While this could model an actual additive effect present in the population, it could also be that some of the chosen regressors only fit noise in the given sample, and do not represent population effects. While there is no way to tell for sure without conducting replication studies on different samples, the possibility of overfitting has to be taken into account, because genetic effects on phenotypes are usually small, and it is likely that deviations of the residuals from the normal distribution occur or that outliers are present in the data. In order to reduce such potential overfitting effects, it might be prudent to restrict the model for small sample sizes to the single genetic variable which is most likely to influence the phenotypes. Furthermore, the method of chapter 2 considers each phenotype separately. But often it is expected that a subgroup of the measured phenotypic variables is subject to the same genetic influence. Treating this group of variables as a multidimensional phenotype is likely to reduce noise, and to improve the robustness of prediction. Moreover, in scenarios like multi-voxel pattern analysis of fMRI data (Norman et al., 2006), the phenotypic vector is very high-dimensional, and the association to the genotype might manifest in a complex pattern. In the following, it will be shown that the above issues can be resolved with the help of a new learning paradigm called target selection.

This chapter is structured as follows: First, the proposed learning paradigm of target selection is introduced. This is followed by the derivation of an objective function for target selection based on the concept of mutual information. Then, a learning machine for target selection is suggested, and the question of model evaluation and significance testing is discussed. Finally, the proposed method is used for re-analyzing the data from the genetic imaging study in chapter 2 under different model assumptions. The goal of this analysis is to test for potential associations between a multidimensional phenotype and one of the genetic variables.

3.2 Target Selection

3.2.1 A New Learning Paradigm

In the target selection paradigm, the following generative model of the data is assumed: There are several discrete random variables $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_m\}$, which take values from different sets of labels $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_m$. These variables are distributed according to $P(\mathbf{Y})$ and have the marginal distributions $P(Y_i),\ i = 1, \ldots, m$. One of these variables, Y_g, is assumed to give rise to the distribution of another set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\} \in \mathcal{X}$ via a set of M_g conditional distribution functions

$$P(\mathbf{X}|Y_g = c_j),\ c_j \in \mathcal{Y}_g, \tag{3.1}$$

where M_g is the number of different labels in $\mathcal{Y}_g$, and where the functions $P(\mathbf{X}|Y_g = c_j)$, $j = 1, \ldots, M_g$ are assumed to be smooth functions of arbitrary shape. Using the Bayes

theorem, the optimal prediction model for Y_g given $\mathbf{X}$ is obtained as

$$P(Y_g = c_k|\mathbf{X}) = \frac{P(\mathbf{X}|Y_g = c_k)P(Y_g = c_k)}{\sum_{j=1}^{M_g} P(\mathbf{X}|Y_g = c_j)P(Y_g = c_j)}. \tag{3.2}$$

In the following we refer to the distribution $P(Y_g|\mathbf{X})$ defined in eq.(3.2) also as the 'true' distribution, since it is the optimal distribution for predicting Y_g under the assumed generative model.

Viewing this setting as a classification problem, $\mathbf{X} = \{X_1, X_2, \ldots, X_d\} \in \mathcal{X}$ can be considered as multidimensional input of a learning machine, while $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_m\}$ forms a set set of target (or output) variables. Under the assumed generative model, all targets $Y_{i\neq g}$ are statistically independent of the inputs, given Y_g:

$$Y_{i\neq g} \perp \mathbf{X} \mid Y_g. \tag{3.3}$$

Let us assume that we are given a dataset consisting of N input-target pairs $D = (\mathbf{x}^{(i)}, \mathbf{y}^{(i)}), i = 1, \ldots, N$, which are assumed to be sampled from the above generative model. The goal of target selection is to recover the optimal prediction model, (3.2), as good as possible given only the dataset D. This involves (i) the selection of a target variable Y_t from $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_m\}$, and (ii) the choice of a probabilistic classification function which predicts the class probabilities of the selected target variable given the inputs.

This task is to be solved by a learning machine, denoted by LM_1, which needs to select the target, as well as the parameters and hyperparameters of the classifier. The learning machine LM_1 will incorporate a second learning machine LM_2, which is responsible for fitting the model parameters θ of a probabilistic classification model specified by a parameterized function $\hat{p}_\theta(Y_t|\mathbf{X})$. We assume that the learning machine LM_2 has a set of hyperparameters ψ. For a selected target variable and given hyperparameters ψ, LM_2 adjusts the parameters θ on the dataset D. An example for such a learning machine LM_2 is a probabilistic neural network with d input neurons and one sigmoid output neuron, a quadratic loss function, and the conjugate gradient algorithm for learning the weight parameters of the network, where the hyperparameters are the number of hidden units. Another example, which will be used in the experiments, is a probabilistic support vector machine (Chang and Lin, 2001) with hyperparameter C.

3.2.2 Objective Function for Target Selection

In the following, we propose an objective function for the target selection paradigm which is based on the idea to chose that model from the given model class in which the input variables provide most information about the target. The informativeness of the input variables for a specific target can be assessed by the concept of mutual information (Cover and Thomas, 1991). The mutual information between two discrete variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass

functions $p(x)$ and $p(y)$ is defined as the Kullback-Leibler divergence between the joint and the product of the marginals,

$$\begin{aligned} I(X;Y) &= D_{KL}(p(x,y)||p(x)p(y)) && (3.4)\\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} && (3.5)\\ &= E_{p(x,y)}\log\frac{p(X,Y)}{p(X)p(Y)}. && (3.6)\end{aligned}$$

The mutual information measures the mutual dependence of the two variables, and can also be expressed via the marginal entropy $H(Y)$ and the conditional entropy $H(Y|X)$,

$$I(X;Y) = H(Y) - H(Y|X). \quad (3.7)$$

Since the entropies measure the amount of uncertainty about a variable, mutual information corresponds to the reduction in uncertainty that the knowledge of one variable provides about the other.

Note that mutual information can be analogously defined for continuous variables (Cover and Thomas, 1991), if the probability mass functions are replaced by densities, the sums are replaced by integrals over dx and dy and the entropies are replaced by differential entropies. In the following, however, for simplicity the discrete notation will be used to cover both cases.

Here, the concept of mutual information is extended to the multivariate case of a set of variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ and a single variable Y:

$$\begin{aligned} I(\mathbf{X};Y) &= H(Y) - H(Y|X_1, X_2, \ldots, X_d) && (3.8)\\ &= D_{KL}(p(\mathbf{x},y)||p(\mathbf{x})p(y)) && (3.9)\\ &= \sum_{\mathbf{x}\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(\mathbf{x},y)\log\frac{p(\mathbf{x},y)}{p(\mathbf{x})p(y)} && (3.10)\\ &= E_{p(\mathbf{x},y)}\log\frac{p(\mathbf{X},Y)}{p(\mathbf{X})p(Y)}. && (3.11)\end{aligned}$$

Since $p(Y,\mathbf{X}) = p(Y|\mathbf{X})p(\mathbf{X})$, eq.(3.11) can be also written as

$$I(\mathbf{X};Y) = E_{p(\mathbf{x},y)}\log\frac{p(Y|\mathbf{X})}{p(Y)}. \quad (3.12)$$

The extended mutual information reflects the reduction in uncertainty about Y due to knowledge of the set of variables $\mathbf{X}$. Applied to the above classification setting, Y corresponds to a target variable $Y_t, t = 1, \ldots, m$, and $\mathbf{X}$ to the input variables. Eq.(3.12) specifies the information gain the input variables provide about target Y_t. We do not know the distribution functions $p(Y_t|\mathbf{X})$ and $p(Y_t)$, however, we can estimate distribution models on the training data D.

The model for the prior probabilities $\hat{p}_\phi(Y_t)$, parameterized by ϕ, is obtained by counting the relative frequencies with which the labels occur in the training data. To

model the conditional probability distribution of the target variable given the inputs, the learning machine LM_2, which was introduced in section 3.2.1, is used. For given hyperparameters ψ, and target variable index t, LM_2 adjusts the parameters θ of a parametric probabilistic model $\hat{p}_\theta(Y_t|\mathbf{X})$ on the dataset D. The resulting model, $\hat{p}_\theta(Y_t|\mathbf{X}; D)$, is just a function of $\mathbf{X}$.

Given a dataset D, the goal of target selection is to let LM_1 chose the target Y_t and the hyperparameters ψ such that the objective

$$G(\psi, t|D) = E_{p(\mathbf{x},y)} \log \frac{\hat{p}_\theta(Y_t|\mathbf{X}; D)}{\hat{p}_\phi(Y_t; D)} \tag{3.13}$$

is maximized. We call this function *expected log-likelihood gain*. The expected log-likelihood gain is a function of the hyperparameters and the target variable index and depends on the used model class as well as on the given dataset. Unlike the mutual information, the expected log-likelihood gain is not non-negative. Eq.(3.13) corresponds to the expected increase in log-likelihood the probabilistic classification model trained on dataset D achieves on data sampled from the true underlying distribution, compared to the estimated prior model.

In practice, this objective is not directly accessible, since the true distribution $p(\mathbf{X}, Y)$ is not known. However, if the mathematical expectation in eq.(3.13) is replaced by the empirical average over a test dataset T of size N_T assumed to be drawn from the distribution $p(\mathbf{x}, y)$, a calculable objective is obtained,

$$G_{emp}(\psi, t|T, D) = \frac{1}{N_T} \sum_{i=1}^{N_T} \log \frac{\hat{p}_\theta(y_t^{(i)}|\mathbf{x}^{(i)}; D)}{\hat{p}_\phi(y_t^{(i)}; D)}. \tag{3.14}$$

We call the function G_{emp} the *empirical log-likelihood gain*. According to the law of large numbers, the empirical log-likelihood gain, eq.(3.14), will converge almost surely to the expected log-likelihood gain, eq.(3.13),

$$G_{emp}(\psi, t|T, D) \overset{\text{a.s.}}{\to} G(\psi, t|D) \text{ for } N_T \to \infty, \tag{3.15}$$

i.e.

$$P\left(\lim_{N_T \to \infty} G_{emp}(\psi, t|T, D) = G(\psi, t|D)\right) = 1. \tag{3.16}$$

It is important that the target selection paradigm is fundamentally different from the feature selection paradigm (Kohavi and John, 1997; Guyon and Elisseeff, 2003). This is illustrated in Figure 3.1. In feature selection, the goal is to select a subset of the input variables which is relevant for predicting a given target, and the performance criterion is the prediction performance on that target. Instead, the aim of target selection is to find the target variable with the most informative input-target relationship, and to fit a parametric function to build a probabilistic model for the chosen target variable conditional on the inputs. Unlike in feature selection, the performance criterion for target selection is not the prediction performance on the selected target Y_k,

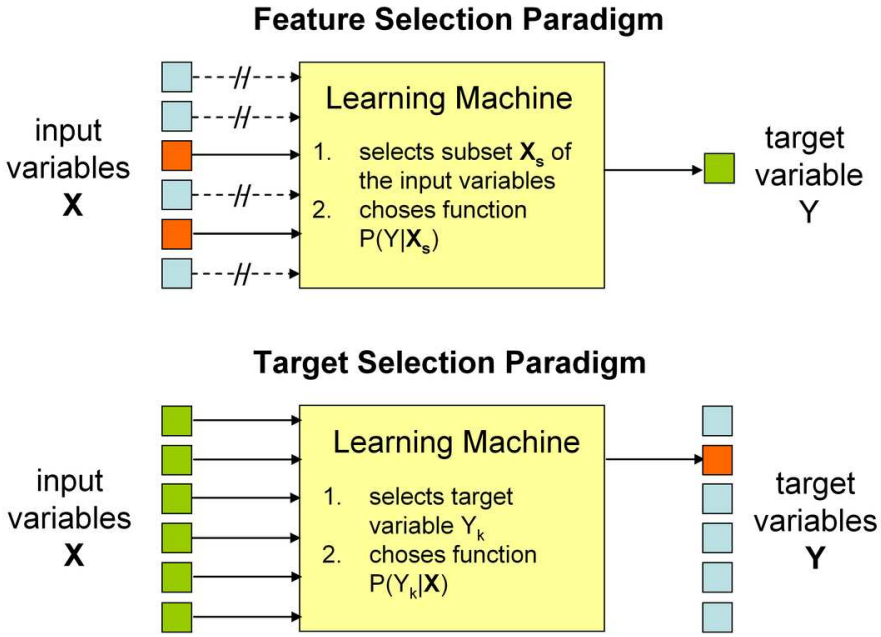


Figure 3.1: Comparison of the feature selection paradigm to the target selection paradigm. Red boxes denote variables selected by the learning machine, blue boxes variables which are not selected and green boxes variables which are always selected.

but the additional information about the target which is provided by the input variables $\mathbf{X}$. While the motivation of feature selection is either to improve the prediction performance of a classifier, or to get an improved explanatory model by retaining only the input variables most relevant for predicting the targets, target selection shares only the second of these motives. Its aim is to find the probably most informative relation between the multidimensional input and one of the target variables, in order to recover the true model as good as possible.

Note, that if an alternative generative model is considered, where only a subset of the variables $\mathbf{X}_g \subset \mathbf{X}$ is influenced by the class variable Y_g, the target selection paradigm can also be combined with feature selection. This is illustrated in Figure 3.2. Then the task of the learning machine would be to select both a target variable Y_k and a subset $\mathbf{X}_s$ of the input variables, and to learn the corresponding probabilistic classification function.

3.2.3 Learning Method for Target Selection

In target selection, the learning machine LM_1 is required to perform two tasks:

1. The target variable Y_t has to be chosen.
2. A probabilistic classification function for the desired target has to be learned.

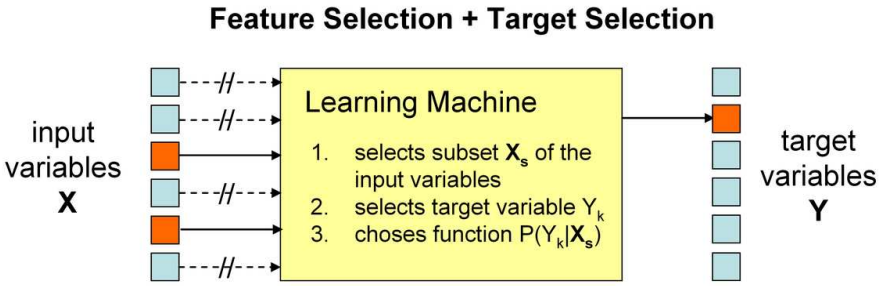


Figure 3.2: Target selection combined with feature selection. Red boxes denote variables selected by the learning machine, blue boxes variables which are not selected.

In addition to the choice of target variable Y_t, also the hyperparameters ψ have to be adjusted. If a separate validation set is available, the optimal values ψ_{opt} and t_{opt} can be found by evaluating the empirical log-likelihood gain, eq.(3.14), which the models trained on the training sample at given values of t and ψ achieve on the validation set. This can be done by using a grid search involving all targets and several candidate values for ψ.

However, if there is not enough data available to split the data into separate training and validation sets of suitable size, leave-one-out cross-validation (LOOCV) has to be used instead. Then, ψ_{opt} and t_{opt} are determined on the given training sample $D = (\mathbf{x}^{(i)}, \mathbf{y}^{(i)}), i = 1, \ldots, N$, via a grid search procedure in which the LOOCV value of the empirical log-likelihood gain, eq.(3.14), is maximized.

In the following, we consider the special case of binary class variables, and suggest an algorithm for the learning machine LM_1 that makes use of LOOCV. It incorporates the learning machine LM_2 which conducts the parameter estimation for given target and hyperparameters. In principle, any probabilistic classification algorithm can be used for LM_2. Here, the C-SVM with probabilistic outputs and linear kernel which is implemented in the LIBSVM (Chang and Lin, 2001) is employed. It provides probabilistic outputs and allows to adjust for unbalanced classes by assigning different values for C to the two classes. This method has only one hyperparameter, $\psi = C$, which penalizes high values of the slack variables; see (Vapnik, 1995; Schölkopf and Smola, 2002) for details. The proposed learning machine LM_1 is described in pseudo code in Algorithm 2.

3.2.4 Model Evaluation and Significance Testing

The generalization performance of the target selection model trained on D using Algorithm 2 can be assessed on an independent test set T by evaluating the empirical log-likelihood gain at the optimal values C_{opt} and t_{opt} chosen on the training set, $\hat{G} = G_{emp}(C_{opt}, t_{opt}|T, D)$.

If the available sample size is small (as it is usually the case with genotype-phenotype

Algorithm 2 Learning Machine for Target Selection

BEGIN PROCEDURE
  for all leave-one-out CV folds $k = \{1, \ldots, N\}$ **do**
    Training set D_k
    Test point T_k
    for $t = \{1, 2, \ldots, m\}$ **do**
      Estimate $\hat{p}_\phi(y_t(T_k); D_k)$
      for $C = \{1, 2, 4, 8, 16\}$ **do**
        Train C-SVM on D_k using Y_t and C
        Calculate likelihood $\hat{p}_\theta(y_t(T_k)|\mathbf{x}; D_k)$
        Calculate $G_{emp}(C, t|T_k, D_k)$
      end for
    end for
  end for
  $\bar{G}_{emp}(C, t) = \frac{1}{N} \sum_k G_{emp}(C, t|T_k, D_k)$
  $(C_{opt}, t_{opt}) = \arg\max_{(C,t)} \bar{G}_{emp}(C, t)$
  Train C-SVM on all samples in D using t_{opt} and C_{opt}
END PROCEDURE

data), cross-validation (CV) has to be used for model evaluation. Therefore, a nested CV procedure is employed: the outer CV loop is used for evaluating the predictive performance, the inner loop for hyperparameter optimization and target selection. In the following we restrict ourselves to the special case of leave-one-out cross-validation, however for larger datasets other CV schemes might be preferred. The final performance measure, the LOOCV estimation of the empirical log-likelihood gain, is calculated as

$$\hat{G}_{LOOCV}(D) = \frac{1}{N} \sum_{n=1}^{N} G_{emp}(C_{opt}^{(n)}, t_{opt}^{(n)}|T_n, D_n), \tag{3.17}$$

where $C_{opt}^{(n)}$ and $t_{opt}^{(n)}$ are the optimal values for C and t obtained with Algorithm 2 on the training set of the n^{th} outer LOOCV loop.

In order to determine whether it is likely that the obtained value of $\hat{G}_{LOOCV}(D)$ could have been achieved even though inputs and targets are in fact independent, a numerical significance test using $\hat{G}_{LOOCV}(D)$ as test statistic can be conducted. This test is based on generating a sufficiently large number of datasets in which the values of the whole set of target variables $\mathbf{Y}$ are randomly permuted. This does not change the empirical distributions $P(X_1, X_2, \ldots, X_d)$ and $P(Y_1, Y_2, \ldots, Y_m)$, so the internal correlations among the members of $\mathbf{X}$ and $\mathbf{Y}$ are not changed. Instead, the purpose is to destroy the correlations between $\mathbf{X}$ and $\mathbf{Y}$. Any remaining correlations are purely by chance, arising from a particular random permutation. The generated datasets therefore represent samples from the factorizing distribution $P(\mathbf{X})P(\mathbf{Y})$, and can be used to assess the H_0 hypothesis of independence.

To do this correctly, the whole nested cross-validation procedure for model learning (including target selection, hyperparameter selection and model parameter fitting), as well as the model evaluation using $\hat{G}_{LOOCV}$ has to be carried out on each of the generated datasets. The p-value of the test is then calculated by dividing the number of permuted datasets on which a value of $\hat{G}$ was achieved which was higher or equal to the one on the original dataset by the total number of permutations. This p-value gives the significance of the learned model in terms of its empirical log-likelihood gain, and can be compared to a chosen significance level α, which bounds the number of type I errors. If the p-value is lower than α, the H_0 hypothesis is rejected.

3.3 Application to Genetic Association Analysis

3.3.1 Introduction

The goal of genetic association studies is to determine whether there is a significant relation between a set of genetic variables and a set of phenotypes. For example, several psychiatric diseases, like alcohol dependence and schizophrenia, are known from twin studies to possess a major genetic component. For these so-called complex diseases, it is assumed that a large number of small genetic effects as well as environmental factors influence the disease risk of an individual. Although several candidate genes have been proposed to have an influence, the genetic disposition for these diseases is yet far from being completely understood. Instead of trying to find a direct association with disease status, a promising strategy is to measure some intermediate phenotypes, which are more directly related to neurobiology. It is considered to be more likely that a genetic effect can be found for these intermediate phenotypes, which can for example be obtained using medical imaging techniques (like MRI, fMRI, EEG), and are usually encoded by a set of real-valued variables.

The measured genetic data usually consists of a set of promising candidate markers, so-called single nucleotide polymorphisms (SNPs). These are point mutations in which a single nucleotide (A, C, G or T) is exchanged and form the main source for genetic differences between individuals. Because there are two chromosomes, two alleles are measured for each SNP. Since it is not known from which of the chromosomes an allele originated, each SNP appears either in one of two homozygous forms (two identical alleles) or in the heterozygous form (two different alleles). In chapter 2, the following encoding scheme for the genetic variables was suggested: Each SNP is represented by two binary variables, coding for the presence of each of the two alleles. Moreover, in order to assess multiplicative interactions between groups of SNPs, a further set of binary interaction variables can be constructed.

To apply the target selection paradigm, the set of binary genetic variables is considered as set of targets $\mathbf{Y}$, and the phenotypic variables are considered as set of real valued inputs $\mathbf{X}$. Then the framework described in section 3.2.4 can be used to test for significant associations.

3.3.2 Experiments

In this thesis, the proposed method is applied to the re-analysis of the data from chapter 2, where a genetic association study was conducted to analyze genetic effects on phenotypes associated with alcohol dependence. The study contained 38 alcohol dependent patients and 41 controls. For details about medical background, data acquisition and statistical analysis see chapter 2.

The genetic data consisted of two sets of markers from two candidate genes, catechol-O-methyltransferase (COMT) and metabotropic glutamate receptor 3 (mGluR3). Six mGluR3 SNPs (rs1990040, rs2214653, rs10238436, rs6947784, rs2299219, rs10266758) and three COMT SNPs (rs2097603, rs4680, rs165599) were incorporated in the analysis. One aim of the study was to look for potential interactions between the COMT-mGluR3 genes, therefore a set of genetic interaction variables was constructed. All 18 alleles of the 9 SNPs and all pair-wise multiplicative interactions which involved one allele from the COMT gene and one allele from the mGluR3 gene were included, yielding a total of 90 binary genetic variables.

As phenotypes, volumetric measurements were obtained using structural MRI and segmentation in a standardized 3D space. The left/total/bilateral volume of hippocampus, nucleus pallidus and nucleus accumbens were included in the genetic association analysis.

In this work, the proposed target selection method was applied to re-analyze the above dataset. For each of the three brain structures a separate test was conducted, and the false discovery rate for these three tests was controlled using the Benjamini-Hochberg-Procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). The three volumetric variables of each brain region were used as input variables, and the 90 binary genetic variables were used as classification targets.

3.3.3 Results

The target selection method was applied to all three brain structures separately for the patient and the control sample. A significant dependency (controlling the false discovery rate at 0.05) was found only for the hippocampus and the patient sample. The selected target variable and the p-value are shown in Table 3.1. The selected genetic variable corresponds to a multiplicative interaction variable between one COMT and one mGluR3 SNP, which suggests the presence of an epistatic interaction between these two systems.

For ease of comparison, the results obtained in chapter 2 using MLGPA are reprinted in Tables 3.2 and 3.3. Models with a significant genotype-phenotype association were found for left, right and bilateral hippocampus, each involving a linear combinations of three genetic variables. The interaction variable which was selected by the target selection model was also present in the regression model, where it was always assigned the largest regression coefficient.

Hippocampus		
Selected Target Variable	$\hat{G}_{LOOCV}(D)$	p-value
rs6947784(C) • rs165599(G)	0.264950	< 0.0005

Table 3.1: Results (Target Selection Method). Significant probabilistic classification model (each allele is described by the name of the SNP and the nucleotide in brackets, the • in the variable names denotes a multiplicative interaction term between two alleles), LOOCV estimation of the predictive empirical log-likelihood gain $\hat{G}_{LOOCV}(D)$, p-value (based on 2100 permutations).

Phenotype	$E_{LOOCV}(D)$	p-value
Right Hippocampus	0.6656240	0.0023
Left Hippocampus	0.715100	0.0057
Bilateral Hippocampus	0.555630	0.0004

Table 3.2: Results from chapter 2: p-values. Estimated generalization mean squared error for significant regression models (controlling the FDR at 0.05) and p-values (10000 random permutations).

Phenotype	Regression Model
Right Hippocampus	$Y = 1.0923 \times$ rs6947784(C) • rs165599(G)
	$+0.75052 \times$ rs2214653(G) • rs165599(G)
	$-1.044 \times$ rs10266758(G)
	$+0.31685$
Left Hippocampus	$Y = 0.90765 \times$ rs6947784(C) • rs165599(G)
	$+0.57617 \times$ rs2214653(G) • rs165599(G)
	$-0.76875 \times$ rs10266758(G) • rs165599(A)
	$+0.097414$
Bilateral Hippocampus	$Y = 1.0873 \times$ rs6947784(C) • rs165599(G)
	$+0.6026 \times$ rs2214653(G)) • rs165599(G)
	$-0.63139 \times$ rs10266758(G)) • rs165599(A)
	-0.063448

Table 3.3: Results from chapter 2: Regression models. Significant Regression models (patient sample).

3.3.4 Discussion

The application of the target selection method to genotype-phenotype analysis was illustrated by re-analyzing data from an imaging genetics study. Using target selection, a significant genotype-phenotype relationship was found between an interaction variable (involving an allele from both the COMT and the mGluR3 gene) and the volume of the hippocampus in alcohol dependent patients. This finding is consistent with the results of the previous analysis in chapter 2, which was based on a regression model and the use of feature selection. The genetic interaction variable found in the target selection model corresponded to the variable with largest regression coefficient in all three regression models. The result obtained from target selection indicates that this interaction variable on its own has already a significant association to the hippocampal volume.

The main difference between the two approaches lies in the model assumptions one wishes to make: The method described in chapter 2 assumes a model which includes additive interactions between the genetic variables. This, however, comes at the cost of a more complex model in which some of the selected regressors might not represent a population effect, but are due to fitting noise in the given sample. The proposed method based on target selection, on the other hand, restrict itself to model a single genetic effect, which reduces the model complexity, and it is more likely that the genetic effect can be reproduced in the population. Moreover, interactions between different phenotypes can be taken into account.

3.4 Summary and Conclusions

In this chapter, target selection was introduced as a new learning paradigm for classification, which can be used for detecting a significant input-target relationship in the data. The difference to the traditional classification paradigm with multiple target variables lies in the fact that the learning machine choses one of the target variables, and learns the prediction function only for the selected target. The objective function for target selection is the expected log-likelihood gain, which was derived from the concept of mutual information.

This specific objective function was chosen since the usual objective of minimizing the total misclassification error is not suitable for the target selection paradigm. The reason for this is that for different classification targets the relative size of the classes will usually vary, i.e. some class labels will appear more often in the underlying population than others. Therefore, the larger classes will have a higher a priori probability, and these prior class probabilities will usually depend on the target variable. In the case of genetic association analysis, this means that some genetic variants will be more frequent in the population than others. If a learning machine always choses the larger class irrespective of $\mathbf{X}$, targets with a very unbalanced class distribution will still achieve a low misclassification error. In the extreme case $P(Y = +1) \rightarrow 1$, all examples in the training and test set will be from the larger positive class, so zero misclassification error

can be achieved by this procedure. However, the goal of target selection is to find a target for which a strong statistical dependency between $\mathbf{X}$ and Y can be established. Therefore, for target selection an objective function must be used which only takes the additional information into account which is provided by the input variables. The expected log-likelihood gain fulfills this criterion.

However, the expected log-likelihood gain is not accessible in practice, since the true distribution is unknown. Instead, the use of the empirical log-likelihood gain on a test dataset was suggested. A learning machine for target selection which is based on the use of a C-SVM with probabilistic outputs as a classifier was proposed. However, the described framework can easily incorporate other probabilistic classification algorithms. The learning machine for target selection maximizes the empirical log-likelihood gain with respect to the selected target and the hyperparameters of the classification function using cross-validation on the training sample.

The proposed method was applied to the field of genotype-phenotype analysis, where it was used to model the effect a single genetic variable has on a multidimensional pattern of phenotypes. In a re-analysis of the data from the genomic imaging study in chapter 2, the finding of a significant genotype-phenotype association for the hippocampal volume could be reproduced. However, the target selection method yielded a sparser genetic model, which included only the regressor with the largest regression coefficient from the regression model learned by MLGPA. In general, the target selection model can be applied to settings where it is assumed that a single genetic variable (which can be a multiplicative interaction variable) influences a multidimensional vectorial phenotype. The genetically influenced pattern for the volumetric variables analyzed in this study was low-dimensional and can be expected to be rather simple, so that a linear classifier already gives good results. However, in tasks like the identification of genetic influences in a multi-voxel pattern analysis of an fMRI study Norman et al. (2006), the phenotypic vectors will be high-dimensional, and the phenotypic patterns influenced by the genotype could be rather complex. Such problems are well-suited for the application of the target selection paradigm.

While a probabilistic SVM was used in the experiments, other probabilistic classification methods could also be applied, e.g. probabilistic artificial neural networks.. Moreover, the target selection paradigm could also be combined with the use of feature selection. Sometimes, the phenotype is influenced by non-genetic variables, such as sex and age. Chance correlations between these covariates and one of the genetic variables might lead to spurious genotype-phenotype associations. However, the target selection paradigm can take this into account by testing for associations between each genetic variable and the covariates, and excluding genetic variables for which a significant association was found from the analysis.

4

Model Comparison in Genomic Imaging Using Information Criteria

4.1 Introduction

In exploratory data analysis, there is a large or even infinite number of hypotheses, and prior knowledge enters mainly via constraints to the functional form or complexity of a predictor. Examples for this have been encountered in chapters 2 and 3, where different assumptions about the general form of the genotype-phenotype relationship were made. In chapter 2 additive and multiplicative effects between alleles at a small number of loci were considered, while in chapter 3 it was assumed that a single genetic variable is associated with a multidimensional phenotype. The choice of candidate polymorphisms and phenotypes included in the analysis was made taking prior knowledge and domain expert knowledge into account. However, the search for the exact model was in both cases exploratory, and a large number of models was included in the analysis without giving preference to any of them. In fact, model selection itself was part of the analysis, either via the mechanism of feature selection or the newly introduced framework of target selection. Note that these methods of analysis result in one final model, chosen by the learning machine, but all other models are discarded.

In contrast to the above exploratory setting, a researcher often has a limited set of certain well-specified candidate hypotheses (or models), for which there is some scientific support, usually from previous studies or the literature. In a hypothesis-driven analysis, a dataset is obtained, and the goal is to use this data for comparing a limited set of models which were chosen a priori. Such a study can be seen as part of a larger modeling process, in which multiple hypotheses are entertained. Whenever relevant experimental or observational data is collected, it will lend more support to

some of the hypotheses and less support to others. Repeated collection of data (also by other research groups) will then over time lead to a refinement of the 'working set' of model hypotheses.

In this chapter, a method for comparing a set of genotype-phenotype association models based on different haplotypes or single SNPs using information criteria is suggested. For this purpose, several models of genotype-phenotype association based on hypotheses derived from former studies needed to be compared on a given dataset, which combined functional magnetic resonance imaging (fMRI) measurements with the genotyping of candidate SNPs.

Since the phenotypes correspond to the blood oxygenation level dependent (BOLD) activity measured by fMRI, they are quantitative variables, and the models take the form of multiple linear regression models with a varying number of free parameters. Moreover, the models are not nested, which means that the regressors of a simpler model are not always included in the more complex ones. The goal is to find the most suitable model that explains the data. Naively, one might say that the model which fits the data best (has the highest likelihood) is also the best model. However, since models of higher complexity also have a higher capacity for fitting the data well, there is the risk that they 'overfit' on the given data sample (Breiman et al., 1984). If that happens, the model will not generalize well to other datasets sampled from the population. In fact, the purpose of statistics lies in the realization of appropriate predictions (Akaike, 1985). The importance of the concept of future observation to clarify the structure of an inference procedure was already realized by Fisher (Fisher, 1935) and Guttman (Guttman, 1967). In this predictive point of view, one is interested in the ability of a model to generalize to the population, not the fit of the model on a specific sample. Therefore, a model of exactly the right complexity needs to be chosen. This is often interpreted as an application of the principle of Occam's razor (Duda et al., 2001), which says that while a model should explain the data well it should also be as simple as possible. Necessarily, the choice of model will depend on the available sample size, since the smaller an effect the more examples are needed to detect it. Thus for comparing a set of candidate genotype-phenotype models on a given data sample, suitable theoretical concepts have to be employed that allow weighing the model fit against the model complexity. Likelihood ratio tests (Wald, 1943) are able to take the model complexity into account. However, they are limited to pairwise comparisons and require nested models, and therefore cannot be applied here. Instead, alternative methods based on information theory or Bayesian analysis need to be employed.

Information criteria allow the selection of a 'best' model from a set of candidates and the ranking and weighting of the remaining models (Burnham and Anderson, 2002). In contrast to likelihood-ratio tests, the model comparison based on information criteria does not require the models to be nested, i.e. the methods can be employed even when the models possess different sets of regressors. Moreover, all models can be compared at once, and there is no need for pairwise comparisons. In this chapter, two information criteria, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), are used to compare different candidate hypotheses. They have been selected

for their thorough theoretical foundations, which will be explained in section 4.2.

The proposed method is applied to a genomic imaging study, which combines genetic analysis with function magnetic resonance imaging (fMRI). The goal of this study was to assess whether haplotype analysis has an advantage over single SNPs methods in a model comparison of the effects of the COMT gene on the central processing of aversive stimuli. Studies on genomic imaging have increased considerably over the last years, and the application of haplotypes also in this field of research has become popular. However, the observation that a combination of different genetic markers increases explained variance of functional brain activation does not necessarily mean that usage of haplotypes is always preferable, since models of higher complexity have a higher capacity for fitting the data well, but might also lead to over-fitting. For this study, catechol-O-methyltransferase (COMT), one of the most extensively studied examples in genetic psychiatric research and genomic imaging was used for comparison of individual single nucleotide polymorphism (SNP) and haplotype analysis.

4.2 Criteria for Model Selection

In the case of multiple linear regression models the model complexity is captured by the number of free parameters k. Let the set of hypotheses be denoted by $\mathcal{M}$. In case the models in $\mathcal{M}$ are nested (each model is a sub-model of all more complex ones) a pairwise comparison between models can be done by using a likelihood-ratio test. Consider a given dataset D with sample size n, and two nested models $M_1 \subset M_2$, where M_1 has k_1 parameters $\boldsymbol{\theta}_1$ and M_2 has k_2 parameters $\boldsymbol{\theta}_2$ ($k_2 > k_1$). Let $p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)$ denote the likelihood of model M_i at the maximum likelihood parameters $\hat{\boldsymbol{\theta}}_i$. Then the log likelihood ratio between the two models is asymptotically χ^2-distributed,

$$-2\ln\frac{p_{M_1}(D|\hat{\boldsymbol{\theta}}_1)}{p_{M_2}(D|\hat{\boldsymbol{\theta}}_2)} \overset{n\to\infty}{\longrightarrow} \chi^2_{k_2-k_1}, \tag{4.1}$$

with the degrees of freedom corresponding to the difference in the number of parameters (Wald, 1943; Akaike, 1985). The likelihood-ratio test rejects the less complex hypothesis M_1, if the value of the left hand side in eq.(4.1) is larger than the $(1-\alpha)$ point in a χ^2 distribution with $k_2 - k_1$ degrees of freedom, where α is the previously chosen significance level, e.g $\alpha = 0.05$. Despite their widespread use, the problem with significance levels in general is that they are chosen arbitrarily: at level $\alpha = 0.05$, a p-value of $p = 0.0499$ will be regarded as a significant finding, while $p = 0.0501$ would be deemed insignificant, although their actual difference is in fact quite negligible. Specific values such as 0.01 or 0.05 have no inherent meaning; they were simply chosen by the scientific community to filter out a sufficient number of false positive results in publications, but using them means introducing an artificial dichotomy between 'good' and 'bad' models, where actually the evidence or support for a model should be viewed as a smoothly varying function of the data.

More natural ways of dealing with this problem are founded in information theory (Cover and Thomas, 1991) or Bayesian analysis (Duda et al., 2001; MacKay, 2003)

and directly compare the support provided by the data for each model within a set of promising candidates. In this thesis, two information criteria will be used for this purpose, AIC and BIC. While AIC and BIC are derived from a different theoretical background, both have a similar form: They consist of the logarithm of the likelihood obtained using the maximum likelihood estimator, penalized by a term which depends on the number of degrees of freedom. In contrast to likelihood-ratio tests, they do not require that the models are nested, and are not limited to pairwise comparisons.

4.2.1 Akaike Information Criterion

The AIC was originally proposed by Akaike (Akaike, 1974). For a probabilistic model M described by a parametric density p_M with k parameters $\boldsymbol{\theta}$ and a given dataset D, it has the form

$$AIC(D, M) = -2 \ln p_M(D|\hat{\boldsymbol{\theta}}) + 2k, \tag{4.2}$$

where n is the sample size and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the parameters. Lower scores are better, i.e. the model which minimizes $AIC(D, M)$ is to be preferred (Akaike, 1985). The first term in eq.(4.2) corresponds to the negative log likelihood function, the second one penalizes the number of parameters.

In the case of regression analysis, which is based on the minimization of mean squared error, AIC takes the form

$$AIC(D, M) = n \ln \sum_{i=1}^{n} \frac{\hat{e}_i^2}{n} + 2k. \tag{4.3}$$

where the $\hat{e}_i$'s are the estimated residual errors under the maximum likelihood model. The number of free parameters k is the total number of regression coefficients including the intercept plus one for the residual variance (Burnham and Anderson, 2002).

The theoretical derivation of the AIC requires the concept of the Kullback-Leibler divergence (KL-divergence). The KL-divergence (Kullback and Leibler, 1951) measures the difference between two probability densities $p(x)$ and $q(x)$ for a continuous random variable X as

$$D_{KL}(p||q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} dx. \tag{4.4}$$

The KL-divergence is always non-negative, and only zero if $p(x) = q(x)$. The term "divergence" is used instead of "distance", since, although it provides a notion of how different to probability distributions are, the KL-divergence is not a metric: it is not symmetric, i.e. in general $D_{KL}(p||q) \neq D_{KL}(q||p)$, and the triangle equation is not fulfilled (Cover and Thomas, 1991).

Let us now assume that the data was generated by drawing independent samples from a hypothetical density $p(x)$, which is called a "generative model", since it is assumed to have generated the data. If $q(x)$ is assumed to be a model of $p(x)$ which is

described by a parameter vector θ, a so-called parametric density model, $q_{\boldsymbol{\theta}}(x)$. Then the KL-divergence $D_{KL}(p||q_{\boldsymbol{\theta}})$ corresponds to the information loss which occurs if the model $q_{\boldsymbol{\theta}}(x)$ is used instead of the generative model $p(x)$.

It can be shown (Burnham and Anderson, 2002) that (under certain regularity conditions) minimizing the AIC corresponds to minimizing an asymptotically unbiased estimator of the expected KL-divergence between the density $p(x)$ and the maximum likelihood estimate $q_{\hat{\boldsymbol{\theta}}_n}(x)$ on an independent and identically distributed (i.i.d.) sample $D = (x_1, \ldots, x_n)$ of size n drawn from $p(x)$,

$$E_D\left[D_{KL}(p\|q_{\hat{\boldsymbol{\theta}}_n})\right] = \underbrace{\int p(x)\ln p(x)dx}_{=const} - E_D\left[\int p(x)\ln q_{\hat{\boldsymbol{\theta}}_n}(x)dx\right]. \tag{4.5}$$

Here, the expectation is denoted by E_D and taken with respect to the distribution of an i.i.d. sample of size n. The first term on the right hand side is a constant which does not depend on $\boldsymbol{\theta}$. The expected KL-divergence, eq.(4.5), describes the expected loss of information about x when a maximum likelihood model estimated on the given dataset is used to approximate reality. Model selection can be done using the AIC by choosing from a given a set of models $\mathcal{M} = \{M_i\},\ i = 1, \ldots, m$ the model M_j which fulfills

$$M_j = \arg\min_i AIC(M_i, D). \tag{4.6}$$

This procedure simultaneously conducts model selection and parameter estimation (Akaike, 1981) and is asymptotically equivalent to cross-validation (Stone, 1977). AIC does not require that the generative model is within the class of candidate models, it just requires that it is close enough in sense of the Kullback-Leibler divergence (Burnham and Anderson, 2002). Note that the AIC is an asymptotic criterion which requires $n \to \infty$ and strictly holds only if $p(x) = q_{\boldsymbol{\theta}_0}(x)$ is an element from the model class. If $p(x)$ is not a member of the model class, but close enough, it still holds in approximation.

Since the AIC does not require nested models and can be used to compare an arbitrary number of different models, it is more generally applicable than likelihood-ratio tests (Akaike, 1981, 1985). Moreover, it is based on a well defined criterion (finding the model with the lowest expected KL-divergence to reality) and does not require the choice of an arbitrary significance level. An interesting question is whether the use of the AIC decision criterion can be viewed in the framework of likelihood ratio test, in order to get an insight into their relation. In the following this is done for the task of deciding between two nested models $M_1 \subset M_2$, where one can directly compare the effect of AIC as model selection criterion to the likelihood-ratio test eq.(4.1). Model M_1 will be rejected according to AIC if

$$AIC(D, M_2) < AIC(D, M_1). \tag{4.7}$$

Δk	α	Δk	α
1	0.1573	8	0.0424
2	0.1353	9	0.0352
3	0.1116	10	0.0293
4	0.0916	20	0.0050
5	0.0752	30	0.0009
6	0.0620	40	0.0002
7	0.0512	45	0.0001

Table 4.1: Comparison between AIC and Likelihood-Ratio Test. Significance levels α of a likelihood ratio test corresponding to the AIC for different values of the difference in the number of free parameters Δk.

Inserting eq.(4.2) yields

$$-2\ln\frac{p_{M_1}(D|\hat{\boldsymbol{\theta}}_1)}{p_{M_2}(D|\hat{\boldsymbol{\theta}}_2)} > 2(k_2-k_1), \tag{4.8}$$

where the left side has the same form as in eq.(4.1). This criterion corresponds to conducting a likelihood-ratio test where the significance level α is not fixed to some arbitrary value (e.g. 0.05), but depends on $\Delta k = k_2 - k_1$. By equating the right hand side of eq.(4.8) with the values of the $\chi^2_{\Delta k}$ statistic, the significance level a specific value of Δk corresponds to in a likelihood-ratio test can be calculated. This was done for various values of k and displayed in table (4.1). This calculation shows that with increasing complexity difference between the two models, the χ^2-test significance level corresponding to the AIC is lowered, so it becomes harder to reject the less complex model M_1 in favor for the more complex model M_2. Note that AIC does not require nested models; they were only used in the above analysis to allow a direct comparison to the likelihood-ratio test.

The individual values of the AIC contain arbitrary constants and depend on sample size, so only their differences are interpretable. The differences to the minimal value within the considered set of m models $\mathcal{M} = \{M_i\}, i = 1 \dots m$,

$$\Delta_i = AIC(D, M_i) - \min_j AIC(D, M_j), \tag{4.9}$$

correspond to the loss of information if model M_i is used instead of the best one in the set. Since the constants and sample size effects have been removed from these rescaled values, they can be used to compare the evidence for models in different scenarios. Using these values, Akaike (Akaike, 1981) defined the likelihoods for specific models as $p(D|M_i) = \exp(-\Delta_i/2)$, which can be normalized over the model set to yield the so-called Akaike weights,

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{j=1}^{m}\exp(-\Delta_j/2)}, \quad i = 1 \dots m. \tag{4.10}$$

In a Bayesian interpretation, the Akaike weights w_i correspond to the probability that the model M_i is the best model with respect to the expected KL-divergence. Then the ratio of two Akaike weights provides an evidence ratio for the respective models.

If for the most complex model in the set the condition $n < 40k$ holds, the asymptotic AIC should be replaced by the AICc (Sugiura, 1978), which includes a small sample correction term and is given by

$$AICc(D, M) = -2 \ln p_M(D|\hat{\boldsymbol{\theta}}) + 2k + \frac{2k(k+1)}{n-k-1}. \tag{4.11}$$

In the large sample limit AICc becomes equivalent to the AIC and retains its asymptotic properties. It is therefore advocated in the literature (Burnham and Anderson, 2002) that AICc should always be used in practice.

In the case of regression analysis, the AICc can be calculated as

$$AICc(D, M) = n \ln \sum_{i=1}^{n} \frac{\hat{e}_i^2}{n} + 2k + \frac{2k(k+1)}{n-k-1}. \tag{4.12}$$

4.2.2 Bayesian Information Criterion

The BIC, also called Schwarz-Criterion (Schwarz, 1978), has the form

$$BIC(D, M) = -2 \ln p_M(D|\hat{\boldsymbol{\theta}}) + k \ln n, \tag{4.13}$$

where p_M is a parametric density model with k parameters $\boldsymbol{\theta}$ and D is a dataset with sample size n. Again, the lower the score the better the model.

In the regression case the BIC can be expressed as

$$BIC(D, M) = n \ln \sum_{i=1}^{n} \frac{\hat{e}_i^2}{n} + k \ln n, \tag{4.14}$$

where the $\hat{e}_i$'s are the estimated residual errors under the maximum likelihood model. The number of free parameters k is the total number of regression coefficients including the intercept plus one for the residual variance (Burnham and Anderson, 2002).

In the following, it is shown how the BIC is obtained as the asymptotic limit of a Laplace approximation of the likelihood of the model parameters. Assuming a set of candidate models $\{M_i\}, i = 1, \ldots, m$, parameterized by $\boldsymbol{\theta}_i \in \Theta_i \subset R^{k_i}$, the posterior probability for the model M_i given the dataset D can be calculated according to Bayes' Theorem as

$$P(M_i|D) = \frac{p(D|M_i)P(M_i)}{p(D)} = \frac{p(D|M_i)P(M_i)}{\sum_{j=1}^{m} p(D|M_j)P(M_j)}, \tag{4.15}$$

where $p(D|M_i)$ is the likelihood that the dataset was generated by model M_i and is given by

$$p(D|M_i) = \int p_{M_i}(D|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i|M_i) d\boldsymbol{\theta}_i, \tag{4.16}$$

with the parameter prior $p(\boldsymbol{\theta}_i|M_i)$. A Taylor approximation of the log likelihood around the maximum likelihood parameter $\hat{\boldsymbol{\theta}}_i$) yields

$$\begin{aligned}\ln p_{M_i}(D|\boldsymbol{\theta}_i) &\approx \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i) + (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)\frac{\partial \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)}{\partial \boldsymbol{\theta}_i} \\ &\quad + \frac{1}{2}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T \frac{\partial^2 \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i^T}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i) \\ &= \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i) + \frac{1}{2}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T \frac{\partial^2 \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i^T}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i), \qquad (4.17)\end{aligned}$$

where the second equality sign holds because at the maximum likelihood solution the gradient of the likelihood with respect to the parameters must vanish. Inserting the exponential of eq.(4.17) into eq.(4.16) gives

$$p(D|M_i) \approx p_{M_i}(D|\hat{\boldsymbol{\theta}}_i) \int \exp\left(\frac{1}{2}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T \frac{\partial^2 \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i^T}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)\right) p(\boldsymbol{\theta}_i|M_i) d\boldsymbol{\theta}_i. \qquad (4.18)$$

If an uninformative, improper prior $p(\boldsymbol{\theta}_i|M_i) = 1$ is chosen for the parameters, the integral in eq.(4.18) corresponds to integrating over a k_i-dimensional unnormalized Gaussian distribution, giving

$$\begin{aligned}p(D|M_i) &\approx p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)(2\pi)^{k_i/2} \left|\det\left(\frac{\partial^2 \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i^T}\right)\right|^{-1/2} \qquad (4.19) \\ &= p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)(2\pi)^{k_i/2} n^{-k/2} \left|\det\left(\frac{1}{n} \cdot \frac{\partial^2 \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i^T}\right)\right|^{-1/2} \qquad (4.20)\end{aligned}$$

Eq.(4.19) can be proven by conducting an orthogonal transformation into the Eigenbasis spanned by the Eigenvectors ξ_j to the Eigenvalues λ_i of $\frac{\partial^2 \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i^T}$, which then becomes a diagonal matrix. Note that since $\hat{\boldsymbol{\theta}}_i$ is a maximum, the Hessian of the log likelihood function is negative definite, and therefore the Eigenvalues λ_j are all negative. The integral can be separated into a product of one-dimensional Gaussian integrals,

$$\begin{aligned}\prod_{j=1}^{k} \int \exp(-\frac{1}{2}|\lambda_j|\xi_j^2) d\xi_j &= \prod_{j=1}^{k} \sqrt{\frac{2\pi}{|\lambda_j|}} = \sqrt{\frac{(2\pi)^k}{\prod_{j=1}^{k}|\lambda_j|}} = \sqrt{\frac{(2\pi)^k}{\left|\prod_{j=1}^{k}\lambda_j\right|}} \qquad (4.21) \\ &= \sqrt{\frac{(2\pi)^k}{\left|\det\left(\frac{\partial^2 \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i^T}\right)\right|}}. \qquad (4.22)\end{aligned}$$

Taking the logarithm of eq.(4.20) yields

$$\ln p(D|M_i) \approx \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i) + \frac{k_i}{2}\ln(2\pi) - \frac{k_i}{2}\ln n - \frac{1}{2}\ln\left|\det\left(\frac{1}{n}\cdot\frac{\partial^2 \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i)}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_i^T}\right)\right|. \tag{4.23}$$

For sufficiently large sample size n this becomes

$$\ln p(D|M_i) \approx \ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i) - \frac{k_i}{2}\ln n. \tag{4.24}$$

Multiplying eq.(4.24) by (-2) yields exactly the BIC, eq.(4.13),

$$-2\ln p(D|M_i) \approx -2\ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i) + k_i \ln n = BIC(D, M_i). \tag{4.25}$$

Therefore, minimizing $BIC(D, M_i)$ corresponds to maximizing the approximated likelihood that the dataset was generated by model M_i. The approximation eq.(4.24) can be plugged into eq.(4.15) to yield an approximate expression for the posterior probability of model M_i,

$$\begin{aligned} P(M_i|D) &\approx \frac{\exp\left(\ln p_{M_i}(D|\hat{\boldsymbol{\theta}}_i) - (k_i/2)\ln n\right) P(M_i)}{\sum_{j=1}^{m}\exp\left(\ln p_{M_j}(D|\hat{\boldsymbol{\theta}}_j) - (k_j/2)\ln n\right) P(M_j)} \\ &= \frac{\exp\left(-0.5BIC(D, M_i)\right) P(M_i)}{\sum_{j=1}^{m}\exp\left(-0.5BIC(D, M_j)\right) P(M_j)} \end{aligned} \tag{4.26}$$

Eq.(4.26) provides the probability that model M_i is the best model of the set (according to the used approximation).

Similar to the analysis carried out for the AIC in section 4.2.1 model selection using the BIC can be directly compared to the likelihood ratio test eq.(4.1) if one considers the special case of two nested models $M_1 \subset M_2$. Model M_1 will be rejected according to BIC if

$$BIC(D, M_2) < BIC(D, M_1). \tag{4.27}$$

Inserting eq.(4.13) yields

$$-2\ln\frac{p_{M_1}(D|\hat{\boldsymbol{\theta}}_1)}{p_{M_2}(D|\hat{\boldsymbol{\theta}}_2)} > (k_2 - k_1)\ln n. \tag{4.28}$$

Note that the left hand side has the same form as in eq.(4.1). The criterion eq.(4.28) corresponds to conducting a likelihood-ratio test with a variable significance level depending on both $\Delta k = (k_2 - k_1)$ and n. Equating the right hand side of eq.(4.28) to the values of the $\chi^2_{\Delta k}$ statistic for specific values of Δk and n yields the significance levels in the corresponding likelihood-ratio test. The results of this analysis are displayed in Table 4.2. This calculation shows that with increasing complexity difference between the two models and with increasing sample size n the χ^2-test significance level corresponding to the BIC is lowered, so it becomes harder to reject the less complex model M_1 in favor of the more complex model M_2.

		n				
		10	100	1000	10000	100000
	1	0.12916	0.031876	0.0085823	0.0024065	0.00069114
	2	0.1	0.01	0.001	0.0001	1e-005
	3	0.074897	0.0031673	0.00012017	4.3409e-006	1.5246e-007
	4	0.056052	0.001021	1.4816e-005	1.9421e-007	2.4026e-009
Δ k	5	0.042107	0.00033375	1.8596e-006	8.8646e-009	3.868e-011
	10	0.010652	1.4036e-006	6.6794e-011	1.9984e-015	0
	20	0.00079296	3.1712e-011	0	0	0
	30	6.4451e-005	7.7716e-016	0	0	0
	40	5.4655e-006	0	0	0	0

Table 4.2: Comparison between BIC and Likelihood-Ratio Test. Shown are the significance levels α of a likelihood ratio test corresponding to the BIC, for different values of difference in the number of free parameters Δk and sample size n.

4.3 Application to Genomic Imaging

4.3.1 Introduction

Genomic imaging has been a productive research field over the last years, and publications in this area have increased dramatically. Catechol-O-Methyltransferase (COMT) on chromosome 22q11, the most relevant dopamine degrading enzyme in the prefrontal cortex, is one of the most extensively studied examples in complex psychiatric diseases. Several studies demonstrated an effect of COMT genotype on frontal activation elicited by working memory function (Egan et al., 2001; Ho et al., 2005) and affective stimuli (Smolka et al., 2005; Drabant et al., 2006; Smolka et al., 2007). A functional relevant single nucleotide polymorphism (SNP) in exon 4 (rs4680,G/A) coding for an amino acid substitution from valine to methionine on position 158 ($Val^{158}Met$) leads to reduced enzyme thermostability. Subjects with the Met allele coding for low COMT expression and presumably high prefrontal extracellular dopamine concentrations display increased working memory performance with less brain activation, suggestive of more efficient information processing (Egan et al., 2001). On the other hand, Met allele carriers have also been suggested to display increased vulnerability for negative mood states, and brain imaging studies showed increased limbic activation elicited by aversive stimuli for this genotype (Smolka et al., 2005, 2007; Drabant et al., 2006).

Whereas initially only discrete genetic variants, mostly SNPs, were analyzed, recent studies in genomic imaging incorporated multiple polymorphisms both within the same gene as well as distributed over several genes (Meyer-Lindenberg et al., 2006; Tan et al., 2007; Buckholtz et al., 2007; Nicodemus et al., 2007). It was assumed that inclusion of multiple genetic variants and haplotypes rather than individual SNPs takes better account of phenotype variablility. Meyer-Lindenberg investigated the effect of

Val158Met in combination with a COMT P2 promoter SNP (rs2097603) as well as a combined analysis with a third SNP located in the 3' region of the gene (rs165599). The study confirmed the hypothesis that optimal prefrontal dopamine concentrations follow an "inverted U shape" dependent on the $Val^{158}Met$ genotype, and that additional genetic variation in the COMT gene can alter dopamine degradation and push subjects above or below optimal dopamine availability. Although the G allele of rs2097603 codes for low COMT expression and thereby increased dopamine availability, subjects with the haplotype G-A (rs2097603-rs4680) display rather inefficient prefrontal response, probably due to too high levels of dopamine. (Nackley et al., 2006) showed that another combination of COMT genetic variants consisting of $Val^{158}Met$ and three adjacent SNPs (rs6269-rs4633-rs4818-rs4680) modulates protein expression according to the stability of mRNA secondary structure of the corresponding haplotype. Specifically, increased COMT activity and consecutive low pain sensitivity (LPS) with Val^{158} was only found when this allele was embedded in the haplotype G-C-G-G (rs6269-rs4633-rs4818-rs4680), whereas the haplotype A-C-C-G displayed the lowest COMT activity and high pain sensitivity (HPS) as a result of its instable mRNA secondary structure. This study highlights the more profound information availed by haplotype analysis compared to single genetic alterations.

The different haplotypes in the COMT gene have never been compared with each other and individual SNPs in a genomic imaging study. The observation that a combination of different genetic markers increases explained variance of functional brain activation does not necessarily mean that modeling genotype effects with more markers is preferable. While models of higher complexity have a higher capacity for fitting the training data well, they might lead to a poor representation of reality, a feature known as overfitting (Breiman, 1994; Bishop, 2006).

In the following study, the above described information criteria are used to evaluate the support which a measured dataset provides for a single SNP model and several haplotype models suggested in the literature for modeling COMT effects on the central processing of affective stimuli.

4.3.2 Materials and Methods

Subjects

60 right-handed healthy volunteers (age 41.0 ± 9.2 [M ± SD] years, 11 females) participated in the study after providing informed, written consent according to the Declaration of Helsinki. The Ethics Committee of the University of Heidelberg approved the study. All subjects were of central European descent. 48 of these subjects have been published in previous studies (Smolka et al., 2005, 2007), while an additional number of subjects (n=12) was recruited for the current study. Standardized clinical assessment with the Structured Clinical Interview I and II (First et al., 1997, 2001) was performed to exclude subjects with a lifetime history of axis I or II psychiatric disorders according to DSM IV and ICD 10. Present drug abuse was excluded with urine tests. Only subjects free of any medication were included. Additionally, we assessed

the level of education, anxiety (SCL-90-R, STAI) (Spielberger et al., 1970; Derogatis, 1983) and depression (SCL-90-R, CES-D, HAMD) (Derogatis, 1983; Hamilton, 1986; Radloff, 1997).

Imaging Study

For emotion induction, we used affectively unpleasant, pleasant and neutral pictures. Each category was represented by 18 pictures. Pleasant and unpleasant cues were taken from the International Affective Picture System IAPS (Lang et al., 1999) in which images are standardized across the dimensions of emotional valence and arousal (Lang, 1995). IAPS catalog numbers were 1440, 2340, 2391, 4220, 4250, 4680, 5260, 5450, 5470, 5480, 5623, 5660, 5830, 7580, 8120, 8190, 8300, 8510 for positive pictures and 2800, 3000, 3015, 3080, 3102, 3140, 3180, 3230, 3261, 3350, 6360, 6570, 9040, 9340, 9520, 9570, 9910, 9921 for negative pictures. Participants were instructed to passively view the stimuli, because even simple rating tasks can alter the brain activation pattern (Taylor et al., 2003).

The stimuli were presented for 750 ms using an event-related design and were arranged in an individually randomized order for each subject. To reconstruct the blood oxygen level dependent (BOLD) event-related time-course, it is necessary to sample data points at different peristimulus time points. This was achieved by a random jitter between intertrial interval and acquisition time, resulting in an equal distribution of data points after each single stimulus. The intertrial interval was randomized between 3 and 6 acquisition times (i.e. 9.9 - 19.8 seconds). During the intertrial interval a fixation cross was presented.

Scanning was performed with a 1.5 T whole-body tomograph (Magnetom VISION; Siemens, Erlangen, Germany) equipped with a standard quadrature head coil. For functional magnet resonance imaging (fMRI), 24 slices were acquired every 3.3 sec (4 mm thickness, 1 mm gap) using an EPI-Sequence (TR = 1.8 ms, TE = 66 ms, $\alpha = 90\,^\circ$) with in-plane resolution of 64 $\times$ 64 pixels (FOV 220 mm), resulting in a voxel size of 3.4 $\times$ 3.4 $\times$ 5 mm^3. FMRI slices were oriented axially parallel to the AC-PC line.

For anatomical reference, we acquired a morphological 3D T1-weighted magnetization prepared rapid gradient echo (MPRAGE) image data set (1 $\times$ 1 $\times$ 1 mm^3 voxel size, FOV 256 mm, 162 slices, TR = 11.4 ms, TE = 4.4 ms, $\alpha = 12\,^\circ$) covering the whole head. After the MRI scan, a subset of the stimuli (8/18 per category) was again presented on a computer monitor for 6 s and assessed for arousal and valence according to the standardized procedure described by (Bradley and Lang, 1994).

Genotyping

For genetic analysis, 30 ml of EDTA blood were collected from each individual. According to prior publications (Meyer-Lindenberg et al., 2006; Nicodemus et al., 2007), six SNPs of the COMT gene were chosen for genotyping (rs2097603, rs6269, rs4633, rs4818, rs4680 [Val^{158}Met], rs165599), see Figure 4.1. Primers were designed for amplification of relevant genomic regions by polymerase chain reaction (PCR); products

were cut by allele specific restriction enzymes and visualized after gel electrophoresis. Primer information and specific assay conditions are available on request. Haplotype construction was performed using HAP Haplotype Analysis Tool (Halperin and Eskin, 2004).

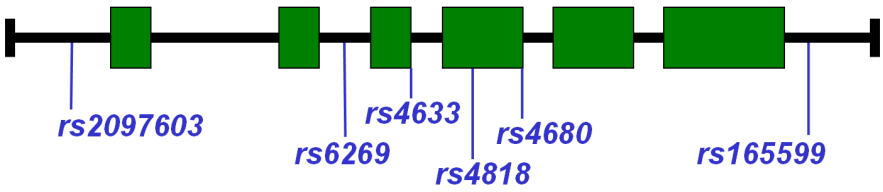


Figure 4.1: Structure of the catechol-O-methyltransferase (COMT) gene
Green boxes depict exons (i.e. the coding sequence which is translated into amino acids). SNPs selected for our study (marked in blue) are scattered throughout the gene, i.e. the promoter region (expression regulating sequence), exons and introns (sequence between exons, some might have regulatory functions, others are functionally inactive).

Data Analysis

Data were analyzed with Statistical Parametric Mapping (SPM5; Wellcome Department of Imaging Neuroscience, London, U.K.). After temporal and spatial realignment, the images were spatially normalized to a standard EPI template using a 12-parameter affine transformation with additional nonlinear components and subsequently re-sampled at a resolution of $3 \times 3 \times 3$ mm^3. The functional data were smoothed using an isotropic Gaussian kernel for group analysis (12 mm FWHM).

Statistical analysis on the first (individual) level was performed by modeling the different conditions (unpleasant, pleasant and neutral pictures; delta functions convolved with a synthetic hemodynamic response function and its time derivative) as explanatory variables within the context of the general linear model (GLM) on a voxel-by-voxel basis with SPM5.

To detect associations between the genotypes and fMRI activation elicited by unpleasant stimuli on a voxel by voxel basis, the contrast images of all subjects (signal change of unpleasant versus neutral pictures) were included in a second level multiple regression analysis with SPM5. All regression models included sex and age as additional covariates, as well as a constant regressor. To assess how much variance is explained by the additional SNPs, we applied five different models (see Table 4.3): The first model (SNP1) only incorporated rs4680 (COMT Val^{158}Met) and genotype was coded by the number of low activity Met^{158} alleles (0, 1 or 2); the second model (SNP 2) incorporated a combination of rs4680 and rs2097603, for the latter SNP, genotype was coded as number of low activity G alleles. Both were informed models, assuming that brain activity elicited by unpleasant stimuli would increase with decreasing COMT activ-

ity. The other three models were based on haplotypes according to (Meyer-Lindenberg et al., 2006) and (Nackley et al., 2006), see Table 4.5. Another haplotype model from

Model	SNPs / Haplotype pattern	k
SNP1	rs4680	5
SNP2	rs4680, rs2097603	6
HAP1	rs2097603-rs4680	7
HAP2	rs2097603-rs4680-rs165599	11
HAP3	rs6269-rs4633-rs4818-rs4680	12

Table 4.3: Different Genetic Models. This table shows the different models, the involved SNPs or haplotype patterns and the total number of free parameters (k).

(Meyer-Lindenberg et al., 2006) rs737865-rs4680-rs165599) was not chosen for our analysis because it showed only a weak effect on activation of the prefrontal cortex during a working memory task. For the selected haplotype models we used haplotypes - 1 regressors depicting the frequency of a specific haplotype (0, 1 or 2) for a subject. In these models, no assumption about the direction of the effect was made.

The analysis of the genotype $\times$ task interaction was restricted to areas showing robust activation related to the task (unpleasant versus neutral stimuli, including age and sex as covariates) as defined by an F-contrast of $p < 0.01$ and a cluster size (K_E) of at least 10 voxels. To estimate the global effect of genotype on brain reactivity, we calculated the number of voxels of the volume of interest (VOI) in which the respective genotype model was significantly associated ($p < 0.05$, uncorrected) with cue elicited brain activity.

Model Comparison Procedure

Two information criteria were used for the model comparison: the Akaike information criterion in its small sample correction (AICc), eq.(4.11), and the Bayesian Information Criterion (BIC), eq.(4.13). Only those voxels in which a task-related activation was found were included in the analysis. For the model comparison it was assumed that the population density can be described by a generative regression model, and all included voxels are considered as different datasets sampled from the same underlying distribution. Altogether, five regression models with different regressors and numbers of parameters were compared. For each model, the maximum likelihood parameters were fitted for each of the included voxels within the general linear model framework of SPM5. The residual mean squared error for each voxel was then used in the calculation of the different information criteria. The free parameters k which is required for the calculation of the information criteria included the number of regressors, the intercept and the residual variance. The information criteria were averaged over all voxels, i.e. over all different datasets sampled from the population density.

SNP	Alleles	Allele frequencies	Published allele frequencies	Genotype frequencies	Published genotype frequencies
rs2097603	A/T*	0.4/0.6	N/A	0.17/0.47/0.37	N/A
rs6269	G/A	0.41/0.59	0.49/0.51	0.15/0.52/0.33	N/A
rs4633	C/T	0.51/0.49	0.48/0.52	0.23/0.55/0.22	0.22/0.53/0.25
rs4818	C/G*	0.59/0.41	0.59/0.41	0.33/0.52/0.15	0.36/0.45/0.19
rs4680	G/A	0.53/0.47	0.48/0.52	0.27/0.53/0.20	0.22/0.53/0.25
rs165599	G/A	0.33/0.67	0.46/0.54	0.13/0.40/0.47	0.22/0.48/0.30

Table 4.4: SNP characteristics. Allele and genotype frequencies of the single nucleotide polymorphisms (SNPs) in our study population, compared to published frequencies from NCBI SNP database (http://www.ncbi.nlm.nih.gov/) in individuals of European descent. If known, alleles are displayed with the ancestral allele first, all other SNPs are marked*. Genotype frequencies are given as homozygous allele 1 / heterozygous / homozygous allele 2. If several database sources are available in NCBI, European HapMap data are chosen. N/A = not available. rs4680 ($Val^{158}Met$) is the functional SNP that has been studied in multiple studies, in our study used as model SNP1. rs2097603 is located in the P2 promoter region of the gene that has also an effect on COMT expression. The combined analysis of rs4680 and rs2097603 was used as model SNP2.

4.3.3 Results

Genotyping

Genotyping results are depicted in Table 4.4. Genotypes of all SNPs were in Hardy-Weinberg equilibrium. Allele frequencies of all SNPs concurred with published data (NCBI SNP database). Haplotype phases were given by HAP Haplotype Analysis Tool, depicting the most probable haplotype. Haplotype frequencies are comparable to published data (Meyer-Lindenberg et al., 2006; Nackley et al., 2006). Identified haplotypes and their frequencies are shown in Table 4.5.

Main effect of the task irrespective of genotype

Significant differences in brain activity elicited by unpleasant stimuli compared to neutral stimuli were found in a total of 5962 voxels ($3 \times 3 \times 3$ mm), i.e. a volume of 161 ml. The main effect of unpleasant versus neutral stimuli comprised a distributed neuronal network of primary, secondary and tertiary visual regions in the occipital, temporal and parietal lobe, bilateral amygdala, hippocampus and parahippocampus, and regions in the left and right medial, ventrolateral and dorsolateral prefrontal cortex (cf. Figure 4.2, left panel). The following analyses of the genotype $\times$ task interactions were restricted to these voxels (VOI).

HAP1 (rs2097603-rs4680)								
AA	AG	TA	TG					
0.28	0.12	0.18	0.42					
HAP2 (rs2097603-rs4680-rs165599)								
AAA	AAG	AGG	AGA	TAG	TGA	TGG	TAA	
0.242	0.042	0.025	0.092	0.025	0.192	0.242	0.142	
HAP3 (rs6269-rs4633-rs4818-rs4680)								
GCGG	ATCA	ACCG	GTCG	ATCG	GCCG	ACGG	ATCA	ATGG
0.392	0.433	0.100	0.008	0.008	0.008	0.008	0.033	0.008

Table 4.5: Selected haplotypes for comparison. Both haplotype models rs2097603-rs4680 and rs2097603-rs4680-rs165599 were studied in (Meyer-Lindenberg et al., 2006). rs6269-rs4633-rs4818-rs4680 comprise the haplotypes examined by (Nackley et al., 2006); GCGG has the highest COMT expression, ATCA has average COMT expression and ACCG has lowest COMT expression according to (Nackley et al., 2006).

COMT genotype and processing of unpleasant visual stimuli

Regression analyses revealed significant associations ($p < 0.05$) between all COMT genotype models and the BOLD fMRI response to unpleasant versus neutral visual stimuli in limbic and prefrontal brain areas as well as in brain areas related to visuo-spatial attention (Fig. 4.2). Depending on the model, brain activation was significantly associated with genotype in 31% (HAP2; 1878 voxels, including 8 haplotypes) and up to 74% (SNP2; 4440 voxels) of the VOI. SNP1, the simplest model incorporating only rs4680 (COMT $Val^{158}Met$), yielded significant associations in 66% (3911 voxels) of the VOI, HAP1, the simplest haplotype model (4 haplotypes) in 42% (2531 voxels) of the VOI, and HAP3, the most complex haplotype model (9 haplotypes) in 44% (2606 voxels) of the VOI.

Comparison of different models

The results of the model comparison are given in Table 4.6. For each of the models, the average values of AICc (eq.(4.11)) and BIC (eq.(4.13)) were calculated. The models with lower values in these information scores should be preferred. The value of the log-likelihood ($LL = n \log \sum_{i=1}^{n} \hat{e}_i^2/n$) is also listed for comparison. A very revealing quantity is the relative belief one should put into the models given the data, according to the information criteria. This is provided in Table 4.6 by the Akaike weights w_{AICc} (eq.(4.10)), which were obtained using the AICc, and the posterior model probabilities $P(Model|D)_{BIC}$ (eq.(4.26)), which arise from the BIC criterion under the assumption of equal prior probabilities.

An interesting result of this analysis is that both of the information criteria assign exactly the same ranking to the five models. This ranking is reflected in the ordering of Table 4.6 from best (top) to worst(bottom). Both assign the strongest belief value

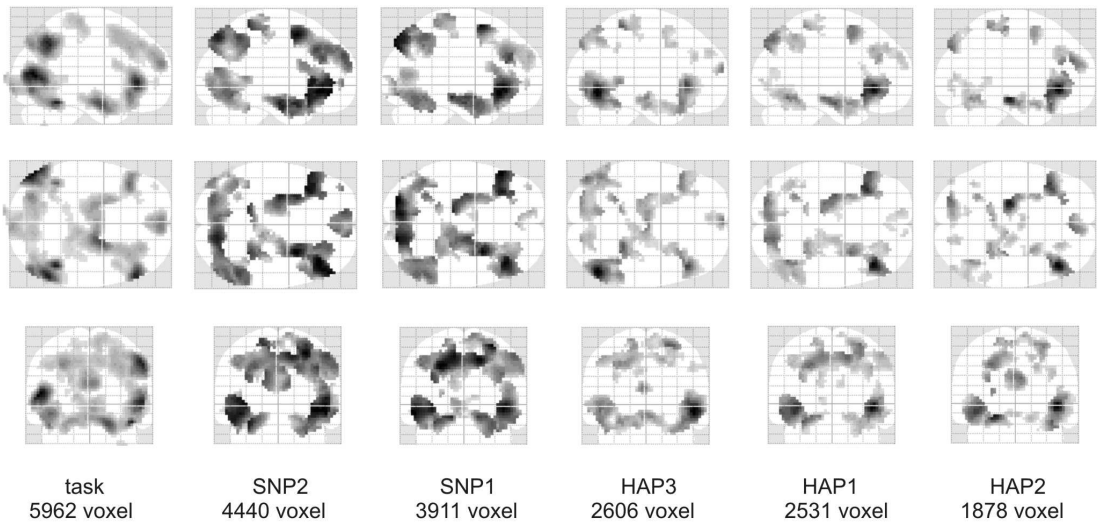


Figure 4.2: Effect of genetic variation of COMT on processing of aversive stimuli
The glassbrain diagram at left depicts the brain regions that are activated by the task itself ($p < 0.05, K_E \geq 10$)) and was used as VOI for the analyses of the effects of the five different genetic models. The other panels show for each of this five models in how many voxels a significant effect of genotype was detected ($p < 0.05, K_E \geq 10$)). SNP1: rs4680; SNP2: rs2097603-rs4680; HAP1: rs2097603-rs4680; HAP2: rs2097603-rs4680-rs165599; HAP3: rs6269-rs4633-rs4818-rs4680.
K_E: extent of cluster in voxels

to model SNP1 (rs4680) (80.8% for BIC and 61.3% for AICc), while the second best model, SNP2, only receives less than half of this value (17.3% for BIC and 30.8% for AICc). The best haplotype based model (HAP1) only gets 1.9% (BIC) and 7.8% (AIC) of belief, the remaining two haplotype models (HAP2 and HAP3) get less than 1 per mill belief for both criteria. If we consider only the log likelihood, without taking the model complexity into account, the SNP1 model still scores best, but HAP1 follows closely and has a higher likelihood than SNP2. However, the information criteria assign a much higher Akaike weight or posterior to SNP2 than to HAP1. As is expected from theory, the AIC preferred more complex models compared to BIC. The fact that both information criteria, though having different theoretical foundations (section 4.2), arive at exactly the same model ranking provides additional support for the results.

4.3.4 Discussion

The method for model comparison based on information criteria was applied to the question whether effects of the COMT gene on central processing of affective cues are best modeled by an individual SNP, a combination of two SNPs or one of several haplotypes suggested in the literature. The main finding of this study is that based on

Model	k	LL	$AICc$	w_{AICc}	BIC	$P(Model\|D)_{BIC}$
SNP1	5	45.993338	57.104449	0.625525	66.465061	0.807627
SNP2	6	44.982722	58.567627	0.300968	69.548789	0.172818
HAP1	7	45.247297	61.401144	0.072984	73.907709	0.019546
HAP2	11	44.866936	72.366936	0.000303	89.904726	0.000007
HAP3	12	42.379705	73.018003	0.000219	91.511840	0.000003

Table 4.6: Results of the Model Comparison using BIC and AICc, showing: Model, number of free parameters (k), value of the log likelihood (LL), value of the Akaike Information criterion ($AICc$), Akaike weights (w_{AICc}), value of the Bayesian information criterion (BIC) and posterior model probability according to the BIC ($P(Model|D)_{BIC}$). See Table 4.5 for the details on the haplotype models.

our data, the strongest belief value was assigned by both AIC and BIC to the model that just included only the (traditionally most often studied) COMT $Val^{158}Met$ SNP (rs4680), while the second best model, which included an additional SNP located in the promoter region, received less than half of this value. This finding may help to explain why many genomic imaging studies found significant effects of the COMT $Val^{158}Met$ polymorphism, even without exploring further genetic variation in the COMT gene (Egan et al., 2001; Ho et al., 2005; Smolka et al., 2005). The initial assumption that additional information on genetic background will be always preferable over a simple genetic model has been refuted in this study.

Note that the more complex an underlying model is, the more data are needed to back it up. Because the noise level and the effect strength are unknown, there is no way of knowing a priori how much data is needed to support a model of a certain complexity. Therefore, it cannot be ruled out that for a replication study with much larger sample size, the model comparison might favor a haplotype based model over the individual SNP model. Thus, for each measured dataset, the model comparison procedures have to be carried out anew. As mentioned before, haplotype-based models suffer from the uncertainty arising in the phase inference. It might be a limitation of our study that we used only the haplotypes with the highest likelihood. If the candidate SNPs are in high LD, uncertainty is low (Balding, 2006) and does not effect the analysis. If the LD is low, the amount of haplotype uncertainty could be directly incorporated into the regression analysis. However, it does not affect the proposed use of information criteria for model comparison.

Altogether, our study elucidates why COMT $Val^{158}Met$ was often successfully correlated with imaging data and suggests a way of finding the best model for genetic analysis. We want to emphasize that our finding does not imply that haplotype-based models are generally inferior to models based on individual SNPs. Haplotype models may be better for explaining COMT effects on cognition performance but not on affective processing. We want to point out, rather, that techniques for model comparison

need to be employed which can establish whether the increase in likelihood potentially achieved with a more complex, haplotype-based model is substantial enough to warrant the increase in model complexity. Whenever haplotype models are employed in genetic studies, we suggest the use of the above technique based on information criteria for model-comparison to identify the most informative model. Although it is tempting to use the model which fits the data best and has therefore the highest likelihood, this may not be the most suitable model for the particular study, genetic variants and phenotypic data.

4.4 Summary and Conclusions

In this chapter, it was shown how model selection techniques based on information criteria can be applied in genomic imaging to assess different models of genotype-phenotype association in a function magnetic resonance imaging (fMRI) study. For this purpose, two model selection criteria, AIC and BIC, were applied to the multiple regression models resulting from a second level analysis of a specific BOLD contrast under different genetic hypotheses. The information criteria used in this analysis were chosen for their theoretical foundations in information theory (AIC) and Bayesian analysis (BIC).

Whenever haplotype models are employed in genetic studies, the proposed technique based on information criteria for model-comparison should be used to identify the most informative model. Although it is tempting to use the model which fits the data best and has therefore the highest likelihood, this may not be the most suitable model for the particular study, phenotypic variables and dataset.

Statistical theory is often interested in the asymptotic behavior of an estimator in the hypothetical case of infinite sample size. If the estimator for an infinite amount of data examples converges to the true value, the estimator is called consistent. It can be shown (Burnham and Anderson, 2002) that BIC is consistent, while AIC is not. This means that if the generative model is contained in the model set, BIC will for infinite sample size always select it as best model, whereas AIC has probability larger than zero to select a more complex model, which, however, will have zero KL divergence to the generative model (Burnham and Anderson, 2002). Note that the likelihood ratio test, eq.(4.1), is also not consistent. The underlying reason for this is that the test always has a probability of type I errors (i.e. false positive rate) of the chosen significance level α, irrespective of the sample size, even if this goes to infinity.

Consistency is concerned with the asymptotic case of infinite sample size and requires that the generative model is in the model set. In reality, sample size is finite, and the generative model, which is a mathematical abstraction in any case, is not necessarily contained in the set. For finite n, the probability that a too simple model is chosen is smaller for AIC than for BIC. Both criteria have their advantages and disadvantages, and both have a different but theoretically meaningful justification. In general, the BIC will result in more parsimonious models than the AIC, which will tend to chose slightly more complex models. If the two criteria disagree on which model is best, or if several

models are assigned almost equal Akaike weights or posterior model probabilities, then more than one model should be considered adequate.

Information criteria like AIC and BIC allow an easy comparison of different models. In comparison to likelihood ratio tests, they can be applied to non-nested model, and do not require significance tests with arbitrarily chosen significance levels. However, this does not justify a procedure in which all possible models are mechanistically enumerated and compared. Instead, the particular choice of candidate models must be based on the domain knowledge and expertise of the researcher and represent the way the researcher is looking at the data (Akaike, 1985; Fienberg, 1980). The model comparison conducted in this chapter meets this requirement, because it is based on a selection of models previously proposed in the literature.

5

Molecule Kernels for QSAR and Genotoxicity Prediction

5.1 QSAR

5.1.1 Introduction

The 3D structure of a molecule is closely related to its physical, chemical and biological properties. This is expressed in the similarity principle: "Similar structures have similar physicochemical properties and biological activities". In quantitative structure-activity relationship (QSAR) analysis the aim is to predict the biological activity, toxicity or mutagenicity of a drug. This is useful in drug discovery during the search for lead compounds, where the aim is to maximize the potency or selectiveness of a drug, while at the same time looking for a compound with good pharmacokinetic properties and minimum toxicity. These depend on the local and global electronic, hydrophobic, lipophilic and steric properties of the compound which are implicitly determined by its 3D structure.

Traditionally, QSAR analysis starts with a representation of the molecular structure, from which a a large number of descriptors are generated, that are concatenated into a descriptor vector. These descriptors replace the initial representation of the molecule. They explicitly encode some aspects of the information which is implicitly contained in the original structure. According to the 'dimensionality' they represent, the descriptors are categorized into different classes: The 0D descriptors contain counts of entities like atoms, elements and bond types, 1D descriptors consist of path and walk counts, 2D descriptors describe the topology of the molecule and are based on the structural formula of a molecule, and 3D descriptors require the reconstruction of the 3D geometry of the compound. These include descriptors obtained by 3D-QSAR methods

using force-field calculations and physicochemical models. On the basis of this descriptor representation, a predictor is learned on the given training data, which assigns a regression value or a class label to a molecule. Since the number of descriptor vectors is very large, often exceeding the available sample size, these predictors usually involve feature selection or feature construction (e.g., principle component analysis) to reduce the input dimensionality. The most popular choice in chemoinformatics is the partial least-squares method (PLS) (Wold et al., 1984, 2001), an embedded method for feature construction in regression tasks.

Recent advances in the field of statistical learning theory (Vapnik, 1998) have led to the development of kernel methods, yielding predictors with very good generalization performance working in high dimensional feature spaces. In general, kernels are functions which take two objects (data points, examples) as input and assign a scalar output value, which is interpreted as a measure of similarity between the objects. The values of the kernel for all pairs of examples in a given dataset are summarized into a matrix called *kernel matrix*. Well-known examples of kernel methods are the support vector machines (Schölkopf and Smola, 2002) for classification and regression. Usually, these approaches are applied to vectorial data; however in the past few years an increasing number of kernel techniques for structured data, like sequences, trees and graphs, have been developed (Bakhir et al., 2007). In QSAR analysis, support vector machines have been mainly applied to descriptor vectors, although recently approaches based on positive-definite graph kernels (Gärtner et al., 2003; Kashima et al., 2003; Ralaivola et al., 2005) have been suggested. These graph kernel methods define a similarity function between two molecules by considering them as graphs, in which atoms correspond to vertices and bonds to edges. However, the runtime complexity of these algorithms often grows quickly with the number of atoms in the molecule. The calculation of an all-subgraphs kernel, which calculates the number of all subgraph-isomorphisms, is practically infeasible (NP-hard) (Gärtner et al., 2003). Other graph kernel approaches count common walks in two graphs (product graph kernels (Gärtner et al., 2003)) or calculate the expectation of a kernel over all pairs of label sequences in two graphs using random walks (marginalized graph kernels (Kashima et al., 2003)). Nevertheless, the runtime complexity of these approaches still scales with $O(n^6)$, where n is the number of atoms in a molecule, making them impractical for larger-size molecules. Recently, computationally more efficient positive-definite graph kernels based on molecular fingerprints have been proposed (Ralaivola et al., 2005). These calculate the similarity of vectors of counts of all labeled paths with a maximum length derived by depth-first searches starting from each vertex of a molecular graph. However, all the above graph kernels make use of path or walk counts, like 1D descriptors, but do not consider the full information contained in the 3D molecular structures.

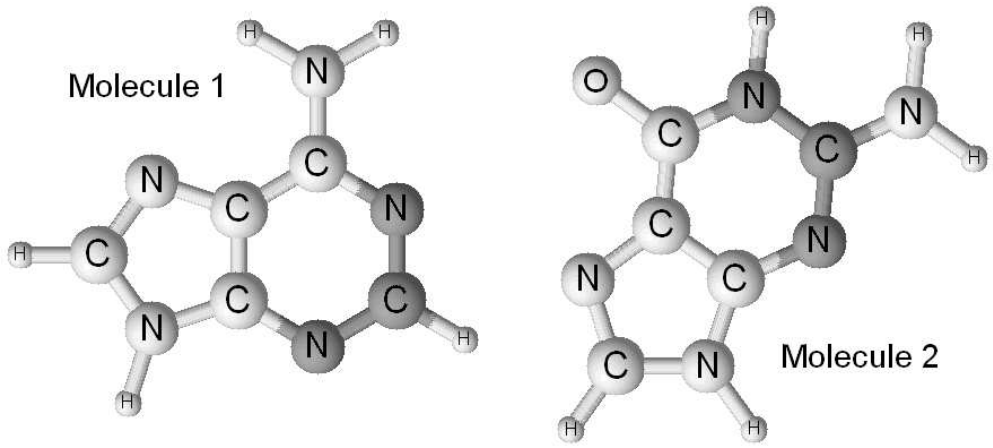
In this chapter a novel kernel method for QSAR analysis is proposed, which is not based on the construction of descriptor vectors, but directly evaluates the similarity between two molecular 3D structures. It makes use of the fact that information about the physicochemical properties is already implicitly contained in the 3D structure. So instead of trying to explicitly extract these properties in form of descriptor vectors, the

similarity between all pairs of molecular structures can directly be used for prediction. To this end, a new family of structural similarity measures called *molecule kernels* is introduced. In contrast to graph kernels, molecule kernels take both topology and 3D geometry of the molecules into account. However, the resulting kernel matrix is not positive definite, a property required by conventional kernel methods like support vector machines. Therefore, the recently proposed potential support vector machine (Hochreiter and Obermayer, 2006) (P-SVM) for dyadic data is used as predictor, which does not require positive definite kernel matrices. If trained on a molecule kernel matrix, the P-SVM implicitly encodes information about certain structural elements or substructures which are relevant for predicting the desired endpoint. Unlike methods based on structural libraries, where the presence or absence of elements from a predefined set of structures is encoded explicitly, in our approach the structural elements are encoded implicitly via the parameters of the predictor and the values of the molecule kernel matrix. Like other 3D QSAR approaches, the proposed method requires suitable 3D conformations, which are assumed to have been determined using geometry optimization techniques or molecular mechanics. However, a spatial prealignment of the compounds with respect to each other or any grid is not necessary.

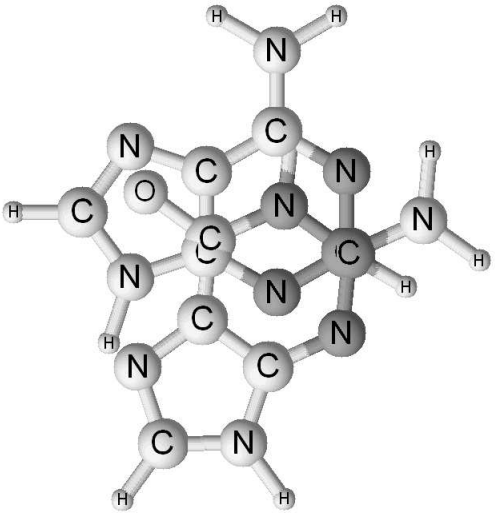
Suitable kernels for molecules need to fulfill two criteria: First, they need to capture specific aspects of the 3D structure of two molecules, which are expressed in form of a similarity function. Second, they need to be efficiently computable. The problem of maximizing a similarity function which depends on the spatial alignment of two compounds leads to a continuous optimization problem with many local optima. If gradient-based optimization techniques are employed, the solution will most likely correspond to a local optimum. A method based on randomized search heuristics has been proposed by Kearsley et al. (Kearsley and Smith, 1990), however, it depends on the choice of an alignment parameter and is not guaranteed to identify the global optimum. The molecule kernels proposed in this chapter transform the continuous optimization problem into a discrete optimization problem by considering the set of all at least locally matching alignments[1]. The smallest units describing a unique 3D alignment of two compounds consists of a pair of ordered bipods (V-shaped subfragments), one from each compound. This is illustrated in figure 5.1, where two molecules are aligned according to the alignment of two matching bipods. Molecule kernels make use of this fact by finding the global optimum in the space of all matching bipod alignments, a problem which can be solved with a runtime complexity of $O(n^4)$.

Moreover, while most other 3D QSAR methods assume the activity of all compounds is mediated by the same interaction mechanism between ligand and receptor protein, the molecular kernel approach does not require this assumption. The reason for this lies in the fact that instead of using a global alignment of all molecules to a common scaffold, receptor, or grid, molecule kernels are based on the optimum pairwise alignments between molecules. The mutual similarities can involve different alignments, which can correspond to different active groups for each pair of molecules.

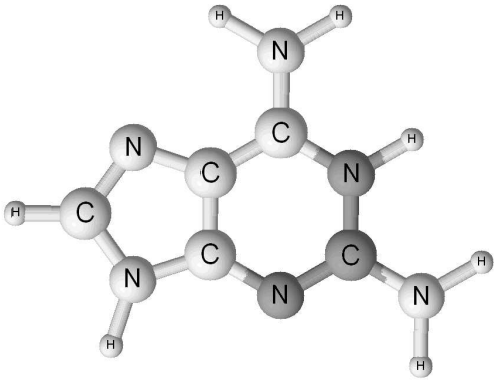
[1] These terms are used here informally to give an intuition, their rigorous definition will be given in the section 5.2.1.



(a) Two Molecules with a pair of matching bipods (marked in dark gray) which will be used for molecular alignment



(b) The same molecules after a translation of molecule 2 such that the central bipod atoms of the two bipods are superposed.



(c) Molecule alignment corresponding to the matching bipod alignment obtained after the rotation which optimally aligns the bipod atoms in a least squares sense

Figure 5.1: Illustration of the concept of a molecular alignment using bipod matching. For ease of visualization this is shown using a simple 2D example. All formal definitions and the details of the alignment procedure will be given in the methods section.

So different mechanisms can be modeled at once, as long as they are all sufficiently represented in the training data.

The predictive performance of the proposed method is compared on publicly available datasets to results from several other QSAR methods which were taken from the literature. In the following, these methods will be briefly reviewed. The technique of hologram QSAR (HQSAR) (Heritage and Lowis, 1999) divides each molecule into

a set of overlapping structural fragments and uses their frequency as 2D descriptors. One widely used 3D-QSAR approach is comparative molecular field analysis (CoMFA) (Cramer et al., 1988), which is based on the assumption that similar steric and electrostatic fields of molecules lead to similar activity. CoMFA requires the spatial superimposition of the molecular structures, which needs expert knowledge, is time-consuming and might introduce user bias. After alignment, a 3D grid is generated around the molecule and local potential fields are calculated at the grid points. Steric fields are modeled by the Lennard-Jones potential, and electrostatic fields are computed using the Coulomb potential of the partial atomic charges. Then the interaction energies between certain probes and the molecule are calculated at the grid points and used as descriptor vectors in order to predict the biological activity. A problem with this approach lies in the abrupt change in potential at the Van-der-Waals surface of the molecule, which leads to a critical dependency of the CoMFA results on grid spacing and the relative orientation of molecules and grid. An alternative technique which is less sensitive to these issues is the comparative molecular similarity index analysis (CoMSIA) (Klebe et al., 1994). It makes use of a Gaussian approximation of the force fields which is not subject to sudden potential changes. In addition to steric and electrostatic fields sometimes also hydrophobic fields and hydrogen bond donor/acceptor fields are modeled. Another 3D QSAR approach is the the GRIND method (Pastor et al., 2000), where GRid INdependent Descriptors encode the spatial distribution of molecular interaction fields (MIFs) based on an autocorrelation function. An extension of this, the anchor-GRIND-method (Fontaine et al., 2005), allows the inclusion of a priori chemical and biological knowledge. The user defines a specific position of the molecular structure (the "anchor point"), which is used as reference in the comparison of the MIFs of the compounds. This allows a better description of the compounds in the vicinity of a common substructure. It requires less human supervision than the previous 3D-QSAR methods, but is only applicable for a set of compounds sharing a common scaffold. Finally, QSAR by eigenvalue analysis (EVA) (Ferguson et al., 1997) uses a descriptor based on molecular vibrational spectra. It has the advantage of being invariant to the spatial alignment of the molecule.

5.1.2 Molecular and Structural Formulas

The *molecular formula* indicates the presence of a certain element in a molecule by the elemental symbol and provides information about the respective number of atoms via a subscript, e.g. H_2O for water. Molecules with the same molecular formula but different topological of spatial arrangement are called isomers.

Isomers fall in several categories: In *structural isomers*, the functional groups lie at different topological positions in the structure or are split up into different subgroups, or hydrocarbon chains have different forms of branching. In contrast to these, *stereoisomers* possess the same topological structure, but some atoms or functional groups have different spatial positions. One subclass are the non-superimposable mirror symmetric *enantiomers*, which often have different chemical properties interacting

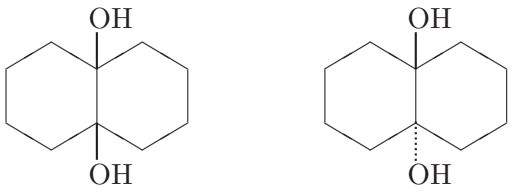


Figure 5.2: The structural formulas of cis- and trans-decalinediol.

with other enantiomers. In pharmacology this is important, since enantiomers often occur in living beings, and sometimes one enantiomer has a desirable effect, while its mirror-symmetric counterpart is toxic. The other subclass of stereoisomers is called diastereomers. A subgroup of this are the *conformational isomers*, which differ by rotation around a single bond. Different conformers can convert into each other by rotation without breaking chemical bonds. Some conformers will have lower energy than others, due to Coulomb, Van-der-Waals or steric interactions with neighboring atoms. In situations where there is a double bond, or sometimes in the presence of ring structures, such a rotation is not possible to actually conduct. Then the isomerism is called *cis-trans isomerism*. Here, 'cis' refers to the case where the substituent groups are oriented in the same direction, and 'trans' to the one where they are oriented in opposite directions.

Structural formulas express information about the topological arrangement of the atoms within a molecule and the nature of bonds (single, double or triple). For organic compounds, a kind of structural formulas often called *skeletal formulas* are usually used as a graphical representation of the topological structure, where each vertex denotes an atom, and each line a bond. The element of an atom is denoted by its elemental symbol, but the elemental symbols of carbon atoms and hydrogen atoms bonded to carbon are left out. Instead, each unlabeled vertex represents a carbon atom and the number of hydrogen atoms attached to a particular carbon atom is equal to four minus the number of bonds drawn to that carbon, since carbon has four covalent bonds. Structural formula cannot distinguish between different conformers. Up and down bond symbols (wedges) can be used to distinguish cis- and trans-configurations. An example for such a structural formula is given in figure 5.2.While this example shows that structural formula can visualize certain 3D topological features, the actual molecular structure of a compound is 3D and its geometry cannot be fully represented by structural formulas. Instead, the relative 3D coordinates of all the atoms needs to be given. *Linear notations* allow a compact representation of chemical structures by alphanumerical strings. An important example is the *SMILES notation*, which was developed in 1986 by Daylight Chemical Information Systems and is often used as a convenient way of entering chemical structures into computer systems. The topology of a molecule can be represented as a graph, whose nodes denote atoms and whose edges

denote chemical bonds, which has lead to the applicability of graph-theoretic methods to computational chemistry. An *adjacency matrix* can be used to describe the topology. A *bond matrix* additionally includes information on the bond order. In computer systems, molecules are often represented using a list of all atoms and a *connection table*, which is a sparse representation of the bond matrix. However, the coding of the molecular graph as matrix does not account for stereochemical information. In order to represent the full 3D molecular structure usually file formats like sd-files or pdb-files are used, which include the 3D position of each atom, and sometimes further information like partial charges.

5.1.3 Geometry Optimization

The 3D coordinates of the atoms can be calculated by determining low-energy conformations using a quantum mechanical approach, where the Schroedinger equation is solved under some approximations, or classical molecular mechanics, where the atoms are considered mass points between which several forces are acting. These give rise to energy terms for bond length, bond angles, torsion angles, as well as Van-der-Waals and Coulomb energy. This energy function can be minimized using procedures like conjugate gradient descent. the coordinates taken from crystallographic databases can be used as initialization. Usually, there is more than one energy minimum, therefore several stable conformations are possible. The space of possible conformations can either be explored by systematic conformation analysis or by random search using on Monte-Carlo-methods or simulated annealing. There are several (commercial) computer programs available for conducting geometry optimization, starting from atom type and connectivity information only, e.g. CORINA. Often, however, only one conformation is given as output. In the context of QSAR analysis, it is not quite clear which conformation the molecule will take in vivo at the place where the biological activity actually occurs. If several possible conformations are available, the QSAR analysis should make use of them. However, if only one conformation is given for each molecule, at least this conformation should have been obtained by the same software under the same conditions.

5.2 Molecule Kernels

5.2.1 Definition

In the following, the mathematical formalism necessary for defining molecule kernels is introduced.

Definition 1 (**Molecule**). *A molecule is an attributed undirected graph* $\mathcal{M} = (\mathcal{A}, \mathcal{B}, \epsilon, \tau, \boldsymbol{\xi})$, *where*

- $\mathcal{A} = \{A_1, \ldots, A_N\}$ *is a set of vertices called atoms*

- $\mathcal{B} \subseteq \mathcal{A}^{[2]} = \{\{A_i, A_j\} | A_i, A_j \in \mathcal{A}, i \neq j\}$ *a set of edges corresponding to the bonds between the atoms*
- $\epsilon : \mathcal{A} \rightarrow \mathcal{E}$ *a mapping of the atoms to a set of labels* $\mathcal{E}$ *corresponding to the chemical elements*
- $\tau : \mathcal{B} \rightarrow \mathcal{T}$ *a mapping of the bonds to a set of labels* $\mathcal{T}$ *corresponding to bond types*
- $\boldsymbol{\xi} : \mathcal{A} \rightarrow \mathbb{R}^3$ *a mapping of the atoms to their 3D coordinates*

The edges of the graph, the bonds, have labels corresponding to the bond types, which can be represented by integer numbers such that $\mathcal{T} = \{1, 2, 3, 4\}$ (1: a single bond, 2: a double bond, 3: a triple bond, 4: an aromatic bond). The vertices of the graph, the atoms, have labels corresponding to the different chemical elements. Therefore we can chose the set $\mathcal{E}$ to consist of all elemental symbols, $\mathcal{E} = \{H, He, Li, Be, B, C, N, O, \dots\}$. Moreover, the atoms have real valued, three-dimensional attribute vectors, corresponding to the three dimensional atomic coordinates in units of angstroms. Note that the rotation and translation of the whole molecule with respect to the global coordinate system is arbitrary.

A molecule contains both topological information, which is determined by the graph structure, as well as geometrical information, which is determined by the coordinate mapping $\boldsymbol{\xi}$. In a QSAR learning task, we are given a dataset $D = \{(\mathcal{M}_p, t_p), p = 1, \dots, m\}$ consisting of pairs of molecules and target values. The target values t_p can be either real valued or binary class labels $(+1, -1)$. To distinguish between different molecules from the dataset, the constituents of molecule $\mathcal{M}_p$ are denoted by superscript indices, e.g. the n^{th} atom in molecule $\mathcal{M}_p$ is denoted by A_n^p, while its coordinates are denoted by $\boldsymbol{\xi}^p(A_n^p)$.

If we are given the coordinates of two molecules $\mathcal{M}_p$ and $\mathcal{M}_q$ in the same coordinate system, a *molecular alignment* is a rigid body transformation (involving only translations and rotations) of the coordinates of $\mathcal{M}_q$.

Definition 2 (Bipod). *A bipod* B_{ijk}^p *is an ordered triplet of atoms from the same molecule* $\mathcal{M}_p$ *connected by two bonds*

$$B_{ijk}^p = \left(A_i^p, A_j^p, A_k^p\right), \quad with \;\; \{A_i^p, A_j^p\} \in \mathcal{B} \;\; and \;\; \{A_j^p, A_k^p\} \in \mathcal{B}, \tag{5.1}$$

for which the vectors $\boldsymbol{\xi}^p(A_i^p) - \boldsymbol{\xi}^p(A_j^p)$ *and* $\boldsymbol{\xi}^{\boldsymbol{p}}(A_k^p) - \boldsymbol{\xi}^{\boldsymbol{p}}(A_j^p)$ *are not colinear* [2].

Thus a bipod is a V-shaped subfragment of a molecule; see figure 5.1(a) for an example. Note that, in general, $B_{ijk}^p \neq B_{kji}^p$, because the triplets are ordered. The middle atom A_j^p in a bipod B_{ijk}^p will be referred to as the *central bipod atom*.

[2] The requirement of non-co-linearity makes sure that a pair of matching bipods can later be used to define a unique spatial alignment (otherwise the alignment would allow arbitrary rotations around the colinear bipod axes)

Definition 3 (**Matching Bipod Alignment**). *Let θ be a constant. Assume there exists a pair of bipods (B^p_{ijk}, B^q_{rst}) belonging to molecules $\mathcal{M}_p$ and $\mathcal{M}_q$ such that $\epsilon(A^p_i) = \epsilon(A^q_r)$, $\epsilon(A^p_j) = \epsilon(A^q_s)$ and $\epsilon(A^p_k) = \epsilon(A^q_t)$. Moreover assume there is a transformation $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3, \mathbf{x} \mapsto \tilde{\mathbf{x}} = \mathbf{RT}(\mathbf{x})$, such that*

$$\begin{aligned} (\boldsymbol{\xi}^p(A^p_i) &- \mathbf{F}\boldsymbol{\xi}^q(A^q_r))^2 \leq \theta \\ (\boldsymbol{\xi}^p(A^p_j) &- \mathbf{F}\boldsymbol{\xi}^q(A^q_s))^2 \leq \theta \\ (\boldsymbol{\xi}^p(A^p_k) &- \mathbf{F}\boldsymbol{\xi}^q(A^q_t))^2 \leq \theta, \end{aligned} \tag{5.2}$$

where $\mathbf{T} : \mathbb{R}^3 \to \mathbb{R}^3, \mathbf{x} \mapsto \tilde{\mathbf{x}}$ is a translation which superposes the central bipod atoms, i.e.

$$\mathbf{T}\boldsymbol{\xi}^q(A^q_j) = \boldsymbol{\xi}^p(A^p_s), \tag{5.3}$$

*and $\mathbf{R} : \mathbb{R}^3 \to \mathbb{R}^3, \mathbf{x} \mapsto \tilde{\mathbf{x}}$ is the rotation around the superposed central bipod atoms which optimally aligns the bipods in a least-squares sense. Then the transformation $\mathbf{F}$ is called a **matching bipod alignment**. If applied to the coordinates of all atoms in molecule $\mathcal{M}_q$, it uniquely specifies a molecular alignment.*

Thus, for a matching bipod alignment the elements of all corresponding atoms in the bipods have to be the identical and if the bipods are aligned 'optimally' (using a translation and a rotation), the squared Euclidean distances between two corresponding atoms must lie below a threshold θ. The introduction of a threshold accounts for small variability in the geometry optimization and numerical inaccuracies. An illustrative example of a matching bipod alignment is given in Figure 5.1.

Given two ordered bipods, the least squares 3D rotation matrix is uniquely determined. The following proposition shows how it can be calculated. Let us assume that the two central bipod atoms have already be aligned by a simple translation $\mathbf{T}$. To simplify the equations we move to a new coordinate system, which has its origin at the position of the aligned central bipod atoms. Let us denote the matrix of the coordinate vectors of the three atoms in bipod B^q_{rst} in this new system by $\mathbf{X}$, a 3×3-matrix, where the rows denote the coordinates in 3D space and the columns the 3 atoms. Equivalently, let $\mathbf{Y}$ represent the coordinates of bipod B^p_{ijk}. Further, let $\mathbf{UWV}^T$ be the singular value decomposition (SVD) of $\mathbf{YX}^T$, where $\mathbf{W}$ is a 3×3 diagonal matrix containing the (non-negative) singular values of $\mathbf{YX}^T$, and where $\mathbf{U}$ and $\mathbf{V}$ are 3×3 orthogonal matrices.

Proposition 1 (**Optimal Rotation Matrix**). *The rotation matrix $\mathbf{R}$ which corresponds to the optimal alignment of two given bipods B^p_{ijk} and B^q_{rst} in a least-squares-sense can be calculated by*

$$\mathbf{R} = \mathbf{U} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(\mathbf{UV}^T) \end{pmatrix} \mathbf{V}^T. \tag{5.4}$$

Proof. The goal is to find the rotation matrix that minimizes the average squared-distances between corresponding atoms in the two bipods. This means we would like to minimize the following cost function

$$\begin{aligned} S &= \mathrm{Tr}((\mathbf{RX}-\mathbf{Y})^T(\mathbf{RX}-\mathbf{Y})) \\ &= \mathrm{Tr}(\mathbf{X}^T\underbrace{\mathbf{R}^T\mathbf{R}}_{=\mathbf{I}}\mathbf{X} - \mathbf{X}^T\mathbf{R}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{RX} + \mathbf{Y}^T\mathbf{Y}) \\ &= \mathrm{Tr}(\mathbf{X}^T\mathbf{X} - 2\mathbf{X}^T\mathbf{R}^T\mathbf{Y} + \mathbf{Y}^T\mathbf{Y}) \end{aligned} \tag{5.5}$$

As a rotation matrix, $\mathbf{R}$ must be a special orthogonal matrix, which has the properties

$$\mathbf{R}^T\mathbf{R} = \mathbf{I} = \mathbf{R}^{-1}\mathbf{R} = \mathbf{R}\mathbf{R}^T \tag{5.6}$$

$$\det \mathbf{R} = +1. \tag{5.7}$$

Eq.(5.6) gives rise to the set of constraints

$$\mathbf{R}\mathbf{R}^T - \mathbf{I} = \mathbf{0}, \tag{5.8}$$

where $\mathbf{0}$ is a 3×3 matrix of zeros.

The cost function eq.(5.5) together with the constraints (5.8) yields the following Lagrangian,

$$\mathcal{L} = \mathrm{Tr}(\mathbf{X}^T\mathbf{X} - 2\mathbf{X}^T\mathbf{R}^T\mathbf{Y} + \mathbf{Y}^T\mathbf{Y}) + \mathrm{Tr}(\mathbf{\Lambda}(\mathbf{R}\mathbf{R}^T - \mathbf{I})), \tag{5.9}$$

where $\mathbf{\Lambda}$ is the matrix of (unknown) Lagrange multipliers. Note that $\mathbf{\Lambda}$ must be symmetric, since $\mathbf{R}^T\mathbf{R} - \mathbf{I}$ is symmetric.

The Lagrangian $\mathcal{L}$ should be minimized with respect to the elements of $\mathbf{R}$.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{R}} &= -2\frac{\partial \mathrm{Tr}(\mathbf{X}^T\mathbf{R}^T\mathbf{Y})}{\partial \mathbf{R}} + \frac{\partial \mathrm{Tr}(\mathbf{\Lambda}\mathbf{R}\mathbf{R}^T)}{\partial \mathbf{R}} \\ &= -2\mathbf{Y}\mathbf{X}^T + (\mathbf{\Lambda} + \mathbf{\Lambda}^T)\mathbf{R} \\ &= -2\mathbf{Y}\mathbf{X}^T + 2\mathbf{\Lambda}\mathbf{R} = 0, \end{aligned} \tag{5.10}$$

where we made use of the fact that $\mathbf{\Lambda}$ is symmetric. It follows that

$$\mathbf{\Lambda}\mathbf{R} = \mathbf{Y}\mathbf{X}^T. \tag{5.11}$$

This can be solved by singular value decomposition,

$$\mathbf{Y}\mathbf{X}^T = \mathbf{U}\mathbf{W}\mathbf{V}^T, \tag{5.12}$$

where $\mathbf{W}$ is a 3×3 diagonal matrix containing the (non-negative) singular values of $\mathbf{Y}\mathbf{X}^T$, and where $\mathbf{U}$ and $\mathbf{V}$ are 3×3 orthogonal matrices. Now we can determine $\mathbf{\Lambda}$ from eq.(5.11) by using the fact that $\mathbf{R}$ must be orthogonal:

$$\begin{aligned} & (\mathbf{\Lambda}\mathbf{R})(\mathbf{\Lambda}\mathbf{R})^T = (\mathbf{Y}\mathbf{X}^T)(\mathbf{Y}\mathbf{X}^T)^T \\ \Rightarrow\ & \mathbf{\Lambda}\mathbf{R}\mathbf{R}^T\mathbf{\Lambda}^T = (\mathbf{U}\mathbf{W}\mathbf{V}^T)(\mathbf{U}\mathbf{W}\mathbf{V}^T)^T \\ \Rightarrow\ & \mathbf{\Lambda}\mathbf{\Lambda}^T = \mathbf{U}\mathbf{W}\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{U}^T \\ \Rightarrow\ & \mathbf{\Lambda}^2 = \mathbf{U}\mathbf{W}^2\mathbf{U}^T \\ \Rightarrow\ & \mathbf{\Lambda} = \mathbf{U}\mathbf{W}\mathbf{U}^T \end{aligned} \tag{5.13}$$

Inserting this result in eq.(5.11) yields

$$\begin{aligned} & \mathbf{\Lambda R} = \mathbf{Y X}^T \\ \Rightarrow \quad & \mathbf{U W U}^T \mathbf{R} = \mathbf{U W V}^T \\ \Rightarrow \quad & \mathbf{U}^T \mathbf{U W U}^T \mathbf{R} = \mathbf{U}^T \mathbf{U W V}^T \\ \Rightarrow \quad & \mathbf{W^{-1} W U}^T \mathbf{R} = \mathbf{W^{-1} W V}^T \\ \Rightarrow \quad & \mathbf{U}^T \mathbf{R} = \mathbf{V}^T \\ \Rightarrow \quad & \mathbf{U U}^T \mathbf{R} = \mathbf{U V}^T \\ \Rightarrow \quad & \mathbf{R} = \mathbf{U V}^T. \end{aligned} \tag{5.14}$$

With eq.(5.14) we have obtained an expression for optimal rotation matrix $\mathbf{R}$. However, so far the constraint eq.(5.7), $\det \mathbf{R} = +1$, has not been used. Therefore the orthogonal solution could still describe a reflection (for $\det \mathbf{R} = -1$). In order to make sure that in fact a rotation matrix is obtained, the following modification should be used (Challis, 1995),

$$\mathbf{R} = \mathbf{U} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(\mathbf{U V}^T) \end{pmatrix} \mathbf{V}^T, \tag{5.15}$$

which ensures that constraint eq.(5.7) is fulfilled and $\mathbf{R}$ is indeed a rotation matrix. □

Definition 4 (**Set of all Matching Bipod Alignments**). *The set of all pairs of matching bipods from two molecules $\mathcal{M}_p$ and $\mathcal{M}_q$ defines a set of molecular alignments, which is called the set of all matching bipod alignments Ω_{pq}.*

Ω_{pq} is a finite subset of the (infinite) set of all possible molecular alignments. It is special in that bipod alignments correspond to the most general locally matching alignments which define a molecular alignment. Larger structures than bipods (like tripods, rings, or certain subfragments) would also allow a unique 3D orientation, but the possible number of alignments is a much smaller subset of Ω_{pq}. Smaller structures than bipods (i.e., pairs of atoms connected by bonds) do not allow a unique alignment, as they do not fix the relative rotation of the molecules around such a bond.

Definition 5 (**Molecule Kernels**). *Let $w_i \in \Omega_{pq}$ denote the i^{th} bipod alignment from the set of all possible matching bipod alignments between two molecules $\mathcal{M}_p$ and $\mathcal{M}_q$. Consider a similarity function $s(\mathcal{M}_p, \mathcal{M}_q, w_i)$, which assigns a similarity value to the molecules at a specific bipod alignment. Then a molecule kernel between the two molecules can be calculated as*

$$k(\mathcal{M}_p, \mathcal{M}_q) = \max_{w_i \in \Omega_{pq}} s(\mathcal{M}_p, \mathcal{M}_q, w_i). \tag{5.16}$$

A molecule kernel corresponds to the maximum similarity value under all possible matching bipod alignments. Different molecule kernels can be defined, based on different choices of similarity function.

In this chapter, two different molecule kernels (MK1 and MK2) are investigated, which will be defined in the following. Assume we are given two aligned molecules $\mathcal{M}_p$ and $\mathcal{M}_q$, whose alignment is specified by $w_h \in \Omega_{pq}$. Let N_p be the number of atoms in $\mathcal{M}_p$, and N_q be the number of atoms in $\mathcal{M}_q$. Let us further assume that the atomic coordinates of the aligned molecules are given in the common reference frame as $\mathbf{x}_i^p,\ i = 1, \ldots, N_p$ and $\mathbf{x}_j^q,\ j = 1, \ldots, N_q$. Let $\mathcal{N}$ be the set of matching atoms

$$\mathcal{N} = \left\{ A_i^p : \min_{j, \epsilon(A_j^q) = \epsilon(A_i^p)} (\mathbf{x}_i^p - \mathbf{x}_j^q)^2 < \theta,\ i = 1, \ldots, N_p,\ j = 1, \ldots, N_q \right\} \tag{5.17}$$

Let the number of matching atoms N_{pq}^h be defined as the cardinality of the set $\mathcal{N}$ under the matching bipod alignment w_h. Using this number as similarity measure, we obtain a kernel which we denote by MK1:

Definition 6 (**Unnormalized Atomwise Correspondence Kernel (MK1)**)**.**

$$k(\mathcal{M}_p, \mathcal{M}_q) = \max_{w_h \in \Omega_{pq}} N_{pq}^h. \tag{5.18}$$

This kernel can also be normalized to the range of $[0, 1]$, by using the Jaccard index as a similarity function. The Jaccard index is a similarity measure for two sets, in which the size of the intersection divided by the size of the union of the sets. This normalizes the atomwise correspondence, taking the size of both molecules into account, and yields a second molecule kernel, which we denote by MK2:

Definition 7 (**Normalized Atomwise Correspondence Kernel (MK2)**)**.**

$$k(\mathcal{M}_p, \mathcal{M}_q) = \max_{w_h \in \Omega_{pq}} \frac{N_{pq}^h}{N_p + N_q - N_{pq}^h}. \tag{5.19}$$

Note that these kernel functions are symmetric with respect to the interchanging of the two molecules. The above definitions of atomwise correspondence kernels allow for some numerical inaccuracies and a certain variability in the atom's position via the threshold $\theta > 0$. This is the same hyperparameter which was already used in the spatial matching of the bipod atoms. It should be small enough that one atom of molecule $\mathcal{M}_q$ which matches one atom of molecule $\mathcal{M}_p$ is matching only that atom and no others. In our experiments, θ was always fixed at $\theta = 0.25$, which is small enough to ensure a unique assignment of corresponding atoms.

The similarity function determines which general aspects of the molecular representation (e.g., atom types, bonds types, chemophysical properties) are modeled by the kernel. The above-defined atomwise correspondence kernels are rather simple examples of molecule kernels, since they take only the spatial superposition of atoms from the same element into account. However, it does not account for the spatial match of atoms belonging to different elements, nor do the bond types enter the analysis. More elaborate types of molecule kernels can be constructed in the above framework

by using similarity functions which take such information into account. In this context it should be mentioned that the adaptation to the chemical space of a particular class of compounds and a particular endpoint is not handled by a specific choice of molecule kernel, but by the learning machine used for prediction (see section 5.2.3).

5.2.2 Properties

Molecule kernels are symmetric

$$k(B, A) = k(A, B), \tag{5.20}$$

and indefinite. The normalized atomwise correspondence kernel is bounded to lie between zero and one. Molecule kernels are guaranteed to find the largest similarity value in which at least two bipods match, since the alignments of all matching bipods are checked out. The runtime complexity of the molecule kernel on two molecules of the same size n depends on the runtime complexity of the calculation of the similarity function in eq.(5.16).

As an example, for the case of the atomwise correspondence molecule kernels the runtime complexity scales with $O(n^4)$. This can be seen as following: Let us assume the worst case scenario, that all atoms in both molecules are from the same element, and that each atom has a degree of d (i.e. d topological neighbors). Then each of the n atoms from molecule $\mathcal{M}_q$ can be centered on each of the n atoms of molecule $\mathcal{M}_p$, yielding n^2 combinations. For each of the atoms in such pair there are $d(d-1)$ combinations of neighboring atoms which can be used to form ordered bipods. Therefore the number of all possible pairs of matching bipods is $(n^2d^2(d-1)^2)$, which is $O(n^2)$. The search for the nearest neighbors would require n^2 operations, leading to a total runtime complexity of $O(n^4)$. This is better than the time complexity of all-subgraph kernels (Gärtner et al., 2003), whose calculation is NP-hard, and of product (Gärtner et al., 2003) and marginalized (Kashima et al., 2003) graph kernels, whose runtime-complexity scales with n^6. In order to illustrate how the runtime complexity of $O(n^4)$ compares to $O(n^6)$, n^4 an n^6 are plotted as a function of molecule size n in Figure5.3. Note that in real datasets, the two molecules are usually of different size and contain more than one element. Thus, the total number of matching bipods is usually quite small, and the nearest neighbors need only to be evaluated for atoms from the same element. This allows the efficient calculation of this molecule kernel, even for molecules containing several hundreds of atoms.

5.2.3 Model Building and Prediction

Kernel matrices resulting from molecule kernels are not necessarily positive semidefinite, and even if they are on the training set, they might not be on the test set. Therefore, standard kernel methods requiring positive-semidefinite kernel matrices cannot be employed for prediction. However, for the recently proposed (Hochreiter and Obermayer, 2006) potential support vector machine (P-SVM) this restriction does not hold.

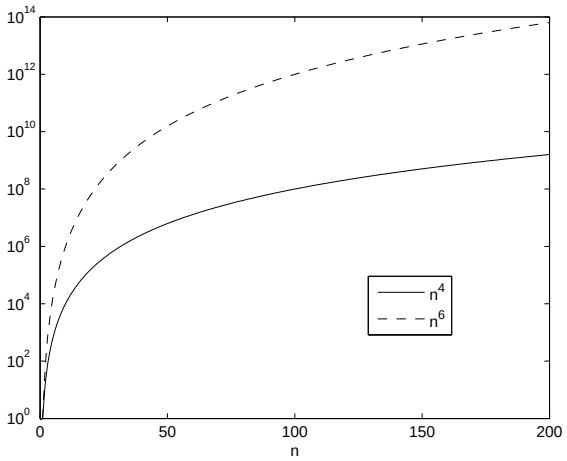


Figure 5.3: Runtime complexity as a function of molecule size
The functions n^6 and n^4, corresponding to the order of the runtime complexity of random walk graph kernels and molecule kernels, respectively, are plotted for growing molecule size n.

The mathematical reason for this is that its (dual) optimization problem only depends on the kernel matrix $\mathbf{K}$ via $\mathbf{K}^T\mathbf{K}$. Therefore, the P-SVM can handle arbitrary kernel matrices, which do not have to be square or positive semidefinite. It can be used for both classification and regression tasks. For the mathematical formulation of the P-SVM see (Hochreiter and Obermayer, 2006). The P-SVM objective function is a convex optimization problem, therefore a global solution exists, which can be found using an efficient sequential minimal optimization (SMO) algorithm[3]. See (Knebel et al., 2008) for details. The result of the optimization is a set of so-called Lagrange-Parameters α_j, one for each training set example, which provide a measure how individual molecules in the training set affect the prediction. The P-SVM usually obtains a sparse solution, which means that many of these α_j values will be zero. Each nonzero α_j corresponds to a molecule, which we will call a "support molecule". The sign of the corresponding α_j serves as class indicator (for classification) or shows whether the respective molecule is associated with increase or decrease in activity (for regression). Its absolute value indicates how relevant a particular molecule is for the prediction (Hochreiter and Obermayer, 2006).

The process of model building and prediction using molecule kernels and the P-SVM is illustrated in figure 5.4. First, the molecule kernel matrix $\mathbf{K}$ on the training set is calculated by evaluating the molecule kernel $k(\mathcal{M}_p, \mathcal{M}_q)$ for all pairs of compounds

[3]The P-SVM software is available under the GNU General Public License from the Neural Information Processing Group at the Berlin Institute of Technology (`http://ni.cs.tu-berlin.de/software/psvm/index.html`)

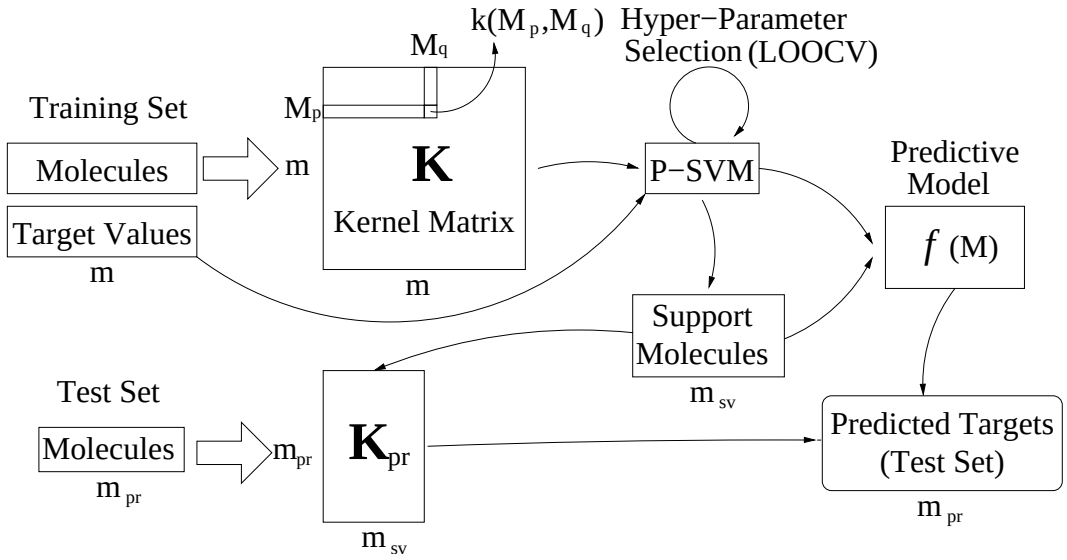


Figure 5.4: Process of model building and prediction using molecule kernels

$\mathcal{M}_p$ and $\mathcal{M}_q$ from the training set. For a training set containing m compounds, this corresponds to a $m \times m$ symmetric matrix with ones on the diagonal, which requires the calculation of $m \cdot (m-1)/2$ scalar kernel values. For the calculation of the kernel matrix, only the molecular structures and not the activity values (or class labels) are needed.

In the next step, a learning machine is trained using both the calculated kernel matrix $\mathbf{K}$ and the activity values (or class labels) in order to build a model. The P-SVM is used as learning machine, which requires the selection of two hyperparameters (C and ϵ). This is done by minimizing the leave-one-out cross-validation (LOOCV) prediction error on the training set over a discrete grid of hyperparameter values. Then the P-SVM is trained at the optimal hyperparameters using the full training set, which yields the final model. The prediction function for a molecule $\mathcal{M}$ is specified via the set of α values and an offset b. It takes the form

$$f(\mathcal{M}) = \sum_{j=1}^{m} \alpha_j \; k(\mathcal{M}_j, \mathcal{M}) + \; b \tag{5.21}$$

for regression and

$$f(\mathcal{M}) = \text{sign}\left(\sum_{j=1}^{m} \alpha_j \; k(\mathcal{M}_j, \mathcal{M}) + \; b \right) \tag{5.22}$$

for classification. Note that only the molecules $\mathcal{M}_j$ corresponding to nonzero α_j, the support molecules, are needed for prediction.

This model can then be used for prediction in the following way: First, the kernel matrix $\mathbf{K}_{pr}$ is calculated between the set of m_{sv} support molecules and the set of m_{pr} molecules for which we wish to predict the activity (or the class label). This is done by evaluating the molecule kernel for all pairs involving a member from each of these sets. The calculation of the $m_{sv} \times m_{pr}$ kernel matrix $\mathbf{K}_{pr}$ thus requires $m_{sv} \cdot m_{pr}$ kernel evaluations. In a regression setting, the P-SVM then yields the m_{pr} unknown activity values via eq.(5.21). In a classification setting, predictions of class labels are obtained via eq.(5.22).

While the choice of kernel influences which general properties of molecules are considered for evaluating the similarity, the adaptation to the chemical space spanned by the compounds in the training set and to a particular endpoint is handled by the P-SVM. Intuitively, this works as following: By generating a linear combination of the kernel values of all support molecules with the respective test molecule, the presence or absence of certain 3D substructures of the support molecules is used to assign a regression value or class label. However, this is done not explicitly, but implicitly by the support-vector machine based on the calculated kernel matrix.

5.3 Explanatory QSAR Models from Molecule Kernels

Descriptor-based QSAR methods encode the molecular properties via a large number of rather diverse descriptor variables, which usually involve a high level of abstraction from the original structure. Usually, these descriptor vectors are embedded into a Euclidean space in which the QSAR models are build. Since the descriptors themselves are very heterogeneous, it is not quite clear how the distances in this space can be interpreted. During the model building process the descriptors are combined in a way which even further obscures the interpretation of the models. For example, if the partial least squares method (Wold et al., 1984, 2001) is used, the prediction is based on a linear combination of the original descriptors. In general, the models based on descriptor vectors allow to make predictions, but are difficult to interpret and do not give insight into the mechanisms underlying the activity.

In contrast to this, the prediction function of the molecule kernel method is expressed as a linear combination of structural similarities to a set of support molecules. These structural similarities can be used together with the corresponding Lagrange multipliers to gain insight into the structural properties which influence the activity.

In this section, a method for building an explanatory QSAR model is suggested, which is based on the molecule kernel method. It allows to calculate the contribution of each single atom to the activity predicted for a molecule. Moreover, a method for 3D visualization of these contributions is suggested.

5.3.1 Building an Explanatory Model

During the calculation of the molecule kernels defined in section 5.2.1, the optimal matching bipod alignment $w_{opt} \in \Omega$ is determined as the one in which the number

of matching atoms N_{pq}^h is maximized. The corresponding set of matching atoms from both molecules is given according to eq.(5.17) as

$$\mathcal{N}(w_{opt}) = \left\{ A_i^p : \min_{j,\epsilon(A_j^q)=\epsilon(A_i^p)} (\mathbf{x}_i^p - \mathbf{x}_j^q)^2 < \theta,\ i = 1,\ldots,N_p,\ j = 1,\ldots,N_q \right\}. \quad (5.23)$$

The cardinality of this set is used in the calculation of the molecule kernel $k(\mathcal{M}_p, \mathcal{M}_q)$. However, this set can further be used in the construction of the explanatory model, as will be explained in the following.

Consider a molecule $\mathcal{M}$ with N atoms. Assume that the set $\mathcal{N}(w_{opt})$ in eq.(5.23) was constructed for every combination of molecule $\mathcal{M}$ with one of the support molecules $\mathcal{M}_p, p = 1,\ldots,m_{sv}$ during the calculation of the corresponding kernel values. These sets are now used to construct for each support molecule a binary vector $\mathbf{m_p}$ of length N, in which the matching atoms are assigned the value 1, and the non-matching atoms the value 0. The resulting vectors $\{\mathbf{m}_p\}, p = 1,\ldots,m_{sv}$ can be used together with the values of the Lagrange multipliers $\{\alpha_p\}, p = 1,\ldots,m_{sv}$ to construct an *influence vector* $\mathbf{s}$, which reflects the contribution of the individual atoms in $\mathcal{M}$ to the prediction. In the case of regression analysis, this will be the influence on the regression value, for classification, this will be the value of the prediction function before the application of the *sign*-function.

In order to construct the vector $\mathbf{s}$, it needs to be taken into account that the prediction function of the P-SVM is actually optimized in a rescaled space, where each row of the kernel matrix on the training set has zero mean and variance one. This rescaling is always possible, and leads to several simplifications in the derivation of the P-SVM. However, one has to keep in mind that the Lagrange multipliers returned by the P-SVM were calculated with respect to the standardized kernel matrix. But for the calculation of the influence vector $\mathbf{s}$, the regression parameters with respect to the original kernel matrix are needed. Therefore, the regression parameters $\tilde{\alpha}_j$ and $\tilde{b}$ obtained by the P-SVM in the form of Lagrange multipliers and offset must be transformed into the space of the original kernel matrix, i.e. into the parameters of eqs. (5.21) and (5.22).

Let

$$\mu_p = \frac{1}{m} \sum_{q=1}^{m} k(\mathcal{M}_p, \mathcal{M}_q) \quad (5.24)$$

and

$$\sigma_p = \sqrt{\frac{1}{m} \sum_{q=1}^{m} (k(\mathcal{M}_p, \mathcal{M}_q) \ - \ \mu_p)^2}, \quad (5.25)$$

be the mean and standard deviation for each column p of the kernel matrix.

Then the prediction function becomes

$$f(\mathcal{M}) = \sum_{p=1}^{m} \tilde{\alpha}_p \frac{(k(\mathcal{M}_p, \mathcal{M}) - \mu_p)}{\sigma_p} + \tilde{b} \tag{5.26}$$

$$= \sum_{p=1}^{m} \underbrace{\frac{\tilde{\alpha}_p}{\sigma_p}}_{\alpha_p:=} k(\mathcal{M}_p, \mathcal{M}) + \underbrace{\left(- \sum_{p=1}^{m} \frac{\tilde{\alpha}_p}{\sigma_p} \mu_p + \tilde{b} \right)}_{b:=} \tag{5.27}$$

for regression and

$$f(\mathcal{M}) = \text{sign} \left(\sum_{p=1}^{m} \tilde{\alpha}_p \frac{(k(\mathcal{M}_p, \mathcal{M}) - \mu_p)}{\sigma_p} + \tilde{b} \right) \tag{5.28}$$

$$= \text{sign} \left(\sum_{p=1}^{m} \underbrace{\frac{\tilde{\alpha}_p}{\sigma_p}}_{\alpha_p:=} k(\mathcal{M}_p, \mathcal{M}) + \underbrace{\left(- \sum_{p=1}^{m} \frac{\tilde{\alpha}_p}{\sigma_p} \mu_p + \tilde{b} \right)}_{b:=} \right) \tag{5.29}$$

for classification.

Thus the values of α_j in eqs.(5.21) and (5.22) can be obtained from the values of $\tilde{\alpha}_j$ in eqs.(5.26) and (5.28) through division by σ_j,

$$\alpha_p = \frac{\tilde{\alpha}_p}{\sigma_p}. \tag{5.30}$$

Using this, the influence vector $\mathbf{s}$, which reflects the individual contribution of each atom in M to the predicted value, can be calculated. For the kernel MK1, it is given by

$$\mathbf{s}^{(1)} = \sum_{p=1}^{m} \frac{\tilde{\alpha}_p}{\sigma_p} \mathbf{m}_p. \tag{5.31}$$

For the kernel MK2, the normalization factor to the number of atoms of both molecules has to be additionally taken into account, which yields

$$\mathbf{s}^{(2)} = \sum_{p=1}^{m} \frac{\tilde{\alpha}_p}{\sigma_p} \frac{1}{N_p + N - k_1(\mathcal{M}_p, \mathcal{M})} \mathbf{m}_p, \tag{5.32}$$

where N_p is the number atoms in support molecule M_p and k_1 is the molecule kernel MK1.

By summing the components of the influence vector $\mathbf{s}$ and adding an offset which does not depend on the molecule $\mathcal{M}$, the prediction for the activity is now obtained as

$$f(\mathcal{M}) = \sum_{j=1}^{N} s_j^{(i)} + (\tilde{b} - \sum_{p=1}^{m} \frac{\mu_p \tilde{\alpha}_p}{\sigma_p}), \; i = 1, 2 \tag{5.33}$$

for regression and as

$$f(\mathcal{M}) = \text{sign}\left(\sum_{j=1}^{N} s_j^{(i)} + \ (\tilde{b} - \sum_{p=1}^{m} \frac{\mu_p \tilde{\alpha}_p}{\sigma_p})\right), \ i = 1, 2 \tag{5.34}$$

for classification. Note that the value for the prediction resulting from eqs.(5.33) and (5.34) is the same as the one obtained by calculating eqs.(5.21) and (5.22), which use the kernel values instead of the influence vector. However, the proposed framework of influence vectors allows direct access to the contributions of the single atoms, which can be used for interpreting the resulting prediction model.

5.3.2 Visualization

The explanatory model described in section 5.3.1 yields an influence vector, which contains the contribution of each atom to the activity prediction. In the following, a visualization technique is proposed which allows an easier interpretation of the influence values.

For this purpose, the 3D structure of the molecule is visualized as a balls-and-sticks model, in which the atoms correspond to spheres, and the bonds to lines connecting the spheres. The components of the influence vector are then represented by the colors of the individual atoms. In order to allow an easy comparison of the results obtained for different molecules, the following color coding scheme is suggested: the colorbar follows a prism spectrum, where zero influence corresponds to a shade of green, and where high positive influences are represented by red colors and high negative influences by blue colors. The color coding is symmetric with respect to zero, and its range is normalized by the largest absolute influence, such that the whole color range is used for each molecule. Therefore, the colors provide a relative measure of influence for comparing the atoms within each individual molecule. Thus, the same color might correspond to different values of influence for different molecules, but the sign of the influence can always be read off by the color. The suggested visualization method was implemented in MATLAB and is illustrated in figure 5.5 on a molecule from the AChE training dataset (see section 5.4).

The top figure visualizes the influences obtained with molecule kernel MK1, while the bottom figure depicts the influences obtained with MK2. For both kernels an oxygen atom with double bond (dark red) and one hydrogen atom (on the left) have the highest positive influence on the activity prediction. A large part of the molecule has a slightly positive influence (orange), some of the atoms do not contribute (green). There are only slight differences between the models obtained by the two molecule kernels.

The explanatory model and its visualization can be used in two ways: First it can be applied on a labeled training dataset, in order to discover likely candidates for pharmacophores, i.e. an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and

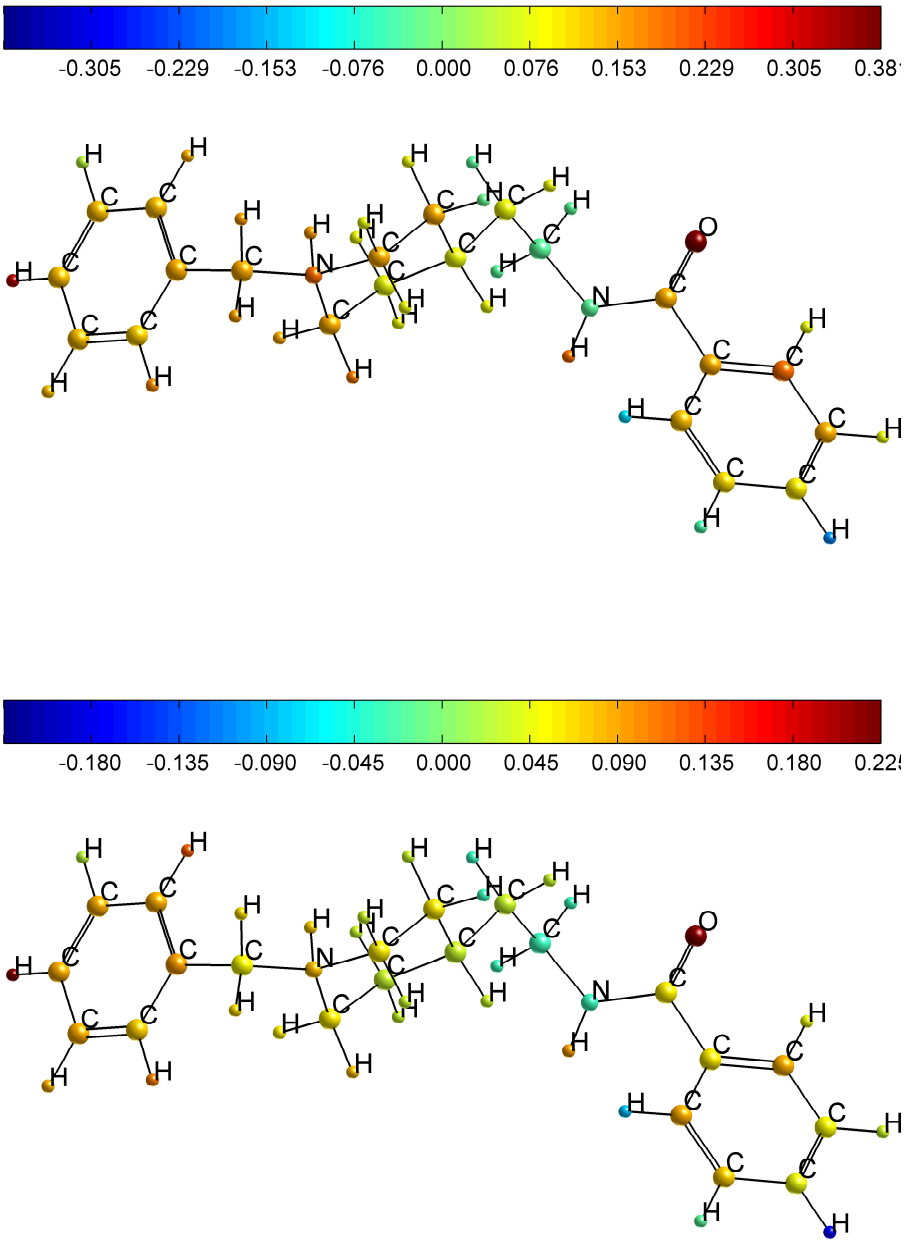


Figure 5.5: Visualization of the explanatory QSAR model obtained by the molecule kernel method for MK1 (top) and MK2 (bottom) on the same molecule.

to trigger (or block) its biological response. On the other hand, it can be employed to show the structural features responsible for the prediction of unlabeled molecules. Thus the proposed molecule kernel method not only provides a predictive model, it can also explain the prediction of a particular model in terms of structural features.

5.4 Application to QSAR Analysis

5.4.1 QSAR Datasets

A crucial requirement for comparing QSAR methods is the availability of enough and suitable data. This makes sure that in the selected training sets the molecular structures and activities do not deviate too far from the test set. This requires also some redundancy in the training data, such that the relevant patterns in the data can be recognized on the training data and used by the QSAR method (learning machine). It is usually assumed that the data points are independent and identically distributed (i.i.d.) samples from an underlying population distribution. The dataset must be large enough that both training and test set are likely to be representative for this distribution.

In the following, the QSAR datasets used in the method comparison will be described. These specific datasets have been chosen for the following reasons: (1) They are large enough to allow a comparison of methods, (2) they are publicly available[4] in structural form (sd files), (3) there are results reported in the literature for specific training and test sets, and (4) the activity values of the training and test sets have similar enough distributions to be representative for the population distribution. To check that the latter holds for the activity values of the regression datasets, the histograms of training and test set were compared, and it was verified that the test set did not contain activities in a range which is not, or insufficiently, represented by the training set.

Details about the mean and standard deviation of the number of atom and bonds in each training and test set are given in Table 5.1. The binding affinity for the regression datasets is measured by pIC50 values, which correspond to the negative logarithm of the inhibitor concentration giving a 50% reduction of specific binding.

Fontaine Dataset (Classification)

This dataset contains 435 molecules of the benzamidine family which is dichotomized into factor Xa inhibitors of low and high affinity. The dataset contains 156 compounds with low activity and 279 compounds with high activity. It was used (Fontaine et al., 2005) for 3D QSAR modeling with the Anchor-GRIND method. There, the dataset was randomly divided into a training set (290 compounds, 99 inactive, 191 active) and a test set (145 compounds, 57 inactive, 88 active) (Fontaine et al., 2005).

[4] All the used datasets are publicly available at `http://www.cheminformatics.org/datasets/`

Dataset	Nr. of atoms (training set)	Nr. of bonds (training set)	Nr. of atoms (test set)	Nr. of bonds (test set)
Fontaine	59.7 ± 11.1	62.5 ± 11.6	60.0 ± 11.1	62.9 ± 11.7
ACE	43.6 ± 18.7	44.2 ± 19.6	40.1 ± 17.9	40.8 ± 18.8
AChE	56.4 ± 5.4	58.9 ± 5.7	56.4 ± 6.3	59.2 ± 6.8
BZR	36.3± 5.8	38.8 ± 6.2	36.4 ± 5.1	38.8 ± 5.4
DHFR	40.7 ± 6.9	42.8 ± 7.1	40.4 ± 7.7	42.7 ± 7.9

Table 5.1: Statistics on datasets

ACE Dataset (Regression)

The ACE dataset comprises a set of 114 angiotensin converting enzyme (ACE) inhibitors was originally published by Depriest et al. (Depriest et al., 1993). The activities range from pIC50 values of 2.1 to values of 9.9. It was used by Sutherland et al. (Sutherland et al., 2004) to compare a large variety of QSAR methods, where it was split into a trainings set containing 76 compounds and a test set with 38 compound.

AChE Dataset (Regression)

The AChE dataset contains 111 acetylcholinesterase (AChE) inhibitors whose pIC50 values lie in the range between 4.3 and 9.5. It has been assembled by Sutherland et al. (Sutherland et al., 2004), who used it in their QSAR method comparison and divided into a trainings set (74 compounds) and a test set (37 compounds).

BZR Dataset (Regression)

The BZR dataset consist of 163 benzodiazepine receptor ligands, whose pIC50 values lie in the range between 5.5 and 8.9. A subset of it was used in (Sutherland et al., 2004) and subdivided into a training set of 98 compounds and a test set with 49 substances.

DHFR Dataset (Regression)

The DHFR dataset consists of 397 dihydrofolate reductase inhibitors (DHFR) with pIC50 values for rat liver enzyme ranging from 3.3 to 9.8. It has been compiled by Sutherland et al. (Sutherland et al., 2004). From the original data, a training set of 237 compounds and a test set of 124 compounds were generated (and a further set of 36 inactive compounds which were not used here).

5.4.2 Assessment of Generalization Performance

All methods used the same training-test set split for evaluating the predictive performance. The data points from the test set were not used at all for model building. This

includes the choice of hyperparameters (parameters which are considered fixed during the learning of the other model parameters, e.g. parameters representing the model complexity or the smoothness of a function), feature selection, feature construction, model selection or parameter fitting. A measure of the generalization performance of the build models was obtained by applying the method to the test set.

The measures of performance for classification problems are based on the numbers of false positives (FP), true positives (TP), false negatives (FN), true negatives (TN), size of positive class (PC) and size of negative class (NC). From these, sensitivity (TP /PC), specificity (TN /NC), balanced error rate [0.5 (FN/PC+FP/NC)] and concordance (TN+TP)/(PC+NC) can be calculated. However, for the method comparison on the Fontaine dataset, only concordance can be used, since only this result is stated in the literature (Fontaine et al., 2005).

For regression problems, the mean squared error (MSE) can be used, which is the average PRESS (predictive sum of errors), defined as

$$PRESS = \sum_{i=1}^{N_{test}} (y^{(i)} - y_{pred}(x^{(i)}))^2. \tag{5.35}$$

In QSAR studies, usually r^2_{pred} is used instead,

$$r^2_{pred} = \frac{SD - PRESS}{SD}, \tag{5.36}$$

where SD is defined as

$$SD = \sum_{i=1}^{N_{test}} (y^i - \bar{y})^2, \tag{5.37}$$

with $\bar{y}$ denoting the average value of y. There are two different conventions in the QSAR literature with respect to the data points used in this average. In one, $\bar{y}$ is the average over the target values of the training set. In the other, $\bar{y}$ is the average over the target values of the test set. In order to allow a comparison of results, we here stick to the first convention which was used in (Sutherland et al., 2004).

5.4.3 Results

All datasets were analyzed using the two molecule kernels defined in eqs.(5.18) (MK1) and (5.19) (MK2). The model building process was carried out as described in section 5.2.3, using the P-SVM as predictor for both classification and regression. The P-SVM implementation by (Knebel et al., 2008) was used in the analysis, which is available under the GNU General Public License from the Neural Information Processing Group at the Berlin Institute of Technology. For the proposed method, there were two hyperparameters of the P-SVM which needed adjustment. The hyperparameter search was carried out systematically via a two-step grid search, using leave-one-out cross-validation on the training set, which is described in the appendix. In the following, this whole model building method is referred to as the "molecule kernel method".

Results of the Molecule Kernel Method

For the classification dataset, the confusion matrix for the prediction of the trained model on the test set is shown in Table 5.2 (MK1) and Table 5.3 (MK2). Table 5.4 lists the resulting performance statistics. For the regression datasets, the fitted predictions for both training and test set are plotted versus the actual activities as scatter diagrams in figure 5.6. As performance statistics, the values for MSE, PRESS, SD and r^2_{pred} are listed in Table 5.5.

		Actual Label		
		+1	-1	$\sum$
Predicted	+1	TP: 87	FP: 6	93
Label	-1	FN: 1	TN: 51	52
	$\sum$	PC: 88	NC: 57	N:145

Table 5.2: Fontaine dataset: Confusion matrix of the molecule kernel method using MK1 on the test set

		Actual Label		
		+1	-1	$\sum$
Predicted	+1	TP: 87	FP: 7	94
Label	-1	FN: 1	TN: 50	51
	$\sum$	PC: 88	NC: 57	N:145

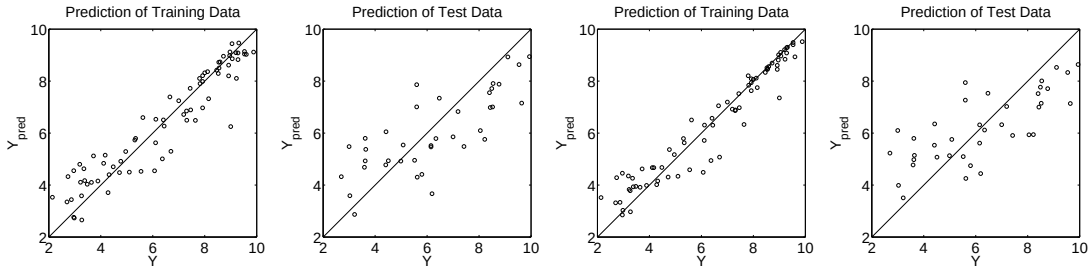
Table 5.3: Fontaine dataset: Confusion matrix of the molecule kernel method using MK2 on the test set

	MK1	MK2
Sensitivity [TP /PC]:	0.987	0.989
Specificity [TN /NC]:	0.895	0.877
Balanced error rate [0.5*(FN/PC+FP/NC)]:	0.067	0.058
Concordance [(TN+TP)/N]:	0.945	0.952

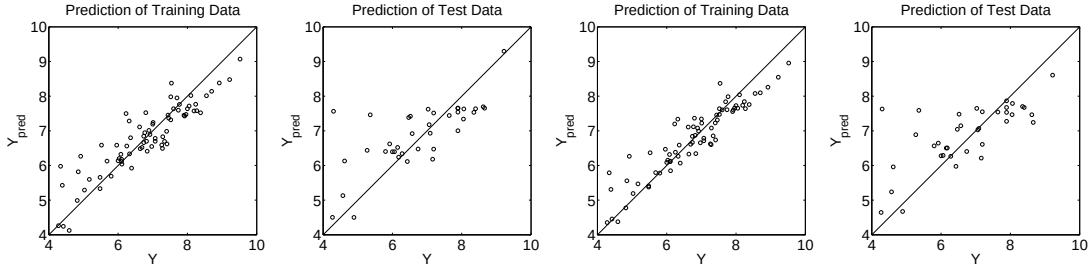
Table 5.4: Fontaine dataset: Predictive performance measures of the molecule kernel method on the test set

	Dataset	ACE	AChE	BZR	DHFR
MK1	MSE	1.906	0.792	0.614	0.687
	PRESS	72.441	29.286	30.062	85.173
	SD	171.878	61.677	45.329	235.561
	r^2_{pred}	0.579	0.525	0.337	0.638
	Dataset	ACE	AChE	BZR	DHFR
MK2	MSE	2.049	0.863	0.591	0.658
	PRESS	77.862	31.877	28.965	81.536
	SD	171.878	61.677	45.329	235.561
	r^2_{pred}	0.547	0.483	0.361	0.654

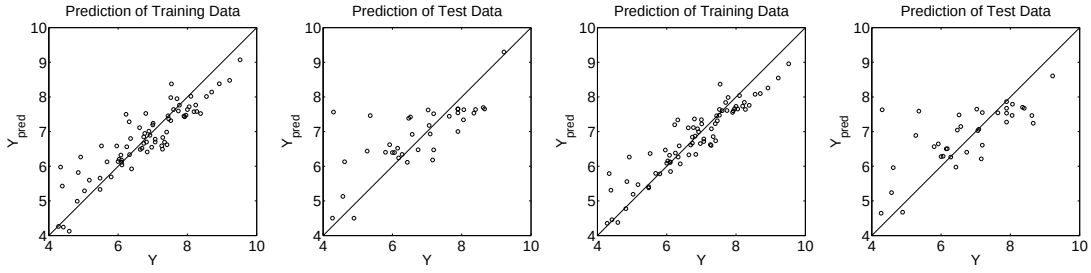
Table 5.5: Regression datasets: Predictive performance measures of molecule kernel methods MK1 and MK2 on test set.



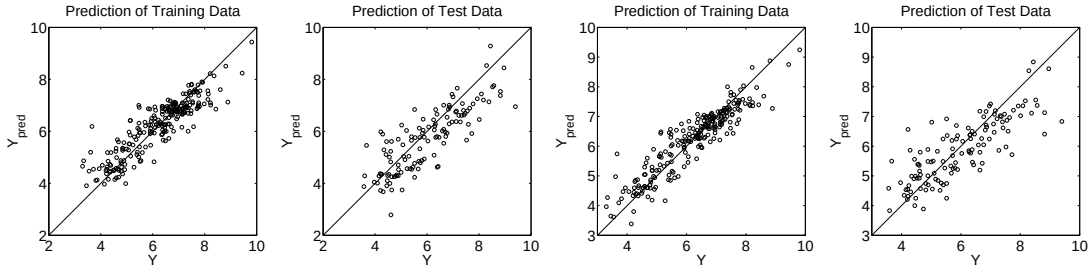
(a) ACE dataset



(b) AChE dataset



(c) BZR dataset



(d) DHFR dataset

Figure 5.6: Application of the molecule kernel method on the regression datasets: Scatter-plot, showing the value of the regression target Y_{pred} predicted by the model against the true value Y. Left: MK1 (training and test set), right: MK2 (training and test set).

Comparison to other QSAR methods

The above results were compared to the results of other QSAR methods taken from the literature. Exactly the same trainings and test sets were used by us and in the respective publications. For the classification dataset (Fontaine dataset), results were available for two variants of the Anchor-GRIND method, the one-block version (using Anchor-MIF descriptors) and the two-block version (using both Anchor-MIF and MIF-MIF descriptors). For details on the methods and their application to the dataset, see the publication by Fontaine et al. (Fontaine et al., 2005) and the references therein. For the regression datasets (ACE, AChE, DHFR) results were available for descriptors calculated with CoMFA, CoMSIA basic (with steric and electrostatic fields), CoMSIA extra (with additional hydrogen-bonding fields, hydrophobic fields, or both), EVA, HQSAR, and traditional 2D and 2.5D descriptors[5] using PLS as predictor. For details on the various methods and their application to the datasets, see the paper by Sutherland at al. (Sutherland et al., 2004) and the references therein. For the comparison of methods on the classification dataset, the correct classification rate (concordance) on the test set is given for each method in Table 5.6. The molecule kernel method achieves 5% prediction error rate which is much lower than the prediction error rates of the Anchor-GRIND methods (16% for the two-block model and 12% for the one-block model). For the regression datasets, the value of r^2_{pred} is given for all compared

Dataset Fontaine	Molecule Kernel MK1	Molecule Kernel MK2	Anchor-GRIND (2 block model)	Anchor-GRIND (1 block model)
Concordance	**0.95**	**0.95**	0.84	0.88

Table 5.6: Classification dataset: The correct classification rate (concordance) is given for different methods. The results for the Anchor-GRIND method are taken from the literature (Fontaine et al., 2005). The best result is printed bold font.

methods in Table 5.7. On the ACE dataset, the molecule kernel method MK1 performs best ($r^2_{pred} = 0.58$), followed by MK2 ($r^2_{pred} = 0.55$). All other QSAR methods except EVA and HQSAR reach r^2_{pred} values around 0.5. On the AChE dataset MK1 gives the best result (r^2_{pred} of 0.50), MK1 scores second best ($r^2_{pred} = 0.48$). Also good results are achieved by CoMFA ($r^2_{pred} = 0.47$), and the two CoMSIA approaches ($r^2_{pred} = 0.44$). Also on the BZR dataset, the two molecule methods show the best performance, $r^2_{pred} = 0.36$ (MK2) and ($r^2_{pred} = 0.34$ (MK1), while all other method yield values of ($r^2_{pred} \leq 0.2$). On the DHFR dataset, the molecule kernel methods perform best with an r^2_{pred} of 0.65 (MK2) and 0.64 (MK1). They are followed by HQSAR ($r^2_{pred} = 0.63$), CoMFA ($r^2_{pred} = 0.59$) and EVA($r^2_{pred} = 0.57$).

[5]Sutherland et al. (Sutherland et al., 2004) use the term 2.5D descriptors to distinguish descriptors which involve straightforward calculations like molecular volume from descriptors based on force-field calculations.

Dataset	MK1	MK2	CoMFA	CoMSIA basic	CoMSIA extra	EVA	HQSAR	2D	2.5D
ACE	**0.58**	***0.55***	0.49	0.52	0.49	0.36	0.30	0.47	0.51
AChE	**0.50**	***0.48***	0.47	0.44	0.44	0.28	0.37	0.16	0.16
BZR	***0.34***	**0.36**	0.00	0.08	0.12	0.16	0.17	0.14	0.20
DHFR	***0.64***	**0.65**	0.59	0.52	0.53	0.57	0.63	0.47	0.49

Table 5.7: Regression datasets: r^2_{pred} for different datasets and methods. All results except for the two molecule kernels are taken from the literature (Sutherland et al., 2004). The best result for each dataset is printed bold font, the second best in bold italics.

5.5 Application to Genotoxicity Prediction

During drug development, not only the potency of a pharmaceutical drug but also its potential genotoxicity has to be taken into account. The gene-damaging potential of a drug needs to be assessed as early in the process as possible, since later failure to pass a regulatory test is very costly for the pharmaceutical company. Therefore, it is advantageous to conduct two *in vitro* tests in an early stage of drug development: The Ames test, which is a bacterial mutagenicity test, and the chromosome aberration (CA) test, which is based on a mammalian cell assay in which the chromosome-damaging potential of chemical compounds is assessed by detecting a microscopically visible formation of aberrant chromosomes. However, the use of the CA test in early discovery stages is rather restricted by costs and compound availability (Rothfuss et al., 2006). Machine learning approaches for toxicity prediction, called *in silico* methods, offer an alternative for high-throughput drug screening. Most machine-learning approaches have focused on the task of predicting the outcome of the Ames test, where relatively good predictive accuracies ($> 70\%$) have been reached (White et al., 2003). However, for the CA test, no similar prediction performance has been achieved (Rothfuss et al., 2006).

One reason for this is that in contrast to QSAR datasets, which usually contain a set of so-called congeneric molecules, i.e. compounds that belong to a well-defined family with a common mechanism, typical toxicity databases span a wide set of substances with various underlying mechanisms (Klopman and Rosenkranz, 1994; Rothfuss et al., 2006). Structural damage can either be caused by direct drug-DNA interactions (e.g. via an electrophilic site which either exists somewhere in the molecule or is produced by its metabolism), as a result of incorrect DNA repair processes or by the interaction of drugs with enzymes involved in DNA replication and transcription. Moreover, the number of chromosomes might change via the interaction with cellular proteins involved in chromosome segregation. Finally, also certain environmental influences (excessive cytotoxicity, pH and temperature) might cause (non-physiological) structural chromosome aberrations in the cell culture.

Another reason is that public CA test data is only scarcely available and the test is not well standardized (Rothfuss et al., 2006), i.e. different cells from different species are used, usually no aberration frequencies are given, and the outcome of the test is judged rather subjectively. This, and the already mentioned non-physiological chromosomal damage most likely lead to a lot of label noise among different data collections.

For these reasons, routinely conducted *in silico* mutagenicity prediction is usually just based on the prediction of the outcome of the Ames test. Therefore, an extension of *in silico* procedures by a prediction of the CA-test results would allow for a more extensive judgment of potential genotoxic effects of candidate substances in early drug development. In this chapter, the molecule kernel method is applied to the problem of CA test prediction.

5.5.1 Chromosome Aberration Dataset

This project was carried out in cooperation with Bayer-Schering Pharma, who compiled a CA dataset from publicly available datasets. Details about the included datasets are listed in Table 5.8. The goal of this study was to apply the molecule kernel method to this data and to compare its performance to the model developed by Bayer-Schering Pharma (Sutter, 2008) using MCASE (MCASE, Beachwood, USA).

Dataset	Description	Experimental System	Neg.	Pos.
(Kirkland et al., 2005)	450 industrial, environmental, and pharmaceutical compounds	CA test on human lymphocytes and cell lines	168	282
(Snyder and Green, 2001)	229 marketed pharmaceuticals	CA test on Chinese hamster ovary cells, Chinese hamster lung cells, V79 cells, MCL-5 human lymphoblastoid cells and human blood peripheral lymphocytes	189	40
(Serra et al., 2003)	383 organic compounds including known carcinogens, drugs, food additives, agrochemicals, cosmetic materials, medicinal products and household materials	CA test on Chinese hamster lung cells	271	112

Table 5.8: Public CA datasets which were combined for the analysis.

MCASE (Klopman and Rosenkranz, 1994) is based on finding topological fragments, ranging from 2 to 10 atoms in length, in combination with 2D distances between atoms, which are statistically correlated with activity (biophores) and inactivity (biophobes), respectively. In addition, the program detects fragments that act as modulators of activity and takes some calculated physicochemical descriptors into account. These can

be used to explain why some of the molecules containing a biophore are not active. A limitation of MCASE is that compounds containing ions, molecular clusters (such as hydrates), and rare atoms (such as Mn, Ca, or K) are not accepted for model generation. Consequently, compounds containing such structural features were automatically eliminated from the training set by the program during model construction. In order to allow a comparison of results between MCASE and the molecule kernel method, the corresponding compounds were also not used for the molecule kernel method, so that the training and test sets were identical. After cleaning and removal of duplicates N = 819 compounds remained, 324 positive ones and 495 negative ones. On this dataset, the classification performance of the methods were compared using 7-fold cross validation. The division into cross-validation folds was the same for all methods.

Ideally, both sensitivity and specificity should be about equal, and as high as possible. A high sensitivity makes sure that possibly genotoxic substances are identified as early as possible in the drug development process. A high specificity is desirable in order to avoid ruling out potential drug candidates by wrongly classifying them as genotoxic.

Since the dataset is rather unbalanced, containing more negative examples than positive ones, application of the P-SVM is expected to result in a high specificity but low sensitivity. This is due to the fact that the P-SVM cost function aims at minimizing the total mean squared deviation of the predictions from the true class labels (+1 and -1) on the training set. A possible solution would be the removal of a sufficient number of compounds from the negative class to balance the dataset. However, this would result in different training and test sets for both methods, and make the results depend on the random subsample which was removed. Therefore, in addition to the analysis using the P-SVM, the molecule kernel matrix was interpreted as vectorial data. This was done by treating the rows of the kernel matrix as feature vectors, each feature measuring the similarity of a molecule to a specific molecule in the training set. In this framework, a C-SVM with RBF kernel was applied using the LIBSVM implementation (Chang and Lin, 2001), which allows assigning different C-values to the two classes to account for the unbalancedness of the dataset. Note that the resulting kernel matrix of the RBF kernel does not represent a kernel between two different molecules anymore, since it depends on the set of molecules in the training set. Instead, it should be considered as a nonlinear transformation used for prediction which is based on the molecule kernel matrix.

5.5.2 Results

The results obtained with the MCASE method and the molecule kernel method under the 7-fold cross-validation are shown in Table 5.9. There are two different MCASE models. The first one (unambiguous) only classifies compounds on the test set where all fragments in the structure are known from the training set, and in which either strong biophores (i.e. pharmacophores or toxicophores) were found, or no biophores were found. The second MCASE model (ambiguous) provides also predictions for some

structures in which the found biophores were not strong (only a weak majority of compounds containing the fragment were positive), or in which no known biophores were found, but some fragments were not present in the training set. Both molecule kernels MK1 and MK2 were used once in combination with the P-SVM and once in combination with the C-SVM. Table 5.9 list the obtained values for sensitivity (TP/PC), Specificity (TN/NC), correct classification rate (TN+TP/(PC+NC)) and coverage (percentage of test set for which a prediction can be made), where TP is the number of true positives, TN the number of true negatives, PC the size of the positive class and NC the size of the negative class.

The average coverage of the unambiguous MCASE model is only about 3/4 of the test set, whereas the ambiguous MCASE model achieves an average coverage of about 98%, which is close to the 100% coverage of the molecule kernel methods; thus the ambiguous model is more suitable for the comparison. However, the prediction performance of the ambiguous MCASE model is only slightly worse than that of the unambiguous one.

For both molecule kernels (MK1 and MK2) the specificity and sensitivity obtained with the P-SVM were better then the results obtained for the two MCASE models, while the coverage was always 100%. As expected, the specificity obtained woth the P-SVM was much higher than the sensitivity, since the data was unbalanced. The C-SVM accounted for unbalanced classes, and obtained a better balance between sensitivity and specificity. Considering the requirement of equally good sensitivity and specificity, the molecule kernel MK2 in combination with the C-SVM performed best, with a sensitivity of 65.5% and a specificity of 74.1%. This model reaches about equal specificity as the ambiguous MCASE model, but has an almost 20% higher sensitivity.

Method	Sensitivity [%]	Specificity [%]	Correct Classification Rate [%]	Coverage [%]
MCASE (unambiguous)	49.5 ± 7.4	77.6 ± 2.6	67.2 ± 1.7	76.9 ± 3.0
MCASE (ambiguous)	46.8 ± 8.3	75.8 ± 3.8	64.6 ± 3.2	98.3 ± 1.0
MK1 + P-SVM	56.6 ± 6.8	81.2 ± 3.3	71.3 ± 2.7	100.0 ± 0.0
MK2 + P-SVM	53.8 ± 7.7	82.3 ± 3.7	70.8 ± 3.2	100.0 ± 0.0
MK1 + C-SVM	63.1 ± 4.2	71.3 ± 4.4	68.3 ± 3.7	100.0 ± 0.0
MK2 + C-SVM	65.3 ± 5.8	74.1 ± 5.2	70.6 ± 4.4	100.0 ± 0.0

Table 5.9: Method comparison on the Chromosome Aberration dataset. Shown are the mean values and standard deviations under 7-fold CV. The MCASE results were obtained by Bayer-Schering Pharma (Sutter, 2008).

5.6 Summary and Conclusions

In this chapter, a new kernel method for QSAR analysis was introduced, which is based on a novel similarity measure (the molecule kernel) and the use of the P-SVM as predictor. Instead of using descriptor vectors, the proposed molecule kernel method implicitly employs similarities in the 3D structures for prediction. In contrast to graph kernel QSAR approaches, which are based on counting paths and walks in the molecular graph, the molecule kernel method takes the 3D geometry of the molecular structure into account. The molecule kernels represent a measure of structural similarity between two given compounds. The P-SVM uses a linear combination of the pairwise similarities of the molecules in the training set to build a predictive activity model. This model will implicitly extract 3D substructures relevant for the prediction of a given endpoint. A necessary precondition for this to work is that the respective substructures are present often enough in the training set, so that the underlying pattern can be learned.

A problem of descriptor-based QSAR models is that the prediction function is expressed in the space of the descriptor variables, therefore the mechanisms underlying the activity of a molecule are hard to investigate. In contrast to this, the prediction function of the molecule kernel method is expressed as a linear combination of structural similarities to a set of support molecules. In this thesis, it was shown how the molecule kernel between a molecule and the set of support molecules can be used together with the respective Lagrange multipliers to build an explanatory QSAR model. Moreover, a visualization technique for this explanatory model was proposed, which is based on color coding the relative influence each atom has on the activity value or class label. This allows to understand why a predicted molecule was assigned a certain target value. If applied to the labeled training set, it might help to gain insight into the general structural properties which influence the activity and might give an idea where modifications to the structure should be applied.

Two different molecule kernels were investigated. Both use similarity functions based on the spatial match of atoms belonging to the same element, which are, however, of different functional form. The molecule was applied in two different application domains relevant in early drug discovery: QSAR analysis and genotoxicity prediction.

In the QSAR domain, the method was compared on four regression and one classification dataset to several state-of-the art descriptor-based QSAR methods. These included approaches based on traditional 2D and 2.5D descriptor vectors as well as a variety of 3D QSAR methods (CoMFA, CoMSIA, HQSAR, EVA). In these experiments, the results of the two proposed molecule kernels using the P-SVM as predictor were consistently better than the results reported in the literature for the other QSAR methods included in the comparison. The empirical evidence suggests that the proposed descriptor-free method offers a promising alternative to existing descriptor-based approaches.

In the genotoxicity prediction domain, the molecule method was applied to the prediction of the outcome of the chromosome-aberration (CA) test. The classification performance was compared to MCASE, a commercial software often used by the phar-

maceutical industry for this task, on a large CA dataset. Compared to MCASE, the proposed method achieved an almost 20% increase in sensitivity at equal specificity. This indicates that the molecule kernel method can also be successfully applied to areas like genotoxicity prediction, where, unlike in receptor-ligand systems, many different (and often unknown) mechanisms are at work.

The two investigated kernels are based on a different similarity measure, and therefore yield different numerical values for the kernel matrix. However, their predictive performance was very similar on all datasets. This could be due to the fact that both kernels use the same level of molecular representation, the atomwise correspondence. These simple kernels were already able to capture enough of the relevant structural similarities to yield excellent prediction performance, which even outperformed much more complex, force-field based approaches on several datasets. Based on the given framework, other molecule kernels can be constructed using more sophisticated similarity measures, which also model bond similarities or electrostatic and steric properties.

As an alignment-free method, molecule kernels do not require any user alignment of the molecules, like CoMFA and CoMSIA, nor the selection of anchor points, as in the Anchor-GRIND method. This saves time, makes the method usable by nonexperts and eliminates potential user bias. In contrast to other 3D QSAR methods, molecule kernels do not require the assumption that there is a single interaction mechanism at the same active site of a macromolecule.

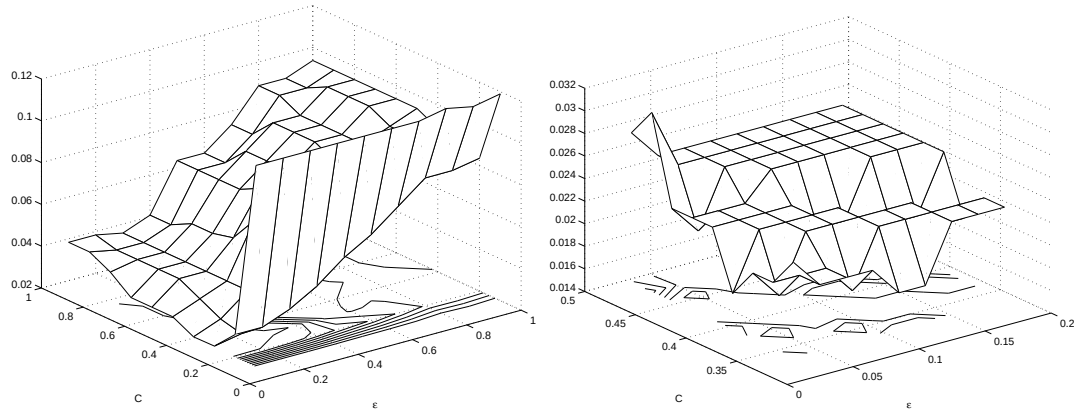
A Appendix

A.1 Hyper-Parameter Optimization on QSAR Datasets

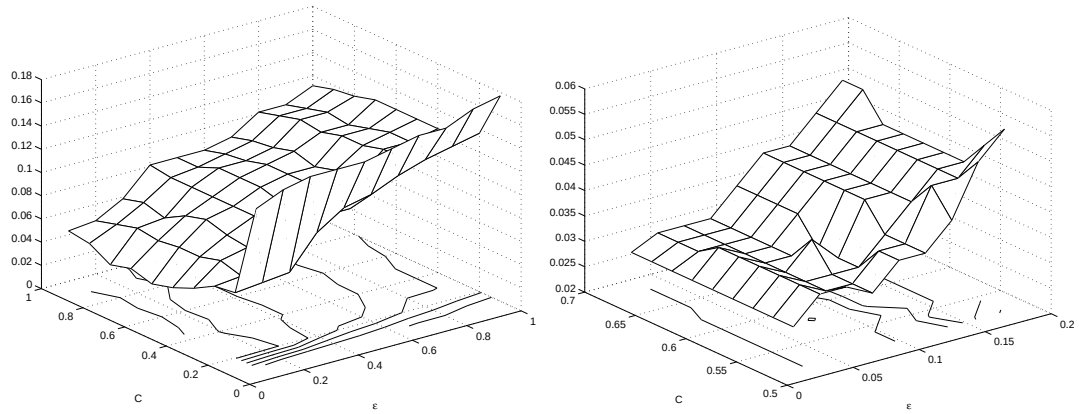
In the following, the results of the optimization of the hyperparameters for the PSVM on the QSAR datasets analyzed in section 5.4 are given for the two molecule kernels MK1 and MK2. The hyperparameters C and ϵ of the P-SVM were selected using a two-phase grid search for the parameter values giving minimal leave-one-out cross-validation (LOOCV) error on each training set.

In the first phase, a rough initial search grid was done, with grid points lying at $\epsilon \in \{0.1, 0.2, \ldots, 1.0\}$, $C \in \{1, 2, \ldots, 10\}$ for regression and $\epsilon \in \{0.05, 0.1, \ldots, 1.0\}$, $C \in \{0.05, 0.1, \ldots, 1.0\}$ for classification. In the second phase, a refined grid search was conducted around the minimum found for the first grid search. The new search grid was chosen as an equally spaced 9×9 grid centered around the found minimum, such that the the start- and endpoints were $1/5 \times$ the previous grid step size away from the points neighboring the minimum in the previous grid.

The result of the hyperparameter selection via a two-phase grid-search on the training set of the classification dataset (Fontaine) is shown in figure A.1. The predictive balanced error rate is shown as a combined surface and contour plot over the hyperparameters C and ϵ of the P-SVM for the first phase (left) and second phase (right) of the grid-search. For the regression datasets (ACE, AChE, BZR and DHFR), the results of the hyperparameter search, conducted using a two-phase grid search on the respective training sets are shown in figures A.2, A.3, A.4 and A.5. The combined surface and contour plots depict the mean squared predictive error as a function of the hyperparameters C and ϵ of the P-SVM.

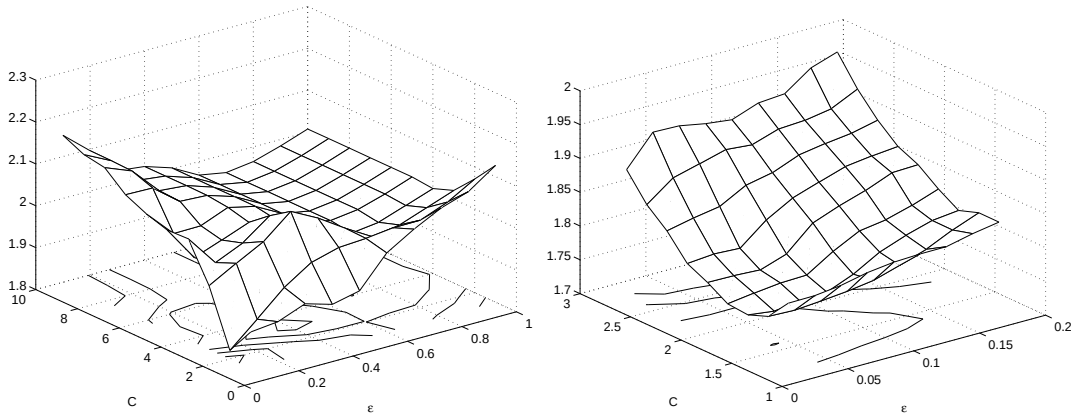


(a) MK1, optimum hyperparameters: $\epsilon = 0.02, C = 0.38$ with a balanced error of 0.018140

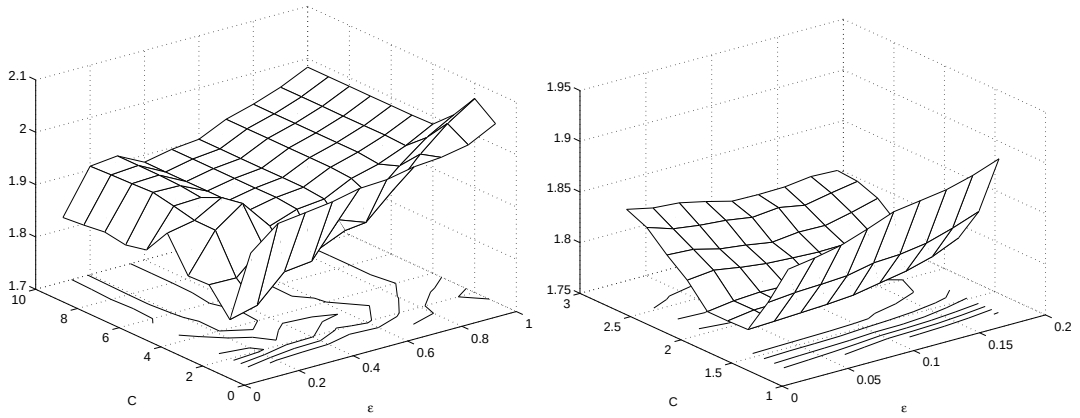


(b) MK2, optimum hyperparameters: $\epsilon = 0.02, C = 0.52$ with a balanced error of 0.028241

Figure A.1: Application of molecule kernel method on the Fontaine dataset: Hyperparameter grid search. Left: phase 1, right: phase 2. The LOOCV balanced error rate is shown as function of the hyperparameters C and ϵ.

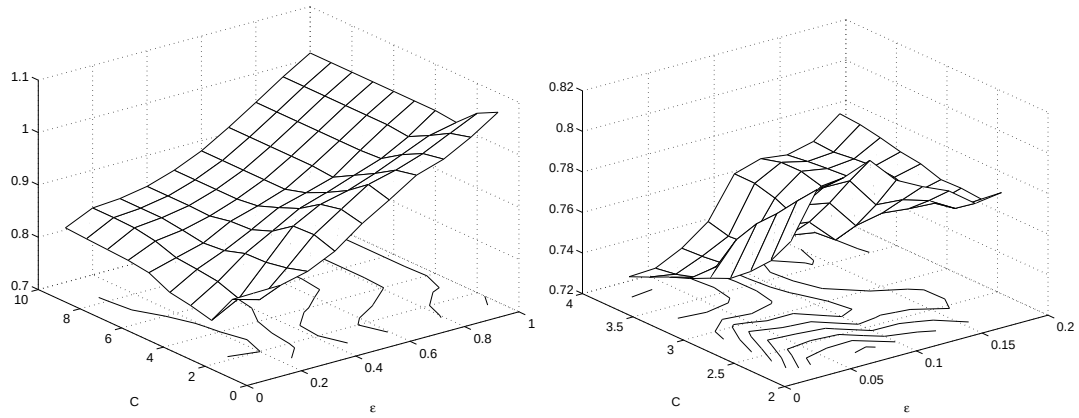


(a) MK1, optimum hyperparameters: $\epsilon = 0.04, C = 1.6$ with a LOOCV-MSE of 1.749284

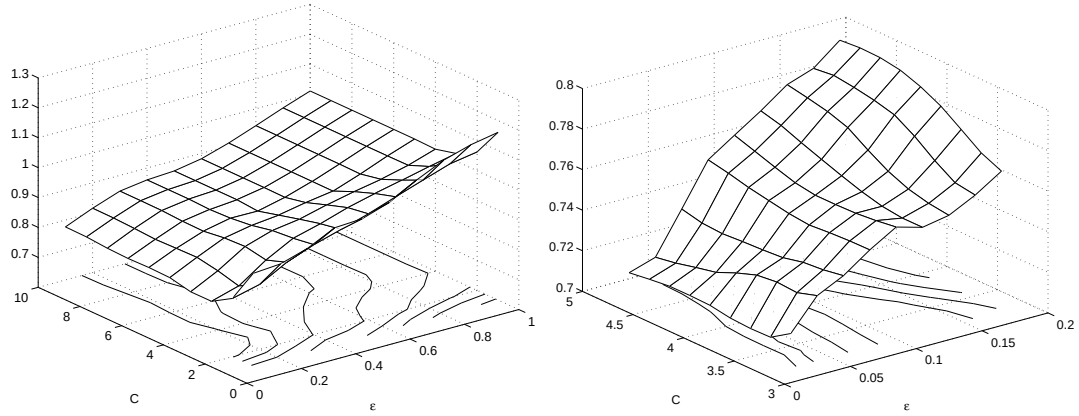


(b) MK2, optimum hyperparameters: $\epsilon = 0.14, C = 2.0$ with a LOOCV-MSE of 1.769261

Figure A.2: Application of the molecule kernel method on the ACE dataset: Hyperparameter grid search. Left: phase 1, right: phase 2.

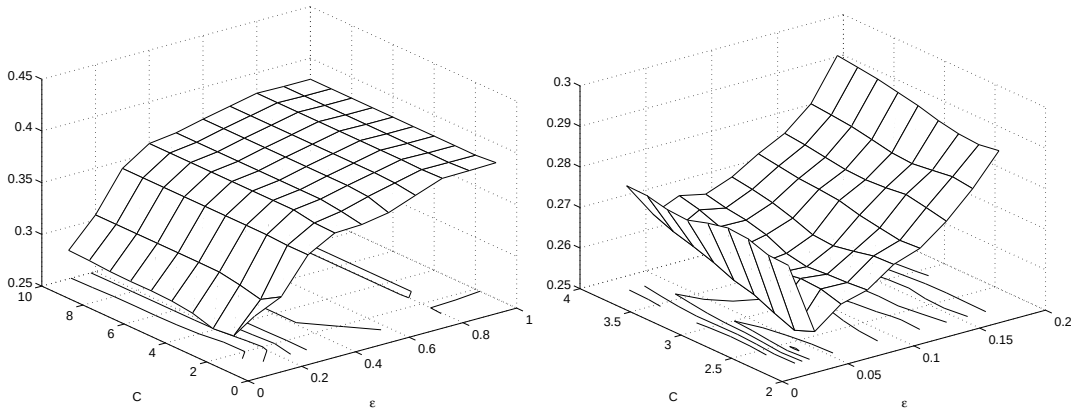


(a) MK1, optimum hyperparameters: $\epsilon = 0.02, C = 3.8$ with a LOOCV-MSE of 0.729378

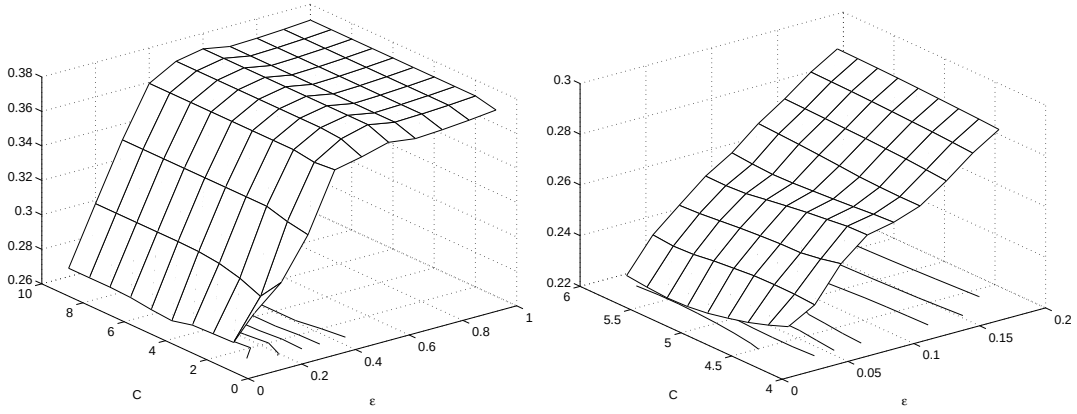


(b) MK2, optimum hyperparameters: $\epsilon = 0.02, C = 3.6$ with a LOOCV-MSE of 0.710225

Figure A.3: Application of the molecule kernel method on the AChE dataset: Hyperparameter grid search. Left: phase 1, right: phase 2.

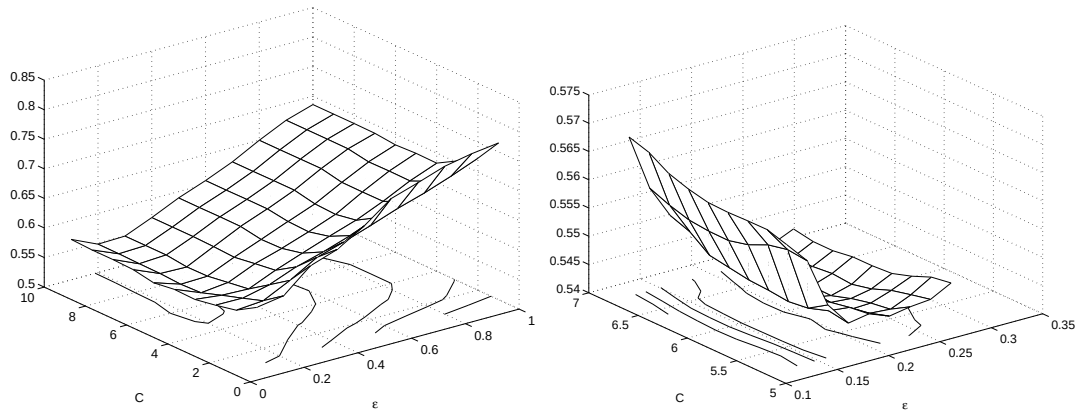


(a) MK1, optimum hyperparameters: $\epsilon = 0.04, C = 2.4$ with a LOOCV-MSE of 0.254526

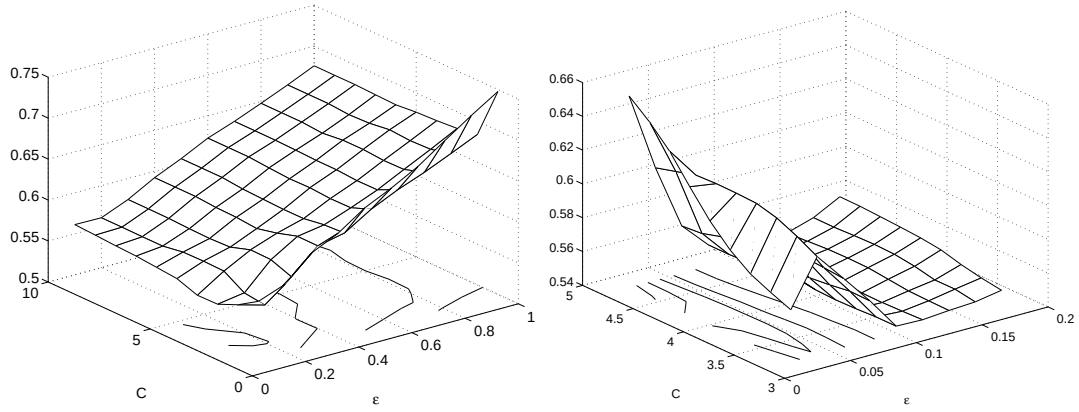


(b) MK2, optimum hyperparameters: $\epsilon = 0.02, C = 5.4$ with a LOOCV-MSE of 0.223651

Figure A.4: Application of the molecule kernel method on the BZR dataset: Hyperparameter grid search. Left: phase 1, right: phase 2.



(a) MK1, optimum hyperparameters: $\epsilon = 0.24, C = 6$ with a LOOCV-MSE of 0.540629.



(b) MK2, optimum hyperparameters: $\epsilon = 0.12, C = 3.6$ with a LOOCV-MSE of 0.543014

Figure A.5: Application of the molecule kernel method on the DHFR dataset: Hyperparameter grid search. Left: phase 1, right: phase 2.

Bibliography

Abarca, J., Campusano, J., Bustos, V., Noriega, V., and Aliaga, E. (2004). Functional interactions between somatodendritic dopamine release, glutamate receptors and brain-derived neurotrophic factor expression in mesencephalic structures of the brain. *Brain Research Reviews*, 47:126–144.

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169.

Afifi, A., Clark, V., and May, S. (2004). *Computer-Aided Multivariate Analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, London, New York, Washington, D.C., fourth edition.

Agartz, I., Momenan, R., Rawlings, R., Kerich, M., and Hommer, D. (1999). Hippocampal volume in patients with alcohol dependence. *Archives of General Psychiatry*, 56:356–363.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control*, 19(6):716–723.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, 16:3–14.

Akaike, H. (1985). A celebration of statistics. In Atkinson, A. and Fienberg, S., editors, *The ISI Centenary Volume*, pages 1–24. Springer Verlag, New York.

Altinbilek, B. and Manahan-Vaughan, D. (2007). Antagonism of group III metabotropic glutamate receptors results in impairment of LTD but not LTP in the hippocampal CA1 region, and prevents long-term spatial memory. *European Journal of Neuroscience*, 26:1166–1172.

Backstrom, P. and Hyytia, P. (2005). Suppression of alcohol self-administration and cue-induced reinstatement of alcohol seeking by the mGlu2/3 receptor agonist LY379268 and the mGlu8 receptor agonist (S)-3,4-DCPG. *European Journal of Pharmacology*, 528:110–118.

Bakhir, G. H., Hofmann, T., Schoelkopf, B., Smola, A. J., Taskar, B., and Vishwanathan, S. V. N. (2007). *Predicting Structured Data*. MIT Press.

Balding, D. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7:781–791.

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

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Beresford, T., Arciniegas, D., Alfers, J., Clapp, L., Martin, B., Du, Y., Liu, D., Shen, D., and Davatzikos, C. (2006). Hippocampus volume loss due to chronic heavy drinking. *Alcoholism Clinical and Experimental Research*, pages 1866–1870.

Bierut, L., Saccone, N., Rice, J., Goate, A., Foroud, T., Edenberg, H., Almasy, L., Conneally, P., Crowe, R., Hesselbrock, V., Li, T., Nurnberger, J. J., Porjesz, B., Schuckit, M., Tischfield, J., Begleiter, H., and Reich, T. (2002). Defining alcohol-related phenotypes in humans. the collaborative study on the genetics of alcoholism. *Alcohol Research & Health*, 26:208–213.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer Science + Business Media, New York.

Bolos, A., Dean, M., Lucas-Derse, S., Ramsburg, M., Brown, G., and Goldman, D. (1990). Population and pedigree studies reveal a lack of association between the dopamine D(2) receptor gene and alcoholism. *Journal of the American Medical Association*, 264:3156–3160.

Bortolotto, Z., Fitzjohn, S., and Collingridge, G. (1999). Roles of metabotropic glutamate receptors in LTP and LTD in the hippocampus. *Current Opinion in Neurobiology*, 9:299–304.

Bowirrat, A. and Oscar-Berman, M. (2005). Relationship between dopaminergic neurotransmission, alcoholism, and reward deficiency syndrome. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 132:29–37.

Bradley, M. and Lang, P. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy & Experimental Psychiatry*, 25:49–59.

Breiman, L. (1994). Bagging predictors. Technical Report 421, Department of Statistics, UC Berkeley. ftp://ftp.stat.berkeley.edu/pub/tech-reports/421.ps.Z.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

Breiter, H. and Gasic, G. (2004). A general circuitry processing reward/aversion information and its implications for neuropsychiatric illness. In Gazzaniga, M., editor, *The Cognitive Neurosciences*, volume III, pages 1043–1065, Cambridge. MIT Press.

Broomhead, D. S. and Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355.

Bruno, V., Battaglia, G., Casabona, G., Copani, A., Caciagli, F., and Nicoletti, F. (1998). Neuroprotection by glial metabotropic glutamate receptors is mediated by transforming growth factor-beta. *Journal of Neuroscience*, 18:9594–9600.

Buckholtz, J., Sust, S., Tan, H., Mattay, V., Straub, R., Meyer-Lindenberg, A., D.R:Weinberger, and J.H.Callicott (2007). fMRI evidence for functional epistasis between COMT and RGS4. *Mol Psychiatry*, 12(10):893–895.

Buckland, P. (2001). Genetic association studies of alcoholism - problems with the candidate gene approach. *Alcohol and Alcoholism*, 36:99–103.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodal Inference: A practical Information-Theoretic Approach*. Springer, Berlin, New York.

Campusano, J., Abarca, J., Forray, M., Gysling, K., and Bustos, G. (2002). Modulation of dendritic release of dopamine by metabotropic glutamate receptors in rat substantia nigra. *Biochemical Pharmacology*, 63:1343–1352.

Caviness, V. J., Kennedy, D., Richelme, C., Rademacher, J., and Filipek, P. (1996). The human brain age 7-11 years: A volumetric analysis based on magnetic resonance images. *Cerebral Cortex*, 6:726–736.

Challis, J. (1995). A procedure for determining rigid body transformation parameters. *J. Biomechanics*, 28(6):733–737.

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Chen, J., Lipska, B., Halim, N., Ma, Q., Matsumoto, M., Melhem, S., Kolachana, B., Hyde, T., Herman, M., Apud, J., Egan, M., Kleinman, J., and Weinberger, D. (2004). Functional analysis of genetic variation in catechol-o-methyltransferase (COMT): effects on mRNA, protein, and enzyme activity in postmortem human brain. *American Journal of Human Genetics*, 75:807–821.

Conn, P. and Pin, J. (1997). Pharmacology and functions of metabotropic glutamate receptors. *Annual Review of Pharmacology and Toxicology*, 37:205–237.

Corti, C., Battaglia, G., Molinaro, G., B.Riozzi, Pittaluga, A., Corsi, M., Mugnaini, M., Nicoletti, F., and Bruno, V. (2007). The use of knock-out mice unravels distinct roles for mGlu2 and mGlu3 metabotropic glutamate receptors in mechanisms of neurodegeneration/ neuroprotection. *Journal of Neuroscience*, pages 8297–8308.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York.

Cowell, R., Dawid, A., Lauritzen, S., and Spiegelhalter, D. (1999). *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Computer Science. Springer.

Cramer, R., Patterson, D., and Bunce, J. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, 110:5959–5967.

De Bellis, M., Clark, D., Beers, S., Soloff, P., Boring, A., Hall, J., and Kersh, A. (2000). Hippocampal volume in adolescent-onset alcohol use disorders. *American Journal of Psychiatry*, 157:737–744.

Dempster, E., Mill, J., Craig, I., and Collier, D. (2006). The quantification of COMT mRNA in post mortem cerebellum tissue: diagnosis, genotype, methylation and expression. *Biomed Central Medical Genetics*, 7:10.

Depriest, S. A., Mayer, D., Naylor, C. B., and Marshall, G. R. (1993). 3DQSARof angiotensin-converting enzyme and thermolysin inhibitors.A comparison of CoMFA models based on deduced andexperimentally determined active-site geometries. *Journal of the American Chemical Society*, 115:5372–5384.

Derogatis, L. (1983). Symptom Checklist 90-Revised/SCL-90-R: Administration, scoring and procedures manual II. In Baltimore, M., editor, *Clinical PsychoMetric Research;*.

Dick, D. and Bierut, L. (2006). The genetics of alcohol dependence. *Current Psychiatry Reports*, 8:151–157.

Drabant, E., Hariri, A., Meyer-Lindenberg, A., Munoz, K., Mattay, V., Kolachana, B., Egan, M., and Weinberger, D. (2006). Catechol o-methyltransferase val158met genotype and neural mechanisms related to affective arousal and regulation. *Arch Gen Psychiatry*, 63(12):1396–1406.

Dravolina, O., Zakharova, E., Shekunova, E., Zvartau, E., Danysz, W., and Bespalov, A. (2007). mGlu1 receptor blockade attenuates cue- and nicotine-induced reinstatement of extinguished nicotine self-administration behavior in rats. *Neuropharmacology*, 52:263–269.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons, New York, 2nd edition.

Edenberg, H. and Foroud, T. (2006). The genetics of alcoholism: identifying specific genes through family studies. *Addiction Biology*, 11:386–396.

Edwards, D. (2000). *Introduction to graphical modelling*. Springer Texts in Statistics, 2nd edition edition.

Egan, M., Goldberg, T., Kolachana, B., Callicott, J., Mazzanti, C., Straub, R., Goldman, D., and Weinberger, D. (2001). Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proc Natl Acad Sci U S A*, 98(12):6917–6922.

Enoch, M., Waheed, J., Harris, C., and Goldman, B. A. D. (2006). Sex differences in the influence of COMT Val158Met on alcoholism and smoking in plains American Indians. *Alcoholism: Clinical and Experimental Research*, 30:399–406.

Enoch, M. A., Schuckit, M., Johnson, B. A., and D.Goldman (2003). Genetics of alcoholism using intermediate phenotypes. *Alcohol Clin Exp Res*, 27:169–176.

Ferguson, A. M., Heritage, T., Jonathon, P., Pack, S. E., and Phillips, L. (1997). EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *Journal of Computer-Aided Molecular Design*, 11:143–152.

Fienberg, S. (1980). *Analysis of Cross-classified Categorial Data*. MIT Press, Cambridge, MA, 2nd edition edition.

First, M., Spitzer, R., Gibbon, M., and Williams, J. (1997). *Structured Clinical Interview for DSM-IV Personality Disorders (SCID-II)*. American Psychiatric Press, Washington D.C.

First, M., Spitzer, R., Gibbon, M., and Williams, J. (2001). *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition With Psychotic Screen (SCID-I/P W/ PSY SCREEN), Biometrics Research*. New York State Psychiatric Institute, New York.

Fisher, R. (1935). The fiducial argument in statistiical inference. *Ann. Eugenics*, 6:391–398.

Fontaine, F., Pastor, M., Zamorra, I., and Sanz, F. (2005). Anchor-GRIND: Filling the gap between standard 3D QSAR and the GRid-INdependent Descriptors. *Journal of Medicinal Chemistry*, 48:2687–2694.

Gärtner, T., Flach, P., and Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. In Schölkopf, B. and Warmuth, M., editors, *Proc. of the Sixteenth Annual Conference on Computational Learning Theory and the Seventh Annual Workshop on Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, Heidelberg. Springer-Verlag.

Goldstein, J., Seidman, L., O'Brien, L., Horton, N., Kennedy, D., Makris, N., Caviness, V. J., and amd M.T. Tsuang, S. F. (2002). Impact of normal sexual dimorphisms on sex differences in structural brain abnormalities in schizophrenia assessed by magnetic resonance imaging. *Archives of General Psychiatry*, 59:154–164.

Goodwin, D. (1975). Genetic determinants of alcohol addiction. *Advances in Experimental Medicine and Biology*, 56:339–355.

Guttman, I. (1967). The use of the concept of future observations in goodness-of-fit problems. *Journal of the Royal Statistical Society, Series B*, 29:38–100.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182. Special Issue on Variable and Feature Selection.

Halperin, E. and Eskin, E. (2004). Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20(12):1842–1849.

Hamilton, M. (1986). The hamilton rating scale for depression. In Sartorius, N. and TA, T. B., editors, *Assessment of depression.*, pages 143–152. Springer, Berlin.

Hariri, A., V.S.Mattay, Tessitore, A., Kolachana, B., Fera, F., Goldman, D., Egan, M., and Weinberger, D. (2002). Serotonin transporter genetic variation and the response of the human amygdala. *Science*, 297:400–403.

Haykin, S. (1994). *Neural Networks : A Comprehensive Foundation*. Macmillan, New York.

Heinz, A., Jones, D., Mazzanti, C., Goldman, D., Ragan, P., Hommer, D., Linnoila, M., and Weinberger, D. (2000). A relationship between serotonin transporter genotype and in vivo protein expression and alcohol neurotoxicity. *Biological Psychiatry*, 47:643–649.

Heinz, A., Schafer, M., Higley, J., Krystal, J., and Goldman, D. (2003). Neurobiological correlates of the disposition and maintenance of alcoholism. *Pharmacopsychiatry*, 36(suppl3):561–70.

Heinz, A., Siessmeier, T., Wrase, J., Buchholz, H., Gruender, G., Kumakura, Y., Cumming, P., Schreckenberger, M., Smolka, M., Roesch, F., Mann, K., and Bartenstein, P. (2005). Correlation of alcohol craving with striatal dopamine synthesis capacity and D2/3 receptor availability: a combined [18F]DOPA and [18F]DMFP PET study in detoxified alcoholic patients. *American Journal of Psychiatry*, 162:1515–1520.

Heritage, T. and Lowis, D. (1999). Molecular hologram QSAR. In *Rational Drug Design: Novel Methodology and Practical Applications*. Oxford University Press: New York.

Hersi, A., Rowe, W., Gaudreau, P., and Quirion, R. (1995). Dopamine D1 receptor ligands modulate cognitive performance and hippocampal acetylcholine release in memory-impaired aged rats. *Neuroscience*, 69:1067–1074.

Hines, L., Ray, L., Hutchison, K., and Tabakoff, B. (2005). Alcoholism: the dissection for endophenotypes. *Dialogues in Clinical Neuroscience*, 7:153–163.

Ho, B., Wassink, T., O'Leary, D., Sheffield, V., and Andreasen, N. (2005). Catechol-O-Methyl transferase Val158Met gene polymorphism in schizophrenia: working memory, frontal lobe MRI morphology and frontal cerebral blood flow. *Mol Psychiatry*, 10(229):287–298.

Hochreiter, S. and Obermayer, K. (2004). Gene selection for microarray data. In Schölkopf, B., Tsuda, K., and Vert, J.-P., editors, *Kernel Methods in Computational Biology*, pages 319–355. MIT Press.

Hochreiter, S. and Obermayer, K. (2005). Nonlinear feature selection with the potential support vector machine. In Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L., editors, *Feature extraction, Foundations and Applications*. Springer.

Hochreiter, S. and Obermayer, K. (2006). Support vector machines for dyadic data. *Neural Computation*, 18:1472–1510.

Hommer, D. (2003). Male and female sensitivity to alcohol-induced brain damage. *Alcohol Research & Health*, 27:181–185.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational facilities. *Proceedings of the National Academy of Sciences of the USA*, 79:2554–2558.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Adaptive and Learning Systems for Signal Processing, Communications, and Control, S. Haykin, editor. John Wiley & Sons, New York.

Jay, T. (2003). Dopamine: a potential substrate for synaptic plasticity and memory mechanisms. *Progress in Neurobiology*, 69:375–390.

Ji, Y., Pang, P. T., Feng, L., and Lu, B. (2005). Cyclic AMP controls BDNF-induced trkb phosphorylation and dendritic spine formation in mature hippocampal neurons. *Nature Neuroscience*, 8:164–172.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1998). An introduction to variational methods for graphical models. In *Learning in Graphical Models*, volume M.I. Jordan, pages 105–162. Kluwer.

Kashima, H., Tsuda, K., and Inokuchi, A. (2003). Marginalized kernels between labeled graphs. In Faucett, T. and Mishra, N., editors, *Proceedings of the Twentieth International Conference on Machine Learning*, pages 321–328. AAAI Press.

Kauhanen, J., Hallikainen, T., Tuomainen, T., Koulu, M., Karvonen, M., Salonen, J., and Tiihonen, J. (2000). Association between the functional polymorphism of catechol-O-methyltransferase gene and alcohol consumption among social drinkers. *Alcoholism: Clinical and Experimental Research*, 24:135–139.

Kearsley, S. K. and Smith, G. (1990). An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. *Tetrahedron Computer Methodology*, 3(6c):615–633.

Kelley, A. (2004). Memory and addiction: shared neural circuitry and molecular mechanisms. *Neuron*, 44:161–179.

Kendler, K., Neale, M., Heath, A., Kessler, R., and Eaves, L. (1994). A twin-family study of alcoholism in women. *American Journal of Psychiatry*, 151:707–715.

Kirkland, D., Aardema, M., Henderson, L., and Mueller, L. (2005). Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens. I. Sensitivity, specificity, and relative predictivity. *Mutation Research*, 584(1-2):1–256.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220:219–227.

Klebe, G., Abraham, U., and Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *Journal of Medicinal Chemistry*, 37(24):4130–46.

Klopman, G. and Rosenkranz, H. S. (1994). Approaches to sar in carcinogenesis and mutagenesis. prediction of carcinogenicity/mutagenicity using multi-case. *Mutation Research*, 305:33–46.

Knebel, T., Hochreiter, S., and Obermayer, K. (2008). An SMO algorithm for the potential support vector machine. *Neural Computation*, 20:271–287.

Knecht, S., Breitenstein, C., Bushuven, S., Wailke, Kamping, S., Floel, A., Zwitserlood, P., and Ringelstein, E. (2004). Levodopa: faster and better word learning in normal humans. *Annals of Neurology*, 56:20–26.

Koehnke, M. (2008). Approach to the genetics of alcoholism: A review based on pathophysiology. *Biochemical Pharmacology*, 75:160–177.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324.

Kohonen, T. (1982a). Analysis of a simple self-organizing process. *Biological Cybernetics*, 43:135–140.

Kohonen, T. (1982b). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Koob, C. (2003). Alcoholism: Allostasis and beyond. *Alcoholism: Clinical and Experimental Research*, 27:232–243.

Kotlinska, J. and Bochenski, M. (2007). Comparison of the effects of mGluR1 and mGluR5 antagonists on the expression of behavioral sensitization to the locomotor effect of morphine and the morphine withdrawal jumping in mice. *European Journal of Pharmacology*, 558:113–118.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Lang, P. (1995). The emotion probe. Studies of motivation and attention. *Am Psychol.*, 50:372–385.

Lang, P., Bradley, M., and Cuthbert, B. (1999). *The International Affective Picture System (IAPS).* Center for Research in Psychophysiology, University of Florida;, Gainsville, FL.

Legault, M. and Wise, R. (2001). Novelty-evoked elevations of nucleus accumbens dopamine: dependence on impulse flow from the ventral subiculum and glutamatergic neurotransmission in the ventral tegmental area. *European Journal of Neuroscience*, 13:819–828.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21(3):105–117.

Lisman, J. and Grace, A. (2005). The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron*, 46:703–713.

Lohoff, F., Lautenschlager, M., Mohr, J., Ferraro, T., Sander, T., and Gallinat., J. (2008). Association between variation in the vesicular monoamine transporter 1 gene on chromosome 8p and anxiety-related personality traits. *Neuroscience Letters*, 434:4145.

MacKay, D. (2003). *Information Theory, Inference and Learning Algorithms.* Cambridge University Press.

Makris, N., Gasic, G., Seidman, L., Goldstein, J., Gastfriend, D., Albaugh, I. E. M., Hodge, S., Ziegler, D., Sheahan, F., Caviness, V. J., Tsuang, M., Kennedy, D., Hyman, S., Rosen, B., and Breiter, H. (2004). Decreased absolute amygdala volume in cocaine addicts. *Neuron*, 44:729–740.

Markou, A. (2007). The role of metabotropic glutamate receptors in drug reward, motivation and dependence. *Drug News and Perspectives*, 20:103–108.

McCulloch, W. S. and Pitts, W. H. (1943). A logical calculus of the idea immanent in nervous activity. *Bull. Math. Biophy*, 5:115–133.

Mechtcheriakov, S., Brenneis, C., Egger, K., Koppelstaetter, F., Schocke, M., and Marksteiner, J. (2007). A distinct pattern of cerebral atrohy in patients with alcohol

addiction revealed by voxel-based morphometry. *Journal of Neurology, Neurosurgery and Psychiatry*, 78:610–614.

Medina, L., Schweinsburg, A., Cohen-Zion, M., Nagel, B., and Tapert, S. (2007). Effects of alcohol and combined marijuana and alcohol use during adolescence on hippocampal volume and asymmetry. *Neurotoxicology Teratology*, 29:141–152.

Meltzer, L., Serpa, K., and Christoffersen, C. (1997). Metabotropic glutamate receptor-mediated inhibition and excitation of substantia nigra dopamine neurons. *Synapse*, 26:184–193.

Meyer-Lindenberg, A., Nichols, T., Callicott, J., Ding, J., Kolachana, B., Buckholtz, J., Mattay, V., Egan, M., and Weinberger, D. (2006). Impact of complex genetic variation in COMT on human brain function. *Molecular Psychiatry*, 11:867–877.

Mohr, J., Jain, B., and Obermayer, K. (2008a). Molecule kernels: A descriptor- and alignment-free quantitative structure activity relationship approach. *Journal of Chemical Information and Modeling*. In Press.

Mohr, J., Puls, I., Wrase, J., Hochreiter, S., Heinz, A., and Obermayer, K. (2006). P-SVM variable selection for discovering dependencies between genetic and brain imaging data. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

Mohr, J., Puls, I., Wrase, J., Priller, J., Behr, J., Kitzrow, W., Makris, N., Breiter, H., Obermayer, K., and Heinz, A. (2008b). Synergistic effects of the dopaminergic and glutamatergic system on hippocampal volume in alcohol-dependent patients. *Biological Psychology*, 79(1):126–136. the first two authors have contributed equally.

Mohr, J., Puls, I., Wrase, J., Vollstädt-Klein, S., Lemenager, T., Volmert, C., Rapp, M., Obermayer, K., Heinz, A., and Smolka, M. (2008c). A model comparison of COMT effects on central processing of affective stimuli - Advantage of haplotype analysis for genomic imaging? The first two authors have contributed equally. In Review.

Mohr, J., Seo, S., and Obermayer, K. (2008d). Automated microarray classification based on P-SVM gene selection. In *Proceedings of the ICMLA '08: The Seventh International Conference on Machine Learning and Applications*, San Diego, CA, USA. the first two authors have contributed equally. In Press.

Mohr, J., Seo, S., Puls, I., Heinz, A., and Obermayer, K. (2008e). Target selection: A new learning paradigm and its application to genetic association studies. In *Proceedings of the ICMLA '08: The Seventh International Conference on Machine Learning and Applications*, San Diego, CA, USA. In Press.

Nackley, A. G., Shabalina, S. A., Tchivileva, I. E., Satterfield, K., Korchynskyi, O., Makarov, S. S., Maixner, W., and Diatchenko, L. (2006). Human Catechol-O-Methyltransferase Haplotypes Modulate Protein Expression by Altering mRNA Secondary Structure. *Science*, 314(5807):1930–1933.

Nagel, B., Schweinsburg, A., Phan, V., and Tapert, S. (2005). Reduced hippocampal volume among adolescents with alcohol use disorders without psychiatric comorbidity. *Psychiatry Research*, 139:181–190.

Nicodemus, K., Kolachana, B., Vakkalanka, R., Straub, R., Giegling, I., Egan, M., Rujescu, D., and Weinberger, D. (2007). Evidence for statistical epistasis between catechol-o-methyltransferase (COMT) and polymorphisms in RGS4, G72 (DAOA), mGluR3, and DISC1: influence on risk of schizophrenia. *Human Genetics*, 120:889–906.

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430.

Oroszi, G. and Goldman, D. (2004). Alcoholism: genes and mechanisms. *Pharmacogenomics*, 5(1037-1048).

Pastor, M., Cruciani, G., McLay, I., Pickett, S., and Clementi, S. (2000). Grid-independent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *Journal of Medicinal Chemistry*, 43:3233–3243.

Peters, J. and Kalivas, P. (2006). The group II metabotropic glutamate receptor agonist, LY379268, inhibits both cocaine- and food-seeking behavior in rats. *Psychopharmacology*, 186:143–149.

Pezawas, L., Meyer-Lindenberg, A., Drabant, E., Verchinski, B., Munoz, K., Kolachana, B., Egan, M., Mattay, V., Hariri, A., and Weinberger, D. (2005). 5-HTTLPR polymorphism impacts human cingulate-amygdala interactions: a genetic susceptibility mechanism for depression. *Nature Neuroscience*, 5:828–834.

Pfefferbaum, A. (2004). Alcoholism damages the brain, but does moderate alcohol use? *Lancet Neurology*, 3:143–144.

Pizzi, M., Consolandi, O., Memo, M., and Spano, P. (1996). Activation of multiple metabotropic glutamate receptor subtypes prevents NMDA-induced excitotoxicity in rat hippocampal slices. *European Journal of Neuroscience*, 8:1516–1521.

Radloff, L. (1997). The CES-D scale: A self report depression scale for research in the general population. *Appl Psychol Meas.*, 1::385–401.

Ralaivola, L., Swamidass, S., Saigo, H., and Baldi, P. (2005). Graph kernels for chemical informatics. *Neural Networks*, 18:1093–1110.

Reischies, F., Neuhaus, A., Hansen, M., Mientus, S., Mulert, C., and Gallinat, J. (2005). Electrophysiological and neuropsychological analysis of a delirious state: the role of the anterior cingulate gyrus. *Psychiatry Research*, 138:171–181.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, U. K.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Rothfuss, A., Steger-Hartmann, T., Heinrich, N., and Wichard, J. (2006). Computational prediction of the chromosome-damaging potentialof chemicals. *Chemical Research in Toxicology*, (19):1313–1319.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(9):533–536.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels – Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge.

Schumann, G., Saam, C., Heinz, A., Mann, K., and J.Treutlein (2005). The NMDA receptor system: genetic risk factor for alcoholism. *Nervenarzt*, 76:1355–1362.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Seidman, L., Faraone, S., Goldstein, J., Kremen, W., Horton, N., Makris, N., Toomey, R., Kennedy, D., Caviness, V., and Tsuang, M. (2002). Left hippocampal volume as a vulnerability indicator for schizophrenia: a magnetic resonance imaging morphometric study of nonpsychotic first-degree relatives. *Archives of General Psychiatry*, 59:839–849.

Serra, J., Thompson, E., and Jurs, P. (2003). Development of binary classification of structural chromosome aberrations for a diverse set of organic compounds from molecular structure. *Chem. Res. Toxicology*, 16:153:163.

Shen, K. and Johnson, S. (1997). A slow excitatory postsynaptic current mediated by g-protein-coupled metabotropic glutamate receptors in rat ventral tegmental dopamine neurons. *European Journal of Neuroscience*, 9:48–54.

Sher, K., Grekin, E., and Williams, N. (2005). The development of alcohol use disorders. *Annual Review of Clinical Psychology*, 1:493–523.

Shigemoto, R., Kinoshita, A., Wada, E., Nomura, S., Ohishi, H., Takada, M., Flor, P., Neki, A., Abe, T., Nakanishi, S., and Mizuno, N. (1997). Differential presynaptic localization of metabotropic glutamate receptor subtypes in the rat hippocampus. *Journal of Neuroscience*, 17:7503–7522.

Skinner, H. and Horn, J. (1984). *Alcohol Dependence Scale: Users Guide*. Addiction Research Foundation, Toronto.

Skinner, H. and W.J. Sheu, W. (1982). Reliability of alcohol use indices. the lifetime drinking history and the mast. *Journal of Studies on Alcohol*, 43:1157–1170.

Smolka, M., Bühler, M., and Schumann, G. (2007). Gene-gene effects on central processing of aversive stimuli. *Mol Psychiatry.*, 12:307–317.

Smolka, M. N., Schumann, G., Wrase, J., Grusser, S. M., Flor, H., Mann, K., Braus, D. F., Goldman, D., Buchel, C., and Heinz, A. (2005). Catechol-O-methyltransferase val158met genotype affects processing of emotional stimuli in the amygdala and prefrontal cortex. *J Neurosci*, 25:836–842.

Snyder, R. and Green, J. (2001). A review of the genotoxicity of marketed pharamceuticals. *Mutat. Research*, 488:151–169.

Spampanato, M., Castello, M., Rojas, R., Palacios, E., Frascheri, L., and Descartes, F. (2005). Magnetic resonance imaging findings in substance abuse: alcohol and alcoholism and syndromes associated with alcohol abuse. *Topics in Magnetic Resonance Imaging*, 16:223–230.

Spanagel, R., Pendyala, G., Abarca, C., Zghoul, T., Sanchis-Segura, C., Magnone, M., Lascorz, J., Depner, M., Holzberg, D., Soyka, M., Schreiber, S., Matsuda, F., Lathrop, M., Schumann, G., and Albrecht, U. (2005). The clock gene Per2 influences the glutamatergic system and modulates alcohol consumption. *Nature Medicine*, 11:35–42.

Spielberger, C., Gorsuch, R., and Lushene, R. (1970). *Manual for the State-Trait-Anxiety Inventory.* Consulting Psychologists Press, Palo Alto, CA.

Stone, M. (1977). Asymptotic equivalence of choice of models by cross-valdiation and akaikes criterion. *Journal of the Royal Statistical Society, Series B*, 39:44–47.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods*, A7:13–26.

Sullivan, E., Marsh, L., Mathalon, D., and Lim, K. (1995). Anterior hippocampal volume deficits in nonamnesic, aging chronic-alcoholics. *Alcoholism: Clinical and Experimental Research*, 19:110–122.

Sutherland, J. J., OBrien, L. A., and Weaver, D. F. (2004). A comparison of methods for modeling quantitative structure-activity relationships. *Journal of Medicinal Chemistry*, 47:5541–5554.

Sutter, A. (2008). Bayer-Schering Pharma (personal communication).

Tan, H., Chen, Q., Sust, S., Buckholtz, J., Meyers, J., Egan, M., Mattay, V., Meyer-Lindenberg, A., Weinberger, D., and Callicott, J. (2007). Epistasis between catechol-O-methyltransferase and type II metabotropic glutamate receptor 3 genes on working memory brain function. In *Proceedings of the National Academy Sciences U.S.A.*, volume 104, pages 12536–12541.

Taylor, S., Phan, K., Decker, L., and Liberzon, I. (2003). Subjective rating of emotionally salient stimuli modulates neural activity. *Neuroimage*, 18:650–659.

Tsai, G., Ragan, P., Chang, R., Chen, S., Linnoila, V., and Coyle, J. (1998). Increased glutamatergic neurotransmission and oxidative stress after alcohol withdrawal. *American Journal of Psychiatry*, 155:726–732.

Vapnik, V. and Chervonenkis, A. (1991). The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3):283–305.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York.

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Transactions of the Mathematical Society*, 54:426–482.

White, A., Mueller, R., Gallavan, R., Aaron, S., and Wilson, A. (2003). A multiple in silico program approach for the prediction of mutagenicity from chemical structure. *Mutation Research*, 539:77–89.

Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems (NIPS) 8*, pages 514–520. MIT Press.

Wittmann, B., Schott, B., Guderian, S., Frey, J., Heinze, H., and Duzel, E. (2005). Reward-related fmri activation of dopaminergic midbrain is associated with enhanced hippocampus-dependent long-term memory formation. *Neuron*, 45:459–467.

Wold, S., Ruhe, A., Wold, H., and Dunn, W. (1984). The collinearity problem in linear regressio. The partial least squares approach to generalized inverses. *SIAM J. Sci. Stat. Comput.*, 5:735–743.

Wold, S., Sjostrom, M., and Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, 58:109–130.

Wong, D., Maini, A., Rousset, O., and Brasi, J. (2003). Positron emission tomography - a tool for identifying the effects of alcohol dependence on the brain. *Alcohol Research & Health*, 27:161–173.

Yao, H., Ding, J., Zhou, F., Wang, F., Hu, L., Sun, T., and Hu, G. (2005). Enhancement of glutamate uptake mediates the neuroprotection exerted by activating group II or III metabotropic glutamate receptors on astrocytes. *Journal of Neurochemistry*, 92:948–961.

Index