eScience Approaches to Model Selection and Assessment

Applications in Bioinformatics

MARTIN EKLUND
Dissertation presented at Uppsala University to be publicly examined in B42, Husargatan 3, Uppsala, BMC, Saturday, November 28, 2009 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

High-throughput experimental methods, such as DNA and protein microarrays, have become ubiquitous and indispensable tools in biology and biomedicine, and the number of high-throughput technologies is constantly increasing. They provide the power to measure thousands of properties of a biological system in a single experiment and have the potential to revolutionize our understanding of biology and medicine. However, the high expectations on high-throughput methods are challenged by the problem to statistically model the wealth of data in order to translate it into concrete biological knowledge, new drugs, and clinical practices. In particular, the huge number of properties measured in high-throughput experiments makes statistical model selection and assessment exigent. To use high-throughput data in critical applications, it must be warranted that the models we construct reflect the underlying biology and are not just hypotheses suggested by the data. We must furthermore have a clear picture of the risk of making incorrect decisions based on the models.

The rapid improvements of computers and information technology have opened up new ways of how the problem of model selection and assessment can be approached. Specifically, eScience, i.e. computationally intensive science that is carried out in distributed network environments, provides computational power and means to efficiently access previously acquired scientific knowledge. This thesis investigates how we can use eScience to improve our chances of constructing biologically relevant models from high-throughput data. Novel methods for model selection and assessment that leverage on computational power and on prior scientific information to "guide" the model selection to models that a priori are likely to be relevant are proposed. In addition, a software system for deploying new methods and make them easily accessible to end users is presented.

Keywords: bioinformatics, high-throughout biology, eScience, model selection, model assessment

Martin Eklund, Pharmaceutical Pharmacology, Box 591, Uppsala University, SE-75124 Uppsala, Sweden

© Martin Eklund 2009

ISSN 1651-6192
urn:nbn:se:uu:diva-109437 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-109437)
"Kaffet värmde gott i magen. Den där underbara kombinationen av koffein och nikotin smakar storstad, dödtimmar på fik, ett stilla bläddrande i en utländsk tidning och tomma dialoger i väntan på något som aldrig kommer inträffa; det är själva möjligheten som får blodet att skälva."

Klas Östergren, Gentlemen
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Additional papers


• B. Rogell, M. Hofman, M. Eklund, A. Laurila, and J. Höglund. The interaction of multiple environmental stressors affects adaptation to a novel habitat in the natterjack toad Bufo calamita. Journal of Evolutionary Biology, accepted.


• M. Junaid, M. Lapins, M. Eklund, O. Spjuth, J.E.S. Wikberg. Proteochemometric modeling of the susceptibility of mutated variants of HIV-1 virus to nucleoside/nucleotide analog reverse transcriptase inhibitors. Submitted.
# Contents

1 Abbreviations and notation ................................................. 9  
1.1 Abbreviations ......................................................... 9  
1.2 Variable notations ................................................... 9  
1.3 High-throughput and large-scale ................................... 10  
1.4 A note on the notation in multilevel models ......................... 10  
2 Introduction .................................................................. 11  
3 Background .................................................................. 13  
  3.1 High-throughput experimental techniques and their applications 13  
  3.1.1 DNA microarrays ............................................... 13  
  3.1.2 Protein microarrays ............................................. 14  
  3.1.3 High-throughput screening .................................... 15  
  3.2 eScience .............................................................. 15  
  3.2.1 Semantic information .......................................... 16  
  3.2.2 High-performance computing ................................. 16  
  3.2.3 Software services ............................................... 17  
  3.3 Model selection .................................................... 17  
  3.3.1 The bias-variance tradeoff .................................... 17  
  3.3.2 Controlling the model complexity ........................... 18  
  3.3.3 Choosing a model ............................................... 20  
  3.3.4 Assessing a model ............................................... 22  
4 Motivation for the work in this thesis .................................. 23  
  4.1 A motivating example: Predicting distant metastasis development in breast cancer patients from gene expression microarray data .. 23  
  4.2 High-throughput biology poses new challenges for model selection and assessment ......................................................... 24  
  4.3 Aims .................................................................. 25  
5 Methods .................................................................... 26  
  5.1 eScience methods .................................................... 26  
  5.1.1 Data integration .................................................. 26  
  5.1.2 High-performance computing ................................. 26  
  5.1.3 Optimization and search algorithms ......................... 26  
  5.2 Bayesian statistics .................................................. 28  
  5.2.1 Markov chain Monte Carlo methods ........................ 29  
  5.2.2 Bayesian model selection and model averaging ............. 30  
  5.3 Disturbing datasets by adding noise ................................ 31  
  5.3.1 Using datasets disturbed by noise in model selection ........ 32  
  5.4 Rich clients .......................................................... 32  
6 Results and discussion ..................................................... 33  
  6.1 Papers I and II ....................................................... 33
1. Abbreviations and notation

1.1 Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>Akaike’s Information Criterion</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>E</td>
<td>Expected value</td>
</tr>
<tr>
<td>GA</td>
<td>Genetic algorithm</td>
</tr>
<tr>
<td>HPC</td>
<td>High-performance computing</td>
</tr>
<tr>
<td>HTS</td>
<td>High-throughput screening</td>
</tr>
<tr>
<td>L</td>
<td>Loss function</td>
</tr>
<tr>
<td>n</td>
<td>Number of rows in a matrix</td>
</tr>
<tr>
<td>$N_p(\mu, \Sigma)$</td>
<td>A $p$-variate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$</td>
</tr>
<tr>
<td>p</td>
<td>Number of columns in a matrix</td>
</tr>
<tr>
<td>PRSS</td>
<td>Penalized residual sum of squares</td>
</tr>
<tr>
<td>$R_g$</td>
<td>Generalization error</td>
</tr>
<tr>
<td>$R_t$</td>
<td>Training error</td>
</tr>
<tr>
<td>RSS</td>
<td>Residual sum of squares</td>
</tr>
<tr>
<td>var</td>
<td>Variance</td>
</tr>
<tr>
<td>$\sim$</td>
<td>&quot;Distributed according to&quot;</td>
</tr>
<tr>
<td>$\Phi(u)$</td>
<td>The normal cumulative distribution function</td>
</tr>
</tbody>
</table>

1.2 Variable notations

Random variables are denoted by uppercase letters, such as $X$, $W$ and $Y$. Independent variables will typically be denoted $X$ or $W$ (if they are errors-in-variables), and dependent variables $Y$. If $X$ is a vector, its components will be accessed by subscripts according to $X_j$. Realizations of random variables are written in lowercase. If a random variable is a vector I will write the realization of it in bold. For example, if $X$ is a vector, the $i$th observed value of $X$ is written as $x_i$ and the whole observed vector as $x$. Matrices are represented by bold uppercase letters. For example, a set of $n$ realizations of a $p$-dimensional random variable $X$ will be represented by the $n \times p$ matrix $X$. Rows and columns in matrices will be accessed by the indices $i$ and $j$, respectively.
All vectors are assumed to be column vectors. So, for example, the $i$th row of the matrix $X$ is $x_i^T$, where $T$ denotes the transpose, and the $j$th column of $X$ is $x_j$.

1.3 High-throughput and large-scale

There is a difference between large-scale and high-throughput experimental techniques, where large-scale means 'massively parallel', whereas high-throughput means 'fast repetition'. For example an experiment where the expression of thousands of genes in one tissue sample is measured simultaneously (i.e. DNA microarrays) is large-scale, whereas measuring the interaction activity of thousands of molecule-protein interactions (i.e. high-throughput screening) is high-throughput. However, large-scale and high-throughput techniques alike generate vast amounts of data and similar computational challenges are faced in the data analysis. Lately, the term high-throughput has been used to encompass both large-scale and high-throughput (see e.g. [1]). Supported by this, I will use high-throughput throughout this thesis both when referring to high-throughput and large-scale in order to avoid overly cumbersome language.

1.4 A note on the notation in multilevel models

I will express multilevel models in the Laird-Ware form [2], i.e:

$$y_k = f(X_k, \beta_k) + \varepsilon_k, \quad k = 1, ..., K$$

$$\beta_k \sim N_p(0, \Sigma)$$

$$\varepsilon_k \sim N_{nk}(0, \Psi)$$

(1.1)

where $K$ is the number of groups in the data matrix $X$, $\beta_k$ is the vector of regression coefficients for group $k$, and $\varepsilon_k$ are the error vector associated with group $k$. $n_k$ is the number of observations in group $k$. Hence,

$$\sum_{k=1}^{K} n_k = n.$$
High-throughput experimental methods, such as DNA and protein microarrays, have become ubiquitous and indispensable tools in biology and biomedicine, and the number of high-throughput technologies is constantly increasing. They provide the power to measure thousands of properties of a biological system in a single experiment and have the potential to revolutionize our understanding of biology and medicine. To translate the wealth of data into concrete biological knowledge, new drugs, and clinical practices, it needs to be carefully analyzed. The statistical analysis of data from high-throughput experiments has therefore become of great interest and importance. However, the analysis is characterized by a number of difficulties. In particular, the huge number of properties measured in high-throughput experiments makes statistical model selection and assessment challenging.

Consider for example the problem of predicting cancer prognosis from gene expression microarray data. We typically measure the expression of around 20,000 genes on (at the most) some hundred patients. This gives us a data matrix with roughly 20,000 columns and only a few hundred noisy observations. Based on this data, there is a distinct risk of overfitting our predictor of cancer prognosis and detecting spurious correlations between the gene expression measurement and the prognosis. It is easy to construct predictors with good performance on training data, but how do we ensure that it also performs well on new data? How do we attack the model selection problem from a computational and statistical point of view in such a setting? And how do we produce estimators of the generalization error with the limited number of observations we have available? These issues are of crucial importance. If we want to use high-throughput data in critical applications, such as cancer prognostication, the models we construct need to reflect the underlying biology and not just be experimental artifacts and hypotheses suggested by the data. To gain acceptance in applications where the subsequent decisions have serious consequences, we need to have a clear picture of how much we can trust predictions from the models.

The main idea in this thesis is that eScience, i.e. the current systematic development of research methods that exploit advanced computational thinking, equip us with new tools for addressing the problem of model selection and assessment when analyzing high-throughput data. eScience allows us to calculate more, and to make use of what we already know when approaching a new modeling problem. High-performance computing resources permit us to apply computationally demanding methods even to the large modeling problems that arises when analyzing high-throughput data. Furthermore, huge amounts of scientific data and information have over the years been collected and made publicly available on the Internet. The development of interoperable machine-to-machine techniques means that we readily can harvest the Internet for this already available information to “guide” our modeling approaches when analyzing a new dataset.
However to make this possible on a practical level is a challenge. In this thesis I present some methods and software tools that are my contributions in this direction.
3. Background

3.1 High-throughput experimental techniques and their applications

Biology and biomedicine have become progressively data-rich subjects. Over the last decade researchers have miniaturized and robotized experimental techniques in order to acquire more data at a perpetually increasing rate. This has provided the possibility of studying parts of living organisms holistically, for example the entire genome, transcriptome, or proteome. The objective is to use the generated data to help us understand living organisms better at a molecular level, to enhance drug development processes, to diagnose and prognosticate diseases, and to make better clinical treatments decisions. To achieve this goal a good infrastructure for managing, integrating, and visualizing the voluminous data is vital. Furthermore, the data needs to be carefully analyzed, and hypotheses about the molecular underpinnings of biological systems under study need to be generated and tested.

A considerable number of high-throughput experimental techniques have been developed over the last decade, such as the ’next-generation’ gene sequencing methods (see e.g. Schuster [3]) and high-throughput massspectrometry (see Hood et al. [4] for a recent example). In this section, I give a brief overview of some high-throughput experimental methods relevant for this thesis.

3.1.1 DNA microarrays

DNA microarrays, first described in a seminal paper by Schena et al. [5], enables simultaneous measurement of the transcription level of every gene within a cell (the transcriptome). A number of different DNA microarray approaches exist, but the basic principle is the same. Short segments of DNA are attached to a solid support (usually a glass slide) in a grid-like pattern. mRNA from a sample is reverse-transcribed to cDNA and labelled with a fluorescent dye. The labelled cDNA is subsequently hybridized to the microarray. After unbound labelled cDNA has been washed off the slide, the fluorescence intensities are quantified in each position in the grid using a laser scanner. The fluorescence intensities reflect the mRNA expression levels of the corresponding gene. For a comprehensive review of the DNA microarray technology, I refer to Stoughton [6].

The dominant use of DNA microarrays is to study changes in the transcription level of genes in a particular tissue or cell type under disease states, during development, or in response to intentional experimental perturbations, such as gene disruptions and drug treatments. The response patterns have helped illuminate mechanisms of disease and identify disease subphenotypes [7], predict disease progression [7], assign func-
tion to previously unannotated genes [8], and group genes into functional pathways [9].

An example of a DNA microarray application that has been studied extensively is breast cancer prognosis prediction from gene expression data. Breast cancer is the most common form of cancer and the second leading cause of death from cancer among women in the Western world [10]. The main cause of breast cancer death comes from its metastases to distant sites. Early diagnosis and adjuvant systemic therapy (hormone therapy and chemotherapy) substantially reduce the risk of distant metastases. However, adjuvant therapy has serious short- and long-term side effects and involves high medical costs. Accurate prognostic tests are therefore of great interest to aid clinicians in deciding which patients are at high risk of developing metastases and hence should receive adjuvant therapy. Microarray gene expression profiling has shown promise to allow for such prognostication based on the expression pattern of certain gene sets (often called ‘signatures’ in the gene expression literature). A number of papers have been published where different signatures have been reported [7, 11–14].

3.1.2 Protein microarrays

Protein-protein interaction elucidation is of substantial importance in biology and medicine. Proteins interact with each other in a highly specific manner, and their interactions play a key role in most cellular processes; in particular, the distortion of protein interfaces may lead to the development of many diseases, such as cancer and Alzheimer’s disease [15]. An integrated view of protein interaction networks is thus required to understand cellular processes and to be able to tackle many human disease conditions. Protein microarrays have emerged as a promising approach for deciphering these networks [15].

Protein microarrays work analogously to DNA microarrays, but a protein library is affixed on a glass slide instead of short segments of DNA. The immobilized library is probed with one or more target proteins and their interactions are analyzed with detection systems similar to the ones used for DNA microarrays. Protein microarrays provides a massively parallel approach to identifying protein-protein, protein-DNA, or protein-small molecule interactions and has a wide variety of applications, including clinical diagnostics and monitoring disease states. A nice overview of the protein microarray technology and its applications was given by Hall et al. [16].

In a recent study, protein microarrays were used to study PDZ protein domains’ interaction with peptides in the mouse proteome on a large scale [17, 18]. PDZ domains mediate protein-protein interactions and are of great interest since they play a key role in the development of multicellular organisms, in which PDZ domains are often found as components of scaffolding proteins involved in cell polarity and intercellular interactions [19]. The biological importance of PDZ domains is further underscored by the identification of various PDZ-containing proteins as human pathogens’ targets [20]. However, despite recognizing their importance and their potential as a drug target, the details of PDZ domains interaction with other proteins have to a large extent remained uncharacterized due to the practical difficulty to investigate a large protein domain family in its entirety. In the study by Stiffler et al. [17], protein microarrays were used to investigate the binding selectivity of 157 mouse PDZ domains with
respect to 217 genome-encoded peptides. PDZ domains have long been thought to cluster into discrete functional classes defined by their peptide-binding preferences. However, contrary to the current paradigm, PDZ domains were in this study found to be evenly distributed throughout the selectivity space, which suggested that they have been optimized across the proteome to minimize cross-reactivity. Stiffler et al. [17] concluded that focusing on families of interaction domains may facilitate the integration of experimentation and modeling and play an increasingly important role in future investigations of protein function.

3.1.3 High-throughput screening

Using robotics and automated data processing, high-throughput screening (HTS) allows researchers to quickly conduct millions of biochemical, genetic or pharmacological tests. Through this process it is possible to rapidly identify active compounds, antibodies or genes that modulate a particular biomolecular pathway. The results of these experiments often provide starting points for drug design and for understanding the interaction or role of a particular biochemical process in biology.

The data generated from HTS may be used to construct quantitative structure-activity relationships (QSAR) [21] or proteochemometrics models [22]. The studied entities (e.g. a protein and an interacting ligand) are in these methods numerically characterized by their physicochemical properties. The numerical characterization may subsequently be correlated to a response variable, such as interaction activity, toxicity, or a biological activity. Such regression models are of use in drug discovery and optimization processes, since they allow for prediction of novel (non-synthesized) chemical entities’ activity and for analyses of which parts of the molecules that are most important for the interaction to occur. Such information can aid in drug lead optimization as well as in determining the active site and the functionally important residues in a protein.

3.2 eScience

Scientific research is to an increasing degree carried out by communities of researchers spanning disciplines, laboratories, and national boundaries. These communities typically use geographically distributed and heterogeneous resources, such as experimental instruments, databases, computational systems, and software. The term eScience has emerged as a consequence of these recent developments of distributed and collaborative research, and refers to computationally intensive science conducted in highly distributed network environments [23]. The recent advances in eScience are made available due to the progress of standardization processes, the increasing awareness of interoperability with respect to data provisioning and software development, and the fast increase of computational power in computers. These advances can roughly be divided into the following three main components: semantic information, high-performance computing (HPC) facilities, and software services (Fig. 3.1).
Figure 3.1: Overview of the eScience concept. eScience seeks to develop the tools and content to support multidisciplinary, collaborative science. Its immediate aims are to find ways of sharing information in a form that is appropriate to all readers - human and machine - as well as to provide software tools and high-performing computational power to integrate, visualize, and analyze the information. Scientists interact over a network cloud, typically the Internet, with collaborators (denoted Et al. in the figure). Software services, databases, and high performance computational resources are all available over the network when needed.

3.2.1 Semantic information

Standardization aims at making information self-contained in the sense that a minimum set of meta-data (data about the data) is available to make it understandable and reproducible. To adhere information with meta-data, a crisp and unambiguous explanation of the meaning of the meta-data is needed. Ontologies (controlled vocabularies) play an important role in this process, since they ensure that a term or a word has a well-defined and crisp meaning. Standardization allows information to become semantic, i.e. that it contains a meaning rather than just being a bag of numbers or words.

In the context of high-throughput biology, standardization affords straightforward use of data and information from different research groups and from heterogeneous sources, since it is made available in a format that is semantically well-defined and thus permits comparisons, reproduction, validation, and analyses. A number of initiatives within the high-throughput biology and bioinformatics communities aim at producing data standards for high-throughput biological experiments. For example, the MGED consortium has developed the Minimum Information About a Microarray Experiment (MIAME) standard [24]. Another example is the Human Proteome Organization’s Proteomics Standards Initiative (HUPO-PSI) to establish standards and controlled vocabularies in proteomics, for instance the Minimum Information About a Proteomics Experiment (MIAPE) [25].

3.2.2 High-performance computing

High Performance Computing (HPC) denotes the infrastructure of hardware and middleware that provides scientists with the computational power, memory, and storage capacity to carry out analyses that are not feasible on personal computers. The term includes parallel computers, distributed computation, GRID-computing, and other systems with large computational power, memory and storage capacity. HPC is used to analyze large quantities of data, perform high-end simulations, and enables detailed
studies of complex problems. The handling of data from high-throughput instruments largely relies on HPC for processing and storage.

3.2.3 Software services
The exponentially increasing amount of scientific data and information that is being published and deposited in public databases means that it is no longer feasible for a researcher to manually keep track of all the information that may be relevant and of use to her. Information must be gathered, sifted and presented to the scientist in an automated and personalized format. Since semantic information has a well defined meaning, it allows computers to conduct these data retrieval and processing tasks using a service-oriented architecture (SOA). SOA makes heavy use of Web services, which are software systems that provides machine-machine interoperable services over a network (often the Internet). Interoperable Web services, available via machine-accessible interfaces, are changing how computational analyses of high-throughput biological data are performed and enables novel research by giving scientists access to resources held on widely dispersed computers, as if they were on their own desktops. The resources can include data collections, very large-scale computing resources, scientific instruments and high performance visualisation.

3.3 Model selection
Paper I-IV in this thesis concern regression models and model selection in a regression setting. Thus, although the model selection problem comes into any modeling situations, I will here introduce it in a regression context.

Suppose that we observe the \( n \times (p + 1) \) data matrix \((X, y)\) from the statistical model

\[
Y = f(X) + \epsilon,
\]

where the random error \( \epsilon \) has \( \text{E}(\epsilon) = 0 \) and is independent of \( X \). We seek a model \( g(X, \beta) \) that approximates \( f(X) \) and which we can use for predicting \( Y \) from values of the input \( X \). Given the dataset \((X, y)\), we want to estimate the parameter vector \( \beta \) in the model \( g(X, \beta) \) to obtain a useful approximation of \( f(X) \). I will often (when no confusion can arise) denote a model by \( g(X) \) and its estimate by \( \hat{g}(X) \), thus omitting the parameter vector for convenience. Analogously, the realizations will often be denoted \( g(x) \) and \( \hat{g}(x) \).

3.3.1 The bias-variance tradeoff
Let \( L(Y, \hat{g}(X)) \) be a loss function for measuring errors between \( Y \) and \( \hat{g}(X) \). Throughout this thesis, we will assume squared error loss, i.e.

\[
L(Y, \hat{g}(X)) = (Y - \hat{g}(X))^2,
\]

which is computationally convenient and by far the most commonly used [26]. We define the generalization error (or test error) \( R_g \) to be the expected prediction error.
over an independent test sample:

\[
R_g = E[(Y - \hat{g}(X))^2] = \int E[(Y - \hat{g}(x))^2 | X = x] \, p(x) dx, \tag{3.3}
\]

where \(p(x)\) is the probability density function of \(X\). We can derive an expression for the expected prediction error of \(\hat{g}(x)\) at \(X = x\), according to:

\[
R_g(x) = E[(Y - \hat{g}(x))^2 | X = x] = E[(Y - g(x) + g(x) - E\hat{g}(x) + E\hat{g}(x) - \hat{g}(x))^2 | X = x] = \sigma_e^2 + [g(x) - E\hat{g}(x)]^2 + [E\hat{g}(x) - \hat{g}(x)]^2 = \sigma_e^2 + \text{bias}^2(\hat{g}(x)) + \text{var}(\hat{g}(x)).
\]

The first term is irreducible and cannot be avoided no matter how well we estimate \(g(x)\). The second term is the squared bias (the amount by which the average of our estimate differs from the true mean) and the last term is the variance (the expected squared deviation of \(\hat{g}(x)\) around its mean). In general, the more complex we choose the model \(g(X)\), the more able it is to adapt to a complex relationship between \(X\) and \(Y\) and the lower the bias of \(\hat{g}(x)\), but conversely, the higher the variance. In between there is an optimal model \(\hat{g}(x)\) that best balances the bias and the variance and gives a minimum generalization error (Fig. 3.2, red curve).

3.3.2 Controlling the model complexity

A common criterion for estimating \(g(X, \beta)\) is to minimize the residual sum of squares (RSS):

\[
\text{RSS}(g(x, \beta)) = \sum_{i=1}^{n} (y_i, g(x_i^T, \beta))^2 = (y - g(X, \beta))^T (y - g(X, \beta)). \tag{3.4}
\]
However, by choosing \( g(X, \beta) \) complex enough, we can make the RSS arbitrarily small (i.e. given any \( \nu > 0 \) we can always make RSS < \( \nu \) by making our model choice \( g(X, \beta) \) complex enough; Fig. 3.2, green curve). A model with RSS = 0 is overfit to the training data and will typically generalize very poorly. In order to obtain useful estimates \( \hat{g}(X) \) of \( g(X) \), we must constrain the eligible solutions to equation (3.4) to a smaller set of functions; by doing so we introduce a little bit of bias in the estimate in order to reduce the variance. Two very common ways of constraining the complexity of \( \hat{g}(X) \) are variable selection and regularization.

### 3.3.2.1 Variable selection

Variable selection constitutes a very important special case of the model selection problem where each model under consideration corresponds to a distinct subset of \( X \). Let \( \gamma \) index the subsets of \( X \) and let \( p_\gamma \) be the size of the \( \gamma \)th subset, the problem is to fit a model of the form

\[
y = g(X_\gamma) + \epsilon,
\]

where \( X_\gamma \) is an \( n \times p_\gamma \) matrix whose columns correspond to the \( \gamma \)th subset. Choosing a suitable subset \( \gamma \) of \( X_\gamma \) may substantially increase the prediction accuracy (i.e. reduce the generalization error). It also gives us more parsimonious models that are easier to interpret.

Variable selection is a rich and very active research field. The theory and the number of available methods is by far exceeding what is possible to give a comprehensive account of here; I refer to George [27] for a brief overview. Variable selection is a major topic of this thesis and I will return to it several times, most notably in Papers I-IV.

### 3.3.2.2 Regularization

Regularization controls the complexity by explicitly penalizing RSS(\( g(x) \)), according to

\[
PRSS(g(x), k) = RSS(g(x)) + kJ(g(x)),
\]

where \( J(g(x)) \) is chosen so that high variance of \( g(x) \) is penalized and \( k \geq 0 \) controls the amount of penalization. Many implementations of this idea exist, for example smoothing splines [28], ridge regression [29], and the lasso [30]. The two latter are very closely related and amounts to finding solutions to the following penalized ordinary least squares criteria:

\[
\hat{\beta}_{\text{ridge}} = \min_{\beta} (y - X\beta)^T(y - X\beta) + k\beta^T\beta
\]

and

\[
\hat{\beta}_{\text{lasso}} = \min_{\beta} (y - X\beta)^T(y - X\beta) + k||\beta||_1,
\]
respectively, where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ (i.e. the $L_1$-norm of $\beta$). The solution to the ridge criterion (3.7) is

$$\hat{\beta}\text{ridge} = (X^TX + kI)^{-1}X^Ty,$$  \hfill (3.9)

where $I$ is the $p \times p$ identity matrix. The solution thus adds a positive constant to the diagonal of $X^TX$ before inversion. This makes the problem nonsingular, even if $\text{rank}(X^TX) < p$. For the lasso criterion no closed form solution exists and we have to resort to numerical optimization techniques.

Both ridge regression and the lasso shrink the regression coefficients towards each other (and towards zero) by imposing a penalty on their size. However, the way that the shrinkage penalizes the regression coefficients is different in the two methods. Because of the nature of the lasso constraint, some coefficients are shrunk to exactly zero (see Tibshirani [30] for a more detailed explanation). The lasso thus works as a combined variable selection and shrinkage method and retains some of the favorable properties of each method.

Regularization techniques assume that $g(X)$ is continuous and exhibits a smooth behavior. This can be seen as that we use some prior belief of the nature of $g(X)$, which hints that regularization is related to Bayesian techniques (see Section 5.2). Indeed, regularization methods can usually be cast in a Bayesian framework (see e.g. [31]).

We use the ridge criterion extensively in Paper I and IV. The lasso plays a minor part in Paper I and III.

### 3.3.3 Choosing a model

The restrictions imposed on $g(X)$ introduced in Section 3.3.2 lead to unique solutions of equation (3.4). However, we can conceive infinitely many possible constraints that give us such solutions to (3.4), so we have merely transferred the model selection problem to choosing a suitable constraint. For example, if we exploit regularization, we need to determine the value of the penalty parameter $k$, and in the case of variable selection we need to decide upon which subset $\gamma$ to use. How do we choose the values of these parameters? And in general, how do we choose among different models $g(X)$? Below I briefly outline three methods for approaching these questions that are relevant for Papers I, II, and IV. Two of the methods are defined in terms of an information criterion, a mechanism that uses data to give each candidate model a score, which we may use to rank the candidates.

#### 3.3.3.1 Akaike’s information criterion

Akaike’s information criterion (AIC) [32] is an approach to model selection that in principle works in any situation where parametric models are compared. According to AIC we should choose a model $g(X, \beta)$ so that it maximizes the following criterion:

$$\text{AIC}(\hat{g}(x)) = 2l(\hat{g}(x)) - 2df(\hat{g}(x)),$$  \hfill (3.10)

where $l$ is the log-likelihood function and $df$ denotes the number of free parameters in a given model.
Analogously to how the RSS can become arbitrarily small by choosing \( g(X) \) complex enough, the log-likelihood is monotonously increasing with increasing model complexity. Directly comparing the attained log-likelihood maxima for different models does thus not suffice for choosing the best model. The second term in AIC punishes too complex models and serves the purpose of striking a tradeoff between model fit and complexity. The AIC method thus has an intuitive appeal in penalizing log-likelihood for complexity. However the choice of the penalty term is not obvious. Provided that the maximum likelihood estimator is used to estimate \( \hat{g}(X) \), it can be shown that the penalty term used in AIC assures that the distance between the chosen model and the true model is asymptotically minimized, as measured by the Kullback-Leibler divergence\(^1\) [34]. Further, within the linear model class and assuming squared error loss, AIC is efficient (conditional on the training data) [34]. This means that the AIC asymptotically with increasing \( n \) selects the model that minimizes the generalization error.

### 3.3.3.2 Bayesian information criterion

The Bayesian information criterion (BIC) [35] is very similar to AIC, but it penalizes the log-likelihood harder (provided that \( n > 8 \)). BIC is defined according to

\[
BIC(\hat{g}(x)) = 2l(\hat{g}(x)) - (\log n)df(\hat{g}(x)).
\]

The name implies that BIC is it related to Bayesian methods for model selection (see Section 5.2). A Bayesian procedure selects the model that is \textit{a posteriori} most likely. Schwartz [35] showed that as \( n \) tends to infinity, the BIC model choice will be the same as the Bayesian choice. However, BIC does not depend on any prior probabilities and is thus valid outside the Bayesian context. (The reason for this is that BIC is derived as the two first terms in a Laplace approximation of the posterior probability of a model. These two leading terms do not depend on the prior. See [35] for a derivation of BIC for models within the exponential family and [34] for a more general derivation.) BIC can be shown to be \textit{consistent}, which means that the BIC model choice, asymptotically in \( n \), is the least complex model that minimizes the Kullback-Leibler divergence. Thus, AIC is efficient and BIC consistent and both properties are attractive for model choice. Interestingly, it can be shown that these two properties can never be combined in the same information criterion [36].

### 3.3.3.3 Cross-validation

An approach to model selection that, at least superficially, is different in spirit from that of AIC and BIC is cross-validation [37]. A reasonable criterion is to choose \( g(X) \) so that the generalization error of \( \hat{g}(X) \) is minimized. However, the generalization error is unknown to us and we thus need to estimate it. Cross-validation implements this idea by directly estimating the generalization error according to following procedure: Randomly allocate the \( n \) observations in a dataset to \( K \) different subsets. Let \( \kappa: \{1, \ldots, n\} \rightarrow \{1, \ldots, K\} \) be an indexing function that indicates the partition to which observation \( i \) is allocated. Denote by \( \hat{g}^{-k}(x) \) the fitted model, computed with the \( k \)th part of the data removed. Then the cross-validation estimate of the generalization error

\(^1\)The Kullback-Leibler divergence is a measure of the difference between two probability distributions [33]
The cross-validation approach to model selection is simply to choose \( g(X) \) so that equation (3.12) is minimized. Common choices of \( K \) are \( K = n \), often called leave-one-out cross-validation, and \( K = 5 \) or \( K = 10 \), denoted five-fold and ten-fold cross-validation.

Intriguingly there exist connections between cross-validation and the AIC and BIC information criteria, despite their seemingly very different motivations. Shao [38] demonstrated that within the linear model, the leave-one-out cross-validation criterion is asymptotically equal to AIC and \( K \)-fold cross-validation, where \( K \) is chosen so that \( n_K/n \to 1 \) as \( n \to \infty \), where \( n_K \) is the number of observations in the left out sample.

### 3.3.4 Assessing a model

Having chosen a final model, model assessment amounts to estimating its generalization error on completely new data, a test set. The test set should thus be a set of observations independent of the data used for the model selection and fitting process, otherwise the test set error of the final chosen model will underestimate the true generalization error. Provided that we have a very large set of observations (i.e. \( n \) is large), the best way to assess a model is to at the onset of the data analysis set apart a sizable number of observations to use for assessment after the entire data analysis process has been conducted.
4. Motivation for the work in this thesis

High-throughput experimental technologies lay the data-collecting foundation for answering many complex biological and biomedical questions. They also present opportunities for the development of new clinical tests for diagnosing and prognosticating diseases, and for enhancing the drug development process. However, merely collecting huge amounts of data does not automatically provide new biological insights or clinical applications. Obtaining large quantities of data is not the same as obtaining large amounts of knowledge. In order to translate the wealth of data into concrete biological knowledge, new drugs, and clinical practices, it needs to be carefully analyzed. The statistical analysis of high-throughput data has therefore become of great interest and importance.

4.1 A motivating example: Predicting distant metastasis development in breast cancer patients from gene expression microarray data

As described in Section 3.1.1, there is a need for improved tests for breast cancer prognosis prediction. Microarray gene expression profiling of breast cancer tumors has shown promise to allow for improved prognostication based on the expression pattern of certain gene sets (signatures). A number of papers have been published where different signatures have been reported, see e.g. [7, 11–14]. However, the prognostic capacity of these signatures when applied to new data have been questioned in several studies, e.g. [39–41]. This criticism is founded in three observations of the published studies:

1. Signatures from different studies share almost no genes, suggesting that the signatures to a large degree depend on the datasets rather than being prognostic for breast cancer [39].
2. The assessment of several prognosis prediction models has been inadequate, since they have relied on a single, very small, test set to estimate the generalization error (see e.g. [7, 11]). In fact, Michiels et al. [40] demonstrated that several of the reported signatures did not perform better than random guessing when a more sturdy assessment of the generalization error was conducted.
3. The estimates of the generalization error has typically not been reported with confidence intervals (see e.g. [7, 11]).

The first point relates to the model selection problem. The theory behind AIC, BIC, and cross-validation is derived as asymptotic results when the number of observations tends to infinity. Here we have a situation with roughly 20,000 variables and only a
few hundred very noisy observations (at best). How do we attack the model selection problem from a computational and statistical point of view in such a setting?

Points number two and three relate to model assessment. Relying on a single small test set for generalization error estimation is inadequate, as demonstrated by for instance Ein-Dor et al. [39] and Michiels et al. [40]. The assumption behind test set assessment relies on having a large enough set of observations to make the variance of the generalization error estimate reasonably small. This assumption is typically not fulfilled when analyzing gene expression microarray data. Furthermore, as emphasized by Wickenberg-Bolin et al. [42], Isaksson et al. [43], and Dobbin [44], any published generalization error estimate should be reported with a "useful" (i.e. not overly conservative) confidence interval to be of any practical value. The questions are however: How do we produce better estimators of the generalization error with the limited number of observations we have available when analyzing high-throughput data? And how do we adhere this estimate with an accurate confidence interval?

These issues are of adamant importance. If we want to use high-throughput data in critical applications, such as breast cancer prognostication, the models we construct need to reflect the underlying biology and not just artifacts and hypotheses suggested by the data. To gain acceptance in applications where the subsequent decisions have serious consequences, we need to have a clear picture of how much we can trust predictions from the models and of the risks of making errors of type I and II (i.e. the risk of false positives and false negatives, respectively).

4.2 High-throughput biology poses new challenges for model selection and assessment

From the discussion in Section 4.1 it is clear that in particular the large number of variables in relation to the number of observations in high-throughput biological datasets presents new problems regarding model selection and assessment. In addition, a number of other properties of high-throughput data complicates the situation even further:

1. High-throughput data is typically very noisy, which means that the independent variables in our models almost always contain errors.
2. The data often contain grouping structures (clusters).
3. There is in general a nonlinear relationship between the independent and dependent variables.
4. The size of the data makes the analysis computationally very expensive.

The main idea in this thesis is that eScience equip us with new tools for addressing the problem of model selection and assessment when analyzing high-throughput data. Briefly: eScience allows us to calculate more, and to make use of what we already know when approaching a new modeling problem.

To put this in a bit more detail: High-performance computing resources permit us to apply computationally demanding methods even to the large modeling problems that arises when analyzing high-throughput data. Further, huge amounts of scientific data and information has over the years been collected and made publicly available on the Internet. This information is to an increasing degree being standardized and semantically described. In combination with the advent of interoperable machine-to-machine techniques, such as Web services, this means that we readily can harvest the
Internet for already available information to "guide" our modeling approaches when analyzing a new dataset.

4.3 Aims

The main objective of the papers included in this thesis is to develop novel methods for model selection, primarily variable selection, that leverage on the possibilities in terms of computational power and distributed information that is provided by eScience. In particular, the aims are to:

- Investigate automated methods for searching the model space.
- Exploit the opportunities provided by standardized data and Web services to retrieve prior information and use the information for guiding model selection.
- Development of variable selection methods that are insensitive to noisy independent variables and nonlinear relationships between independent and dependent variables.
- Develop a framework for deployment of novel methods to make them accessible to end users (i.e. the experimental scientists).
5. Methods

5.1 eScience methods

5.1.1 Data integration
Data integration is the process of combining heterogeneous data or data residing in different sources. Interoperable Web services are extremely useful for performing this task due to their modularity, reusability, ease of maintenance, standardized ways of interaction, and simplicity. In Paper II we make extensive use of Web services for data integration and data processing. In short, we combine protein sequence information with structural information from the Protein Data Bank [45] by using the Sequence Annotated by Structure Web service [46, 47], and we merge gene expression microarray data with previously acquired gene data by text mining all free fulltext articles in PubMed. The latter Web service workflow is outlined in Figure 5.1(a) with a bit more detailed description.

5.1.2 High-performance computing
In Paper I, II, and IV the high-performance computing facility at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) was used for calculations (Fig. 5.1(b)). The UPPMAX resources consist of several computer clusters, some comprising up to 800 processor cores. The UPPMAX resources are located on a private network (one has to be an Uppsala University employee to use the resources) and are accessed via a secure shell connection.

5.1.3 Optimization and search algorithms
Paper I makes heavy use of two different optimization methods: a genetic algorithm (GA) and a brute-force method.

5.1.3.1 Genetic algorithms
A genetic algorithm (see [49] for a comprehensive treatment) is a stochastic search technique for finding exact or approximate solutions to optimization and search problems. A typical genetic algorithm is defined by a genetic representation of a given solution (normally termed a chromosome in the GA context). This mean that a vector \( z^u_i \) specifies the numerical representation of the \( i \)th chromosome at generation \( u \), and an objective function (or "fitness function"), \( h(z^u_i) \to \mathbb{R} \) evaluates the fitness of a chromosome. The GA is initiated by setting up a random population that contains \( N \) number of trial chromosomes. New solutions are generated by mutation or recombination of existing solutions and are selected for the next generation with a probability
Figure 5.1: Overview of the data integration used in Paper II. (a) Data integration via Web services. Using the Gene Expression Omnibus (GEO) Web service [48] we downloaded five breast cancer gene expression datasets and their associated clinical data, which we integrated with previously acquired information about the genes association with breast cancer. To do this we used three Web services in conjunction: NetPath (http://www.netpath.org/), DictService (http://services.aonaware.com/) and Entrez Utilities (http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/v2.0/DOC/esoap_help.html). NetPath is a curated resource containing genes that in the literature have been reported as being transcriptionally regulated by cancer-signalling pathways. All genes in the downloaded array data that were listed in NetPath were retrieved using the NetPath Web service. The retrieved genes were used to text mine PubMed for all free fulltext articles where the gene name was mentioned in combination with breast cancer using the Entrez Utilities Web services, after first discarding genes with names that represented English words (using the dictionary definition Web service DictService) to reduce the number of spurious hits. (b) Computation on a high-performance computer. To use the large quantities of data collected in workflows such as the one outlined in (a), access to HPC resources is convenient, sometimes crucial.
given by
\[ \frac{h(z'_i)}{\sum_{i=1}^{N} h(z'_i)}. \tag{5.1} \]

The process is continued through a number of generations until an optimal or acceptable solution has been found. Genetic algorithms of this type can be shown to converge with a probability 1 to the global optimal solution as \( u \to \infty \) [50].

### 5.1.3.2 Brute-force search

A brute force search systematically tests an exhaustive list of all possible candidates for the solution to a given search or optimization problem and checks whether each candidate satisfies the problem’s statement.

### 5.2 Bayesian statistics

Mathematical statistics uses two major paradigms: the frequentist and the Bayesian. The frequentist paradigm evaluates procedures based on imagining repeated sampling from a particular model (the likelihood), which defines the probability distribution of the observed data conditional on unknown parameters. Properties of the procedure are evaluated in this repeated sampling framework for fixed values of the unknown parameters. The Bayesian approach requires a sampling model and, in addition, a prior distribution on all unknown quantities in the model. The prior and the likelihood are used to compute the conditional distribution of the unknowns given the observed data (the posterior distribution).

In the Bayesian approach, in addition to specifying the model for the observed data \((X, y)\) given a vector of unknown parameters \(\beta\), usually in the form of a probability distribution \(p(x, y \mid \beta)\), we thus suppose that \(\beta\) is a random quantity as well, having a prior distribution \(\pi(\beta)\). Inference concerning \(\beta\) is then based on its posterior distribution, given by Bayes’ theorem:

\[ p(\beta \mid x, y) = \frac{p(x, y, \beta)}{p(x, y)} = \frac{p(x, y, \beta)}{\int p(x, y, \beta) d\beta} = \frac{p(x, y \mid \beta) \pi(\beta)}{\int p(x, y \mid \beta) \pi(\beta) d\beta}. \tag{5.2} \]

Note that the denominator in equation (5.2) does not depend on \(\beta\), and we may thus express equation (5.2) according to

\[ p(\beta \mid x, y) \propto p(x, y \mid \beta)\pi(\beta), \tag{5.3} \]

or in words: "the posterior is proportional to the likelihood times the prior".

The Bayesian approach has several advantages over frequentist statistical methods. We exploit three of these advantages in Paper II, namely:

1. Bayesian methods may be applied to problems whose structure is too complex to be dealt with using classical statistical methods [51].
2. Bayesian methods correctly summarize the predictive uncertainties of a model [52].
3. Bayesian methods provide the possibility of formally incorporating prior information in the data analysis [53].

The points above are of obvious importance in the analysis of high-throughput data: (1) Allows us to make Bayesian models increasingly complex to accommodate for the complexity of biological questions. One example is to accommodate for the grouping structures in the data which are often encountered in the analysis of high-throughput data (see Section 4.1). (2) Means that we can obtain reliable and "useful" Bayesian confidence intervals (often denoted credible intervals) of the estimates of model performance (see e.g. Isaksson et al. [43]). (3) The impact of this point is tremendous when modeling high-throughput data, since it allows us to coherently make use of the prior information that we can collect with eScience techniques (see Section 5.1.1).

5.2.1 Markov chain Monte Carlo methods

As seen from equation (5.3), Bayesian methods amount to the determination of posterior distributions, which boils down to the evaluation of complex, often high-dimensional integrals. In all but the simplest model settings these integrals are analytically intractable and thus need to be tackled numerically. Over the last two decades Markov chain Monte Carlo (MCMC) methods have been extensively used for this task.

Suppose we wish to compute a complex integral

$$\int_{a}^{b} h(x)dx$$  \hspace{1cm} (5.4)

If we can decompose $h(x)$ into the product of a function $f(x)$ and a probability density function $p(x)$ defined over the interval $[a, b]$, then

$$\int_{a}^{b} h(x)dx = \int_{a}^{b} f(x)p(x)dx = E_{p(x)}[f(x)],$$  \hspace{1cm} (5.5)

which we, if we can sample a large number of observations $\{x^1, ..., x^n, ..., x^U\}$ from $p(x)$, can approximate arbitrarily well according to

$$E_{p(x)}[f(x)] \approx \frac{1}{U} \sum_{u=1}^{U} f(x^u)$$  \hspace{1cm} (5.6)

by virtue of the law of large numbers. This is referred to as Monte Carlo integration. However, a problem with Monte Carlo integration is that it may be difficult to obtain samples from some complex probability distribution $p(x)$. Attempts to solve this problem are the roots of MCMC methods. There exists several MCMC-methods; here focus on describing the Metropolis-Hastings (MH) algorithm [54, 55], which is the most general MCMC-algorithm (other popular methods, such as Gibbs sampling [56], are special cases of the MH algorithm [57]) and which is the algorithm we use for fitting the Bayesian models in Paper II.
Given a symmetric proposal distribution \( q(x, y) \) and an arbitrary starting value \( x^0 \), the Metropolis-Hastings algorithm is as follows:

1. Given \( x^{u-1} \), generate \( \tilde{x} \sim q(x^{u-1}, x) \).
2. Compute
\[
\rho(x^{u-1}, \tilde{x}) = \min \left( \frac{p(\tilde{x})/q(x^{u-1}, \tilde{x})}{p(x^{u-1})/q(\tilde{x}, x^{u-1})}, 1 \right).
\]
(5.7)
3. With probability \( \rho(x^{u-1}, \tilde{x}) \), accept \( \tilde{x} \) and set \( x^u = \tilde{x} \); otherwise reject \( \tilde{x} \) and set \( x^u = x^{u-1} \).

It is fairly easy to demonstrate that the Metropolis-Hastings algorithm generates an aperiodic and irreducible Markov chain whose stationary distribution is the distribution \( p(x) \) [57], and hence will converge almost surely [58]. However, a key issue in the successful implementation of Metropolis-Hastings (or any other MCMC sampler) is the number of runs (steps) until the chain converges. Typically the first few thousand samples are discarded (the burn-in), and a test for convergence is used to assess whether stationarity has indeed been reached. In Paper II we used the potential scale reduction factor [59] for checking convergence.

5.2.2 Bayesian model selection and model averaging

The standard Bayesian approach to model selection is to choose the model with the highest Bayes factor (BF):
\[
BF = \frac{p(X, y \mid g_n)}{\sum_{m=1}^M p(X, y \mid g_m)},
\]
(5.8)
where \( \{g_1, \ldots, g_M\} \) is a set of candidate models and
\[
p(X, y \mid g_m) = \int p(X, y \mid g_m, \beta) \pi(\beta_m) d\beta_m,
\]
(5.9)
where \( \beta \) is the parameter vector related to model \( g_m \). To choose a single model, e.g. by maximizing the Bayes factor, is reasonable in some problem domains where there may be underlying scientific reasons why one particular model should be preferred in a model selection process. However, when analyzing high-throughput data, the model selection is typically rather arbitrary in that a (usually) large number of models may fit equally well. To choose one single model as the “correct” one may be suboptimal, since further inference conditional on the chosen model would ignore the uncertainty in the model selection process itself [53]. This difficulty can be handled by averaging over a number of models, called model averaging, which we use in Paper II. For example, the variance in generalization error estimates can sometimes be reduced substantially by this technique [26].

Model averaging is straightforward in a Bayesian setting. Let \( \delta \) some quantity of interest (e.g. a vector of regression coefficients). Given a set of prior probabilities
\( \{\pi(g_1), \ldots, \pi(g_M)\} \) for each model, the posterior distribution of \( \delta \) is

\[
p(\delta \mid X, y) = \sum_{m=1}^{M} p(\delta \mid g_m, X, y)p(g_m \mid X, y), \tag{5.10}
\]

where \( p(\delta \mid g_m, X, y) \) is the posterior for \( \delta \) under the \( m \)th model, and \( p(g_m \mid X, y) \) is the posterior probability of \( m \)th model, computable according to

\[
p(g_m \mid X, y) = \frac{p(X, y \mid g_n)\pi(g_n)}{\sum_{m=1}^{M} p(X, y \mid g_m)\pi(g_m)}, \tag{5.11}
\]

where \( p(X, y \mid g_m) \) is the marginal distribution of the data under the \( m \)th model, given in equation (5.9). The posterior probability in equation (5.10) is in general readily approximated with MCMC-methods, such as the Metropolis-Hastings algorithm.

### 5.3 Disturbing datasets by adding noise

Resampling techniques, such as the bootstrap [60, 61], for disturbing datasets to estimate properties of the sampling distribution of an estimator has existed in statistics for some time. The SIMEX procedure, a novel way of exploiting disturbed datasets, was introduced by Cook and Stefanski in [62]. SIMEX controls for the attenuation effect of least squares parameter estimates in errors-in-variables models by perturbing the dataset by adding noise to the independent variables (as opposed to resampling like in bootstrapping). Consider the errors-in-variables model

\[
y_i = g(x^T_i, \beta) + \varepsilon_i, \tag{5.12}
\]

where \( i = 1, \ldots, n \) and \( x^T_i \) denotes the true but unobserved value of the independent variable. Instead we observe \( x^T_i \) with an error:

\[
w^*_i = x^T_i + \eta_i, \tag{5.13}
\]

where the measurement error \( \eta_i \) is assumed to be independent of \( x^T_i \) and \( E(\eta_i) = 0 \). It can be shown that the standard least squares estimates are biased in this setting [63]. SIMEX corrects for this bias in the following way: The measurement error \( \eta_i \) is increased by adding pseudo errors to one independent variable at the time according to

\[
w^*_j = w_j + \sqrt{\lambda} \varepsilon^*, \]

where \( \lambda \in \{\lambda_1, \ldots, \lambda_N\} \) is a sequence of positive numbers, and \( \varepsilon^* \) is a vector of pseudo errors with \( E(\varepsilon_i^*) = 0 \), independent and identically distributed from a symmetric distribution with mean zero, unit variance, and finite fourth moment. For every value of \( \lambda \) we fit a model and record the model parameter estimates. Given the parameter estimates for all values of \( \lambda \), it is possible to extrapolate backwards to obtain an asymptotically unbiased estimate of the model parameters [62].
5.3.1 Using datasets disturbed by noise in model selection

The SIMEX idea of disturbing data by introducing noise has sparked a number of approaches for utilizing this idea in a model selection setting. We can roughly divide these approaches into two main categories:

1. The first approach is based on adding pseudo errors to the dependent variable. Variable selection methods can be very sensitive to small changes in the data [30]. To control for this sensitivity, Breimann [64] proposed to repeatedly add pseudo errors to the dependent variable and perform the variable selection. By averaging over the selected independent variables in each repetition, he showed that an unstable variable selection procedure could be stabilized (compare model averaging, Section 5.2.2). Shen and Ye [65] used addition of random errors to the response variable to estimate the model degrees of freedom. In [66], Lou et al. added noise with stepwise increasing variance to the response variable in order to tune parameters in variable selection methods.

2. In the other group of approaches new pseudo variables (i.e. a simulated variable containing only noise) are generated and added to the dataset. In [67], Wu et al. proposed this method for controlling the false selection rate, i.e. the proportion of unimportant variables included in a selected model. This method was further developed by Johnson [68].

In contrast to these approaches, in Paper III and IV we study the effect of adding controlled amount of noise to the independent variables to aid variable selection. The idea is founded in the intuition that adding noise to an unimportant independent variable should not affect the residual sum of squares of a model fitted to the data.

5.4 Rich clients

A rich client is a computer program in client-server architecture networks that provides rich functionality independently of the central server. In contrast to web-based systems, rich clients run on the local computer and hence take full advantage of modern laptops and workstations and allows for tight integration with the operating system. However, they still have the option to invoke and use remote services and resources (e.g. databases, software services, and HPC resources). The rich client concept has been implemented in rich client platforms (RCP), which are frameworks of a minimum set of functionality that can be used for the development of new software. One of the most used and furthest developed RCPs is the Eclipse rich client platform (http://wiki.eclipse.org/index.php/Rich_Client_Platform), which is used extensively in Paper V.
6. Results and discussion

6.1 Papers I and II

Papers I and II concerned the development of the $C_1C_2$ and "eScience-Bayes" frameworks for simultaneous model selection and assessment. These papers aimed to improve on standard methods for selecting a model and assessing its generalization error (i.e. to address the points (1), (2), and (3) in Section 4.1). In addition, Paper II also dealt with the grouping structure often exhibited by high-throughput data (i.e. point (2) in Section 4.2).

6.1.1 Model selection

Both the $C_1C_2$ and the eScience-Bayes frameworks used a complete separation of model selection from model assessment, which was achieved by reiterating a partitioning of the dataset into a training and a test set (used for model selection and assessment, respectively; see Fig. 6.1). However, the way that the model selection problem was attacked in the two papers differed quite substantially. As discussed in Section 4.2, eScience provides us with tools that can be used to improve our chances to make a good model choice, namely: (1) increased computational power, and (2) methods for accessing and retrieving prior information about the data analysis problem under study. In Paper I we exploited the first of these tools, whereas in Paper II we used both of them.

In Paper I we let optimization algorithms (a genetic algorithm and a brute-force method) traverse the model space and search for the best model (according to the Bayesian information criterion), conditional on the training set in each iteration. In Paper II we used Web services to retrieve background information about which variables that a priori were likely to be important (see Fig. 5.1). This information was used to elicit prior distributions reflecting our a priori belief about each variable's importance, and these distributions were then used to fit Bayesian regression models. The prior information was used to "guide" a Metropolis-Hastings algorithm to sample from the posterior distribution of models (see equation (5.10)) which we a priori considered likely. Hence, in Paper II we in fact used a combination of model choice (via variable selection) and model averaging, the rationale being that since there was a large number of models which were all reasonably plausible and we had no reason to trust any one of them to be the "correct" model. The use of prior information to aid model choice helps to guard against type III errors (i.e. hypotheses suggested by the data). When analyzing high-throughput data, we often have a large number of variables in relation to the number of available observations, which increases the risk of model overfitting and inflated generalization errors since the risk of chance correlations between independent variables and response is high. By using...
Figure 6.1: Schematic overview of the C\textsuperscript{1}C\textsuperscript{2} and eScience-Bayes frameworks for simultaneous model selection and assessment presented in Papers I and II. The red box contains the data partitioning that ensures that the model selection (blue box to the left) always is independent of the model assessment (blue box to the right). The oval represents that the procedure of data partitioning, model selection and assessment is repeated a number of times. To aid the model selection we have networked computational power via HPC resources and prior information, which is available in public repositories in the Internet and retrievable via Web services (WS) (Fig. 5.1).

Previously acquired scientific knowledge about the importance of each variable, the data analysis can be specified to give a higher chance of including the \textit{a priori} important variables while still allowing \textit{a priori} irrelevant variables to enter the model if they are strongly supported by the data (see Section 6.1.5 and Paper II for more details). This reduces the risk of detecting spurious correlations since we rely on hypotheses about the importance of each variable which were formulated \textit{before} seeing the data.

Despite the obvious differences in the approach to model selection used in Paper I and II, there are also a few similarities. For example, we use a similar method for walking through the model space: A genetic algorithm is in fact \textit{almost} an MCMC-methods (see e.g. [69]). When using the GA we look for one "best" model (conditional on the training data in each iteration). This \textit{roughly} equates to finding the mode of the posterior distribution in equation (5.10) (as $n \rightarrow \infty$), whereas we with the Metropolis-Hastings algorithm sample from the full posterior distribution.

6.1.2 Generalization error estimates

The C\textsuperscript{1}C\textsuperscript{2} and the eScience-Bayes frameworks addressed the problem of estimating the generalization error in small-sample problems in the same way, which was done by splitting the dataset into training- and test set, and using the training set for the \textit{entire} model selection and estimation process, and the test set \textit{only} for model assessment (i.e. a complete separation of model selection from model assessment). To avoid the problem discussed in Section 4.1 of basing the generalization error estimate on a single, small test set, this process was repeated (Fig. 6.1). This idea is far from new. In fact, it was introduced already in 1974 in a paper by Stone [37], where
cross-validation was first formally discussed. However, it seems as though it has been somewhat forgotten when analyzing high-throughput data (see the discussion in e.g. [70–72]), which may have led to quite substantial positive biases in the generalization error estimates in many studies [26].

6.1.3 Confidence intervals of generalization error estimates

In Paper I we used the different hold-out estimates of the generalization error obtained from repeatedly partitioning the dataset to construct confidence intervals, whereas in Paper II we used Bayesian credible intervals. It was shown in Isaksson et al. [43] that confidence intervals based on hold-out samples in small-sample problems are too wide due to the small test sets used. The observed variance between the hold-out estimates are inflated and might be completely dominated by the variance contribution from the small test sets used. Isaksson et al. [43] recommended using a Bayesian approach as the best available method. The method used in Paper II for constructing confidence intervals of the generalization error estimate thus represents an improvement compared to Paper I.

6.1.4 Testing of the $C_1^1C_2$ framework

The $C_1^1C_2$ framework was tested on simulated data and on a quantitative structure-activity relationship dataset (see Section 3.1.3) published by Selwood et al. [73]. Two different optimization methods, a genetic algorithm and a brute-force method, were used to search for an optimal model according to the Bayesian information criterion within a linear class of models estimated using the ridge criterion. The optimization methods were used to select a variable subset and for determining the level of penalization (i.e. the parameter $k$ in equation (3.7)).

The $C_1^1C_2$ framework was found to perform well at finding the true model in terms of choosing the correct variable subset and producing reasonable choices for the penalizing parameter. This was also found to be true in rather complex situations, like when the independent variables were highly correlated and when the number of observations was less than the number of variables. The $C_1^1C_2$ framework was also found to give accurate estimates of the generalization error. Using the genetic algorithm worsened the model choice, but not the generalization error estimates, compared to using the brute-force method. This observation may be explained by the fact that more variables were included in the model when the GA was used than compared to the brute-force method. I.e., in the test performed in Paper I, the chance of including the correct variable subset seem to have been roughly the same when using the GA and the brute-force method, but an unnecessarily large subset was included when the GA was used.

6.1.5 Testing of the eScience-Bayes framework

The eScience-Bayes approach was demonstrated on the problems of breast cancer prognosis prediction from transcriptomic data (see Sections 3.1.1 and 4.1) and of studying PDZ domain-peptide interactions based on proteomic data (see Section 3.1.2). In the breast cancer demonstration, we modeled whether development of
distant metastases occurred within five years or not from the gene expression data. In
the PDZ domain demonstration, we aimed to predict whether or not a given pair of a
PDZ domain and a peptide interacts or not. Both problems thus has a binary response
variable and were modeled using similar multilevel Bayesian probit regression
models (see Section 1.4), according to (somewhat simplified, see Paper II for details):

\[
\begin{align*}
Pr(y_{ki} = 1) &= \Phi(x^\gamma_{ki} \beta_k), \\
\beta_k^\gamma &\sim N_p(0, \Sigma^\gamma), \\
\Sigma^\gamma &\sim W^{-1}(J^\gamma, \text{rank}(J^\gamma) + 1), \\
\sigma_j^2 &\sim (1 - p_j) \delta(0) + p_j \cdot N(0, \tau^2)
\end{align*}
\]

(6.1)

where \( Pr \) denotes probability, and \( \Phi \) is the normal cumulative distribution function,
\( N_p \) the multivariate normal distribution, \( \delta \) the one-point distribution, and \( W^{-1} \) the
inverse Wishart distribution. \( n_k \) denotes the number of observations in group \( k \). \( J^\gamma \)
is the diagonal matrix with the non-zero elements in the vector \((\sigma_j^2)_{j=1,\ldots,p}\) as diagonal
elements. Similarly, \( x^\gamma_{ki} \) and \( \beta_k^\gamma \) are the elements of \( x_{ki} \) and \( \beta_k \), respectively, that cor-
respond to the non-zero elements in \((\sigma_j^2)_{j=1,\ldots,p}\) (\( x_{ki} \) relates to observation \( i \) in group
\( k \) and \( \beta_k \) are the regression coefficients related to group \( k \)). \( K \) is the total number of
groups in the data matrix \( X \) and \( \tau \) is some parameter.

The groupings in the datasets in each demonstration correspond to different labs
producing the gene expression data and to the fact that different PDZ domains are
differently promiscuous binders, respectively. The groupings give rise to a correlated
error structure when modeling the data, which is an assumption violation in standard
(non-multilevel) models. Model 6.1 controls for this effect, which in general mean
that we get better estimates of the model parameters [52]. In addition, it also means
that we analyze the data on different levels and thus get more information from the
modeling (see Paper II for more details).

The "guiding" of the Metropolis-Hastings algorithm discussed in Section 6.1.1 is
accomplished by the last line in the model specification. The parameter \( p_j \) controls
the probability of a given variable \( j \) to be included in the model or not. The more
prior information supporting the importance of variable \( j \), the higher the \( p_j \) and thus
the higher the probability of variable \( j \) to enter the model. In Paper II, we showed that
the gain from using prior information was significant and substantial (Fig. 6.2).

6.2 Papers III and IV

In Papers III and IV we investigated the problem of variable selection in quite general
models, such as nonlinear errors-in-variable models (see equations 5.12 and 5.13),
thus addressing points (1) and (3) in Section 4.2. We introduced the SimSel approach
to variable selection in Paper III and further developed it in Paper IV.

The principle of SimSel is to fit an approximative model that we use as a "gauge"
for determining which independent variables that are important for modeling the re-
response. We study how the quality of the approximative model’s fit is affected when
the independent variables are disturbed by pseudo errors, the main idea being that dis-
Figure 6.2: Receiver operating characteristics (ROC) curves, i.e. true-positive fraction (TPF) plotted as a function of the false-positive fraction (FPF) \( \text{TPF} = \frac{TP}{TP + FN} \) and \( \text{FPF} = \frac{FP}{FP + TN} \), where TP=true positives, FP=false positives, FN=false negatives, and TN=true negatives. Red lines represent that prior information was used and green lines that prior information was not used (see Paper II for information about the blue lines). The area under the curves (AUC) are 0.76 (0.71; 0.80) and 0.57 (0.55; 0.58), respectively (numbers in parentheses show the Bayesian credible intervals, shown as dotted lines in the figure).

turbing unimportant variables does not deteriorate the quality of the fit. To determine which variables that are significantly important, we compare the effect of disturbing the independent variables against the effect of adding errors to a simulated pseudo variable that is known to be unimportant.

Consider the model \( y = g(X_\gamma) + \varepsilon \) (see Section 3.3.2.1). We are interested in the subset \( \gamma \) of the variables \( x_1,...,x_p \) that are important for modeling \( y \). Let \( z \) be a simulated pseudo variable that is independent of \( X \) and \( y \), and let \( H(X, z, \beta) \) be a quadratic approximation of \( g(X_\gamma) \) that includes the linear, mixture, and quadratic terms of the independent variables including the pseudo variable. Set

\[
H(x_1,...,x_p,z,\beta) = H_\beta.
\]

Now, assume without loss of generality that \( x_1 \) is the variable under consideration to include in our model. For each positive parameter \( \lambda \in \{\lambda_1,...,\lambda_K\} \) we generate independent pseudo errors \( \varepsilon^* = (\varepsilon_1^*,...,\varepsilon_n^*)^T \) and add them to \( x_1 \). Set

\[
H(\lambda)\beta = H(x_1 + \sqrt{\lambda}\varepsilon^*,...,x_p,z,\beta)
\]

and minimize the residual sum of squares:

\[
\text{RSS}(\lambda) = \min_{\beta} \|y - H(\lambda)\beta\|^2,
\]

where \( \| . \| \) denotes the Euclidean norm in \( \mathbb{R}^n \). We now determine whether the variable \( x_1 \) is important by studying \( \text{RSS}(\lambda) \): if \( \text{RSS}(\lambda) \) is increasing with increasing \( \lambda \) we
conclude that $x_1$ is important and if $\text{RSS}(\lambda)$ is constant with increasing $\lambda$ we classify $x_1$ as unimportant (Figure 6.3).

The SimSel idea is implemented in the SimSel algorithm, where the procedure outlined above is repeated $M$ number of times (typically around 100). This enables us, for each variable $x_1, \ldots, x_p$, to estimate the distribution of the $F$-statistic of the regression line depicted in Figure 6.3. These estimated distributions may be compared with the estimated distribution of the $F$-statistic related to the pseudo variable, which is known to be unimportant. The comparison is implemented in the SimSel algorithm as the heuristic simulation test illustrated in Figure 6.4 (see Paper III and IV for details).

The SimSel method is founded on the following theoretical result, which was derived in Paper III: Under the assumption that $(H^TH)^{-1}$ exists, it holds that

$$\frac{1}{n} \text{RSS}(\lambda) = \frac{1}{n} \text{RSS} + \hat{\beta}^T_{(1)} D(\hat{\lambda}) \hat{\beta}_{(1)} + o(1) + o_{P^*}(1), \quad (6.2)$$

where $\hat{\beta}_{(1)}$ is the vector of regression coefficients related to $x_1$ (see Paper III for a proof). The remainder terms $o(1)$ and $o_{P^*}(1)$ tends to zero for $n \to \infty$ with probability 1 and in probability $P^*$, where $P^*$ is the distribution of the pseudo errors, respectively. Thus, if $\hat{\beta}_{(1)} = 0$, the left side of equation (6.2) is constant with increasing $\lambda$ (compare the cyan line in Fig. 6.3), whereas if $\hat{\beta}_{(1)} \neq 0$, the left side of equation (6.2) increases with increasing $\lambda$ (compare the red line in Fig. 6.3). In Paper IV, we derived a similar expression where the ridge criterion was used instead of the residual sum of squares:

$$\text{RIDGE}(k, \lambda) = \text{RIDGE}(k) + \hat{\beta}^T_{\text{ridge}(1)} (R(\lambda)^{-1} + H_{(p+2, p+2)})^{-1} \hat{\beta}_{\text{ridge}(1)} + o_{P^*}(1),$$

where $\text{RIDGE}(k)$ is the ridge criterion with penalty parameter $k$ (see Paper IV for a proof and for details about the term $R(\lambda)^{-1}$). Thanks to the penalization in the ridge

Figure 6.3: Illustration of the main principle of SimSel. The figure shows $\text{RSS}(\lambda)$ for two variables, one unimportant (cyan) and one important (red).
Figure 6.4: Illustration of the estimated distribution of the F-statistic originating from the regression line $\text{RSS}(\lambda)$ on $\lambda$ for a variable $x_i$ and the pseudo variable $z$. $\hat{f}_i(F)$ and $\hat{f}_{p+1}(F)$ denote the estimated distribution of the F-statistic corresponding to $x_i$ and $z$, respectively. $\hat{F}_i$ and $\hat{F}_{p+1}$ are the estimated distribution functions related to $\hat{f}_i(F)$ and $\hat{f}_{p+1}(F)$. $\alpha_1$ and $\alpha_2$ are parameters in the SimSel algorithm. (a) We consider $x_i$ to not be significantly important if $\hat{F}^{-1}_i(\alpha_2) < \hat{F}^{-1}_{p+1}(1 - \alpha_1)$. (b) Conversely, if $\hat{F}^{-1}_i(\alpha_2) > \hat{F}^{-1}_{p+1}(1 - \alpha_1)$ we consider $x_i$ to be significantly important. The parameters $\alpha_1$ and $\alpha_2$ thus control the risk of making errors of type I and II, respectively. I.e. the probability that the test incorrectly estimates $x_i$ as important when it is in fact unimportant is less than or equal to $\alpha_1$, and the probability that the test incorrectly estimates $x_i$ as unimportant when it in fact is important is less than or equal to $\alpha_2$. $\alpha_2$ thus regulates the power of the test. Adjusting $\alpha_1$ provides a way to deal with the multiple testing problem when $p$ is large. Typical choices of $\alpha_1$ and $\alpha_2$ are the classical 0.05 and 0.2, respectively.

6.2.1 Testing of SimSel

The SimSel algorithms developed in Paper III and IV were evaluated on both simulated and real data. The simulated data was constructed in order to test the SimSel algorithms in linear and nonlinear modeling settings and in errors-in-variable models, both when $H$ did and did not have full rank. We found that SimSel performed well, approximately equal to the advanced method of multivariate adaptive regression splines (MARS) [74]. In general we observed that SimSel (with default parameter settings) was more conservative (i.e. selected fewer variables) than MARS and performed slightly better in simple settings, such as when the true underlying function was linear. Given the very adaptable nature of the MARS method and the quite rugged second-order approximation employed in SimSel, this observation was not unexpected. Thus, con-
versely, in situations where the true underlying function is highly irregular, we would expect MARS to perform better than SimSel.

SimSel was also tested on part of the Selwood data (see Section 6.1.4) and on the prostate cancer dataset published in [75]. These datasets have been extensively analyzed (see e.g. [76] and [26], respectively) and it is well known which variables that are important and which are not. The results found with the SimSel algorithm conformed to what was previously known about the importance about the variables in these datasets. Moreover, the SimSel algorithm presented in Paper IV was tested on the part of the protein-protein interaction data published in Stiffler et al. (2007) [17] that was measured as pKi (i.e. had a continuos experimental end-point; see also Paper II). We fitted a random forests [77] model to the interaction data before and after variable selection by SimSel and estimated the generalization error by ten-fold cross-validation in each case. The generalization error estimates were found to be essentially the same (4.23 and 4.24, respectively) despite that only one seventh of the variables were used in the latter model. This result further supported the conclusion that SimSel retains relevant variables.

### 6.2.2 A brief discussion of the properties of SimSel

SimSel uses an approximative model as a "gauge" to estimate which independent variables that are relevant for modeling the response. As an approximative model, we used a quadratic approximation. This choice is analytically tractable and has worked well when testing SimSel in both linear and nonlinear contexts. The use of an approximative model makes SimSel quite rugged and insensitive to model misspecifications and assumption violations, which is an important and interesting property in real-world applications. However, it also means that SimSel is not a complete modeling framework (i.e. it does not try to accurately estimate the regression coefficients of important variables, like for instance the lasso [30] and MARS [74] do). Instead it acts as a filter that retains the variables that are important, and thus has to be used in combination with a modeling method.

In SimSel we do not, as opposed to most variable selection methods, study how a model assessment criterion is affected by excluding an independent variable. Instead we add pseudo errors to them. This affords repetition of the procedure for each variable and thus enables estimation of the distribution of the assessment criterion (compared to the point estimate obtained if a variable is excluded). This is an interesting property since all models are approximations of the true underlying functional relationship between independent and dependent variables. Thus, it is of interest to study the dispersion of the assessment criterion, which is possible when we estimate its distribution.

SimSel as presented in Paper III and IV was shown to work well on simulated and real data, and in complicated situations like nonlinear errors-in-variable models. However, the main point of the two papers is to present an interesting and novel application of disturbing datasets by noise. I believe that simulation techniques based on disturbing data by pseudo errors and pseudo variables constitute an entire flora of statistical methods that is as rich and general as for example the bootstrap. A large number of extensions to SimSel and new applications are indeed conceivable and will be pursued in the future.
6.3 Paper V

Bioinformatics is a cross-disciplinary research field. Any bioinformatics project therefore requires the usage of a number of different techniques from different research areas and of multiple (almost always incompatible) software systems. Typically development of new software and scripts for file formatting and data handling is also needed. These processes are time consuming, tedious, and substantially increases the risk of making errors at some step in the data management or analysis.

We therefore recognized that a flexible and integrative software system was needed from which all these tasks could be performed and to which new functionality from different research fields easily could be added (for example novel data analysis methods, such as the ones presented in Paper I-IV). This would enable bioinformatics projects to become more reproducible and reusable, less prone to errors, applicable with a much higher throughput, and allow for a higher degree of synergy between different research groups and fields. Consequently, Bioclipse was initiated (Fig. 6.5). Bioclipse is an open source workbench for the life sciences built as a rich client based on Eclipse (see Section 5.4) and contains features for importing, converting, editing, analyzing, and visualizing small molecules, proteins, sequences, spectra, and data matrices (Fig. 6.6(a)). In addition, many existing algorithms and standalone components have been integrated into the workbench, providing a very rich set of functionality. Bioclipse also has the possibility to interact with remote Web services (Fig. 6.5).

6.3.1 Bioclipse and data analysis

Bioinformatics attracts researchers from many fields such as computer science, statistics, and mathematics who find interesting problems to which they can apply their methods. The technical level required to use these methods is typically high and it is virtually impossible for scientists to both master the data generation and the data analysis parts of a high-throughput experiment. The demand for data analysts that are adequately trained for using new methods greatly exceeds supply and it is likely to
remain so for the foreseeable future [72, 78]. In order to exploit the full potential of high-throughput biology and biomedicine, it is thus crucial that the methods designed for high-throughput data analysis become available to the scientists who generates these data, i.e. the experimental biological and biomedical researchers. In fact, Miron and Nadon [78] states that the future success of high-throughput methods depend on "that scientists become less dependent on external expertise [...]". They continue by stating that "professional-grade statistical-analysis software that is both current and comprehensive will be needed to accomplish this".

The modular architecture of Bioclipse was already from the start of the project tailored to not only allow data management, but also data analysis. In fact, the scripting language available in Bioclipse seamlessly interacts with the statistical programming language R [79]. For example, the Metropolis-Hastings algorithms in Paper II and the SimSel algorithms in Paper III and IV were implemented and tested in R, thus making them accessible from within the Bioclipse workbench. However, R has a rather steep learning curve if one is not already acquainted with programming and statistics. In parallel with the development of Bioclipse I have therefore developed P, a library written in Java (http://www.java.com) for statistical learning (in fact, the analyses in Paper I were conducted in P), and jointly with others Bioclipse-P, a prototype graphical interface to access the functionality in P from Bioclipse (Fig. 6.6(b); unpublished work in progress).

Currently, work is being initiated to develop a general object model for representing data, statistical models, and algorithms that encompasses both Bioclipse-P’s and R’s data and object representations (and, in addition, other software and programming languages for statistical modeling, in particular the Hierarchical Bayes Compiler (HBC) and Bayesian inference Using Gibbs Sampling (BUGS), for easy specification of MCMC-algorithms). The object model will be implemented as a middleware that will handle transfers from the Bioclipse-P representation of data, statistical models, and algorithms to the corresponding R representations, and vice versa (Fig. 6.6). The middleware will leverage on an already existing library called ’rJava’ (http://www.rforge.net/rJava/) for executing R code from within a Java program (Fig. 6.7) and the functionality will be made available within the graphical interface in Bioclipse-P, with all technical details hidden from the user. This will result in an inte-
Bioclipse is today a rich and extensible workbench for management and visualization of biological and chemical entities and data. The process of developing the data analysis layer for Bioclipse intended from the onset is well under way, although unfortunately the time of my thesis work was not sufficient to provide the full integration of the tools into the Bioclipse workbench needed to accomplish the "professional-grade statistical-analysis software that is both current and comprehensive" envisioned by Miron and Nadon [78].

6.4 Final comments

Gottfried Leibniz is said to have been the last *homo universalis* to master all contemporary sciences [80]. The complexity and breadth of today’s science makes it impossi-
ble for any single scientist or research group to encompass all knowledge even within a given research field. The Internet, where scientist collaboratively and collectively can build up all contemporary knowledge, has instead become the modern polymath. However, as opposed to Leibniz, the Internet is (yet) not capable of using this knowledge to reason or draw conclusions. And this is where the challenge lies. We want to exploit the information and knowledge available on the Internet when analyzing data from a new biological or biomedical (high-throughput) experiment to infer new knowledge and to deepen our understanding of biology and medicine. In my opinion, eScience tools coupled with a Bayesian approach to data analysis available from within an intelligible system, such as Bioclipse, constitute a tenable approach to make this possible. I believe that analogously to how Leibniz during his life time advanced virtually all sciences in giant leaps, this possibility will have an immense impact on the efficiency of the scientific process and on the way science is conducted. The difference is that the process is not dependent on the brilliance of one mortal person, but is implemented in a sustainable system and may thus be made less transient and more reliable and constant. It is my hope that the work of this thesis is a step in the direction that will realize this system to salutary ends for science.
7. Acknowledgements

I would like to thank the following people who directly or indirectly have had a part in completing this thesis:

My supervisor, Jarl Wikberg. Thank you for giving me this opportunity and for *always* (literally 24/7) be available for discussions, questions, advice, and proofreading. Your laissez-faire supervising style suits me to a tee.

My main collaborator, Silvelyn Zwanzig. I *really* enjoy working together with you! Also, thank you for inviting me to come to Berlin to finalize the SimSel manuscripts when time was in great demand and short supply.

All the tax payers who have paid my salary - I hope to repay you all pretty soon. The work in this thesis was supported by grants from Swedish VR-M (04X-05957) and Uppsala University (KoF 07).


My fellow PhD-students, especially Fredrik Jernerén for always providing CL-watching opportunities and an extra bed in Uppsala when I needed one.

The administrative and technical staff at the Department of Pharmaceutical Biosciences, in particular Agneta Hortlund, Erica Johansson, and Magnus Jansson.

My main off-topic collaborator, Björn Rogell. You wrote in your acknowledgement that our collaboration was the best you have "started after seven beers at snärran". I can only agree, although I thought it was at V-Dala and after 10-12 beers...

Hanna, Darta, Kim, and Nervall (yes, I know that you’re called Löfqvist these days but...well, you know, you are still Nervall somehow). Onsdagslunchen, whatever day it has occurred on, has always been an excellent retreat away from manuscript typing, thinking, calculating, coding, and other hard labour.

All my friends around the world for providing brilliant times like drinking vodka with KAK-people in Polish restaurants, celebrating weddings in the Swiss Alps and the Norwegian fiords, jogging on a crisp Autumn day, eating sklep-food by the road somewhere in Eastern Europe, ...
My family, David, Britta, and Göran, for stimulating and encouraging analytical and independent thinking. And in general for being my family with everything that involves in terms of for example assistance in various situations, such as when a bunch of ceilings have to come down etc.

Ola. Well, what can I say, really. To do a PhD together with my best friend has been bloody brilliant. We started with a little too much ping-pong and ended with working a little too many late nights and weekends. But, to paraphrase Frankie-boy, we did it our way.

Sara. Thank you for your emotional support and for putting up with lost weekends and odd working hours. You make me a happy person - I love you.

091014, Heleneborgsgatan 14, Stockholm, with an Adam-cup of La Bomba and with pale autumn sun sifted through the hibiscus falling on my laptop.
Bibliography


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy 112

Editor: The Dean of the Faculty of Pharmacy

A doctoral dissertation from the Faculty of Pharmacy, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy”.)