UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 401

Fusing Domain Knowledge with Data

Applications in Bioinformatics

CLAES ANDERSSON

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2008

ISSN 1651-6214
ISBN 978-91-554-7094-4
urn:nbn:se:uu:diva-8477

Dissertation presented at Uppsala University to be publicly examined in Fåhraeussalen, Rudbecklaboratoriet, hus C:5, Dag Hammarskjölds väg 20, Uppsala, Thursday, March 13, 2008 at 09:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Andersson, C. 2008. Fusing Domain Knowledge with Data. Applications in Bioinformatics. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 401. 55 pp. Uppsala. ISBN 978-91-554-7094-4.

Massively parallel measurement techniques can be used for generating hypotheses about the molecular underpinnings of a biological systems. This thesis investigates how domain knowledge can be fused to data from different sources in order to generate more sophisticated hypotheses and improved analyses. We find our applications in the related fields of cell cycle regulation and cancer chemotherapy. In our cell cycle studies we design a detector of periodic expression and use it to generate hypotheses about transcriptional regulation during the course of the cell cycle in synchronized yeast cultures as well as investigate if domain knowledge about gene function can explain whether a gene is periodically expressed or not. We then generate hypotheses that suggest how periodic expression that depends on how the cells were perturbed into synchrony are regulated. The hypotheses suggest where and which transcription factors bind upstreams of genes that are regulated by the cell cycle. In our cancer chemotherapy investigations we first study how a method for identifiyng co-regulated genes associated with chemoresponse to drugs in cell lines is affected by domain knowledge about the genetic relationships between the cell lines. We then turn our attention to problems that arise in microarray based predictive medicine, were there typically are few samples available for learning the predictor and study two different means of alleviating the inherent trade-off betweeen allocation of design and test samples. First we investigate whether independent tests on the design data can be used for improving estimates of a predictors performance without inflicting a bias in the estimate. Then, motivated by recent developments in microarray based predictive medicine, we propose an algorithm that can use unlabeled data for selecting features and consequently improve predictor performance without wasting valuable labeled data.

Keywords: cell cycle, cancer chemotherapy, predictive tests, performance estimation, bioinformatics

Claes Andersson, The Linnaeus Centre for Bioinformatics, Box 598, Uppsala University, SE-75124 Uppsala, Sweden

ISSN 1651-6214
ISBN 978-91-554-7094-4
urn:nbn:se:uu:diva-8477 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-8477)

UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 314

Fusing Domain Knowledge with Data

Applications in Bioinformatics

CLAES ANDERSSON

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2008

ISSN 1651-6206
ISBN 978-91-554-7094-4
urn:nbn:se:uu:diva-8461

Dissertation presented at Uppsala University to be publicly examined in Fåhraeussalen, Rudbecklaboratoriet, hus C:5, Dag Hammarskjölds väg 20, Uppsala, Thursday, March 13, 2008 at 09:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Andersson, C. 2008. Fusing Domain Knowledge with Data. Applications in Bioinformatics. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 314. 55 pp. Uppsala. ISBN 978-91-554-7094-4.

Massively parallel measurement techniques can be used for generating hypotheses about the molecular underpinnings of a biological systems. This thesis investigates how domain knowledge can be fused to data from different sources in order to generate more sophisticated hypotheses and improved analyses. We find our applications in the related fields of cell cycle regulation and cancer chemotherapy. In our cell cycle studies we design a detector of periodic expression and use it to generate hypotheses about transcriptional regulation during the course of the cell cycle in synchronized yeast cultures as well as investigate if domain knowledge about gene function can explain whether a gene is periodically expressed or not. We then generate hypotheses that suggest how periodic expression that depends on how the cells were perturbed into synchrony are regulated. The hypotheses suggest where and which transcription factors bind upstreams of genes that are regulated by the cell cycle. In our cancer chemotherapy investigations we first study how a method for identifiyng co-regulated genes associated with chemoresponse to drugs in cell lines is affected by domain knowledge about the genetic relationships between the cell lines. We then turn our attention to problems that arise in microarray based predictive medicine, were there typically are few samples available for learning the predictor and study two different means of alleviating the inherent trade-off betweeen allocation of design and test samples. First we investigate whether independent tests on the design data can be used for improving estimates of a predictors performance without inflicting a bias in the estimate. Then, motivated by recent developments in microarray based predictive medicine, we propose an algorithm that can use unlabeled data for selecting features and consequently improve predictor performance without wasting valuable labeled data.

Keywords: cell cycle, cancer chemotherapy, predictive tests, performance estimation, bioinformatics

Claes Andersson, The Linnaeus Centre for Bioinformatics, Box 598, Uppsala University, SE-75124 Uppsala, Sweden

ISSN 1651-6206
ISBN 978-91-554-7094-4
urn:nbn:se:uu:diva-8461 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-8461)

‘Oh my God, it’s full of stars!’
David Bowman
2001: A Space Odyssey

List of papers

This thesis is based on the following papers, which will be referred to by the Roman numerals assigned below:

I In Vitro Drug Sensitivity-Gene Expression Correlations Involve a Tissue of Origin Dependency.
C.R. Andersson, M. Fryknäs, L. Rickardson, R. Larsson, A. Isaksson, and M. G. Gustafsson.
Journal of Chemical Information and Modeling, 2007, *47*, 239-248.

II Bayesian detection of periodic mRNA time profiles without use of training examples.
C.R. Andersson, A. Isaksson, M.G. Gustafsson
BMC Bioinformatics, 2006, 7:63.

III Revealing cell cycle control by combining model-based detection of periodic expression with novel cis-regulatory descriptors.
C.R. Andersson, T.R. Hvidsten, A. Isaksson, M.G. Gustafsson, and J. Komorowski.
BMC Systems Biology, 2007, 1:45.

IV A Maximum Entropy Empirically Based Prior can Improve the Credibility Interval for the Error Rate of a Single Classifier.
M.G. Gustafsson, U. Wickenberg-Bolin, M. Wallman, H. Göransson, M. Fryknäs, **C.R. Andersson** and A. Isaksson.
Submitted.

V Feature Selection using Classification of Unlabeled Data.
C.R. Andersson, R. Larsson, A. Isaksson and M.G. Gustafsson.
In manuscript.

Contents

Abbreviations

BIC	Bayesian Information Criterion
Cdk	Cyclin-Dependent Kinase
ChIP	Chromatin ImmunoPrecipitation
DLD	Diagonal Linear Discriminant
DNA	Deoxyribonucleotide Acid
FMCA	Fluorometric Microculture Cytotoxicity Assay
G0	Gap 0
G1	Gap 1
G2	Gap 2
MAP	Maximum A Posteriori
mRNA	messenger Ribonucleotide Acid
MEECI	Maximum Entropy Empirically based Credibility Interval
MTT	3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide
PCR	Polymerase Chain Reaction
PLS-DA	Partial Least Squares-Discriminant Analysis
SVM	Support Vector Machine

Introduction

Over the last decade researchers have miniaturized the molecular biologists' analytical tools in order to perform massively parallel analyses (Fodor, Rava et al. 1993). The first and foremost example is the mRNA microarray that in a single analysis measures expression of tens of thousands of different transcripts (Schena, Shalon et al. 1995). Massively parallel techniques are typically used to generate hypotheses for further investigation. For instance, mRNA microarrays can be used to generate hypotheses about what molecular pathways are involved in a phenotypic trait or a disease's etiology. This can be done with a genome-wide comparison against a control group that provides a list of genes differentially expressed betwixt the groups and associates genes with group differences. The mRNA microarray was the first massively parallel technique to reach wide-spread use but many have followed such as genome-wide location analysis aka ChIP-on-chip (Buck and Lieb 2004), comparative genome hybridization (Albertson and Pinkel 2003), and single nucleotide polymorphism array analysis (Chee, Yang et al. 1996). This thesis investigates different ways in which data obtained from such high-throughput analyses can be combined with background knowledge about the biology (domain knowledge) to analyze and generate sophisticated hypotheses about the molecular underpinnings of biological systems. The background knowledge we use include experimentally determined facts about the systems, e.g. gene functions, as well as ancilliary experimental data. We found the applications for our methods in two related areas of research: regulation of the cell cycle and cancer chemotherapy.

In Paper I we investigate an approach for analyzing in vitro chemosensitivity profiles across a cancer cell line panel together with mRNA microarray profiles of the cell lines. By using a simple visualization the investigator may identify groups of co-regulated genes that appear associated with chemoresponse to compounds that have similar chemosensitivity profiles. This suggests a relationship between a biological pathway and compounds with similar mechanisms of action. In principle the same relationship could be discovered by piecing together lists of genes differentially expressed between cell lines sensitive and resistant to the compounds, but such an approach would be much more laborious. A key point in Paper I is that domain knowledge in the form of genetic relationships between the cell lines must be accounted for in order to provide

an unbiased analysis. Inclusion of domain knowledge in integrative analyses of biological systems is a recurrent theme in this thesis.

In Papers II and III we study the cell cycle in the budding yeast *Saccharomyces cerevisiae*. In Paper II we propose a detector of periodicity that is derived from Bayesian principles and uses user-supplied domain knowledge about the period time. After evaluating the detector on simulated data we apply it to microarray time series analyses of synchronized yeast cultures. We then analyze to what degree putative binding sites for transcription factors can explain the appearance of periodic expression. Our analysis provides hypotheses about which motifs confer periodic expression. We also study to what degree domain knowledge about cell cycle genes explain periodicity as predicted by the detector. In Paper III we study whether combinations of *cis*-regulation descriptors explain the appearance of periodic expression that depends on the synchronization method used. The *cis*-regulation descriptors are integrated from genome-wide location analysis of transcription factor binding and putative binding sites for transcription factors. Not only does our analysis provide some systems-wide observations on the overall connectivity of gene regulation, but the hypotheses generated take the form of statements about how a gene's expression behaves under different experimental conditions. Each hypothesis suggests which transcription factor needs to bind to what motif in order for a gene to exhibit phase specific expression. Importantly, we demonstrate that by describing time profiles of gene expression on a semantic level (periodic expression) we are able to provide sophisticated hypotheses about cell cycle regulation that focus on known cell cycle related *cis*-regulation descriptors.

In Paper IV and V we return to the context of cancer chemotherapy but our findings are much more general. Specifically the research originated from problems that arise in the construction of predictors of chemoresponse from mRNA microarray data. Although the situation is improving as the price and complexity of microarray analysis drops there are typically few samples available for the design and evaluation of classifiers. The investigator faces a trade-off between how good the predictor will be (number of samples allocated to design) and how well its performance is estimated (number of samples allocated to validation). In Paper IV we investigate whether better performance estimates can be obtained by using information from independent tests of the predictor on design data as prior knowledge. This prior knowledge, expressed as a probability distribution function of classification error rates represents information about how difficult the problem of classification is, i.e. the prior is specific to the domain of the application. In Paper V we demonstrate how we can integrate additional unlabeled data in the design of classifiers, thus making full use of all data available. The method should be particularly useful when the data used for design comes from a different distribution than data the classifer should be applied to, a situation faced when designing classifiers of chemoresponse from cell line data and applying them to patient data.

In the following chapters I will briefly review the biological and biomedical context the papers originated within, followed by a short introduction to the different computational methods used, methods for generating the high-throughput data analyzed and a discussion of each the papers. Bioinformatics is an inter-disciplinary subject so the background is presented on a level suitable to all interested readers with pointers to additional information for readers with special interests.

Biological and Biomedical Context

This thesis investigates how domain knowledge can be used to integrate heterogeneous types of high-throughput data in a number of specific applications. Our applications fall within two related biological and biomedical contexts: in Papers II and III we study the cell cycle in *S. cerevisiae*; Paper I investigates analysis of gene expression-chemosensitivity associations and Papers IV and V were prompted by investigations into the design of predictors of cancer chemosensitivity.

The Cell Cycle

Mitosis is the process by which two identical cells are formed from a mother cell. Its molecular regulation is highly conserved in eukaryotes. For a full description the reader should see any textbook on molecular cell biology, e.g. “Molecular Biology of the Cell" (Alberts 2002). Briefly, cell division was first observed using light microscopy and was seen to cycle between two phases dubbed interphase and mitosis (M-phase). Interphase does not have any morphological characteristics, but the M-phase can be further subdivided based on morphological changes (see Figure 2a). First come prophase which is recognized by the condensation of chromatin and a dissolving nuclear envelope. Then follows metaphase in which the fully condensed chromosomes align at the equatorial plane of the cell in a structure called the metaphasic plate. At each pole, structures called polar bodies attach through microtubuli to the centrosomes of the chromosomes. Metaphase is followed by anaphase which is characterized by the chromosomes being pulled apart. The cycle ends after telophase, where two distinct cells and the formation of nuclear envelopes in each of the daughter cells can be recognized.

Interphase can be further subdivided by events taking place at the molecular level (see Figure 2b). Obviously, the genome must be replicated prior to division. Replication is prepared for in Gap 1 (G1), the first stage of interphase. A copy of the genome is then synthesized in S-phase which is followed by Gap 2 (G2) in which the cell prepares for mitosis. Incidentally, the quiescent state in which the cell is not committed to mitosis is called Gap 0 (G0). The cell cycle is a carefully concerted process and the molecular regulation is carried out by cytoplasmic proteins. A group of proteins called cyclins rise and fall in concentration in the different stages of the cell cycle. Cyclin D concentration increases in G1, cyclins E and A in S-phase and cyclins B and A in M-phase. In addition, there are a number of kinases that depend on cyclins for activation, the cyclin dependant kinases (cdk). By transferring phosphate moieties they activate proteins that control cell cycle processes.

Cell division is a precarious undertaking and cells have a number of checkpoints to ensure high fidelity of replication. If the cell fails beyond recovery at these checkpoints it enters apoptosis (programmed cell death). For instance, the process is stopped if DNA damage is detected either prior to (G1 checkpoint), during, or immediately after synthesis (the G2 checkpoint). In addition there is a checkpoint in M-phase that arrests the cell in metaphase if a microtubule fails to attach to a chromosome. Understanding these mechanisms is of great medical interest for the treatment of cancer as is illustrated in the next section. The core machinery has been intently studied, but much remains to be discovered about the cell cycle, in particular about events downstream of the cell cycle regulators which are studied in Papers II and III.

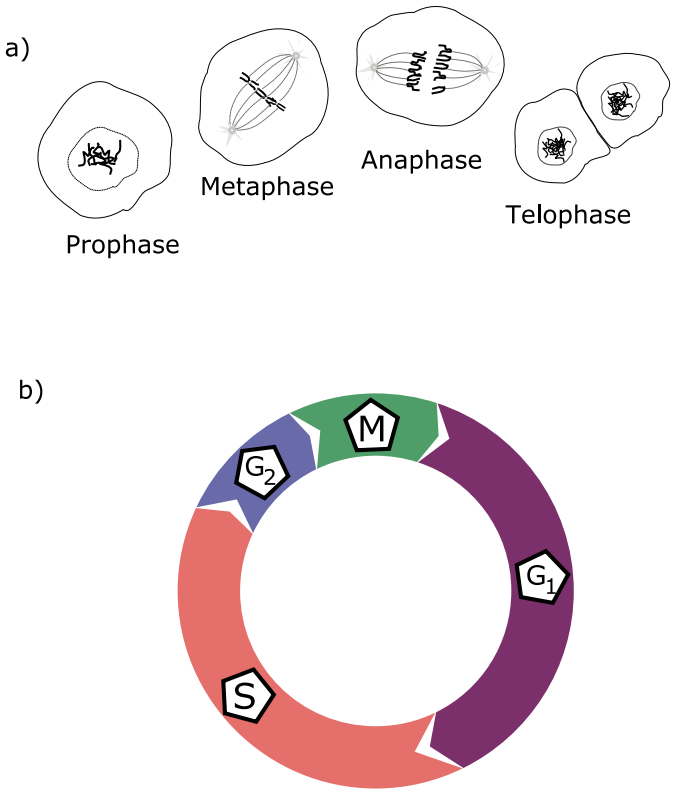

Figure 1. a) Stylized representations of the phases of mitosis as seen in a light microscope. b) Graphical representation of the chronological order of the cell cycle phases.

Cancer Chemotherapy

The overall structure and function of organs and tissues is maintained by controlling cell replication by e.g. contact inhibition. Occasionally control over the carefully concerted cell replication machinery is lost and a clone will start to proliferate. The loss of control may be due to either an activating mutation of a proto-oncogene or a loss-of-function mutation in a tumor suppressor gene. This is not an uncommon event, but the immune system has cells with an innate ability to eliminate cells that do not respect tissue boundaries. However, if an uncontrolled growth evades the immune system a cancerous growth may develop. It is difficult to say at what stage a new growth becomes a cancer tumor and pathologists usually characterize suspected cancer tumors by the degree of de-differentiation in the growth. If the growth has lost all phenotypic characteristics of the original tissue it is a clear sign of an emerging cancer. Clinically cancer typically presents symptoms due to interference with the surrounding tissue, the notable exception being endocrine tumors that may produce a plethora of symptoms by overproducing different hormones. For an excellent review of cancer biology, see (Hanahan and Weinberg 2000).

Treatment of solid cancers usually starts with surgical removal of the tumor mass followed by chemotherapy, for hematological malignancies chemotherapy is the first line treatment. The majority of cancer chemotherapies target dividing cells in general causing the well known side effects of nausea (due to loss of gastrointestinal epithelia) and hair loss. Most cancer chemotherapies work by triggering apoptosis by causing damage either to microtubuli or DNA, causing the cell to fail irrevocably at the cell cycle checkpoints. There are four classic mechanisms of action for cancer cytostatics: microtubule inhibitors, topoisomerase I and II inhibitors, antimetabolites and alkylating agents. Microtubule inhibitors act by either destabilizing or hyperstabilizing the tubulin polymers causing the cells to fail in M-phase. The topoisomerase inhibitors prevent the cells from replicating the DNA. Antimetabolites are nucleotide analogs that prevent further replication by inhibiting enzymes that catalyze production of deoxyribonucleotides, the building blocks of DNA needed for synthesis of a new DNA strand. Alkylating agents cause direct damage to the DNA by cross-linking strands and thus preventing further replication. Although targeted drugs such as tyrosine kinase inhibitors are becoming available, most chemotherapy is based on drugs having one of the above mechanisms of action.

The most common reason for failed treatment of cancer is drug resistance where the cancer cells either acquires or presented with mechanisms for evading chemotherapy. Cells may for instance express drug efflux pumps such as the Multi-Drug Resistance transporter that remove the drug from the cytosol. By analyzing chemoresponse data together with mRNA expression data it is possible to identify pathways that confer resistance as well as sensitivity, Paper I analyzes one method for doing that.

The phenomenon of drug resistance motivates current best clinical practice that uses a combination of drugs with different mechanisms of action. Thus the cancer cells must have several different mechanisms of resistance to escape treatment. However, even if originating within the same tissue, each individual instance of cancer develops against the patient's unique genetic background. Even if two therapies have shown similar effects on overall survival clinical experience shows that individual patients may benefit from one therapy but not the other. It is hoped that overall cancer survival rates can be improved by selecting therapy on a patient to patient basis.

Cell culture based drug resistance tests such as the fluorometric microculture cytotoxicity assay can be used to select the appropriate therapy (Larsson and Nygren 1993) but has thus far failed to gain wide-spread acceptance in the clinic. Unfortunately the number of drugs that can be evaluated is usually severely limited by the amount of tissue available. However, it has recently been suggested that response to therapy could be predicted from microarray analysis of cancer cells (Hess, Anderson et al. 2006; Potti, Dressman et al. 2006; Dressman, Berchuck et al. 2007). Since a microarray analysis requires far less tissue, this would open up the possibility of evaluating all approved drugs for effect on a patient to patient basis. Issues arising in the design of predictors of cancer chemosensitivity motivated the research presented in Papers IV and V.

Elements of Learning from Data

For the purposes of this thesis, bioinformatics is the science of analyzing and testing hypotheses using models constructed from the voluminous datasets generated in molecular biology. The sheer amount of information available means processing must be done computationally. Throughout this thesis we employ computer algorithms for the construction of models from data, i.e. machine learning. In Papers II and IV we present new algorithms derived using the Bayesian formalism of probability which I describe below, followed by a brief description of different methods of machine learning.

Statistical Inference

Probability theory plays a central role in life sciences as the formalism of statistical inference: the process of drawing conclusions from data, or more specifically, the process of drawing conclusions about a population using data collected from a sample of the population. For conclusions to be objective a formal procedure is needed. In the common school of statistics the basic procedure for stating that some effect is visible in the data is as follows. A mathematical model is stated that describes how frequently the effect would appear by chance if it is actually absent. The hypothesis that there is no effect is called the null hypothesis. Then the model is used to calculate the probability that the observed effect would occur by chance, the p-value. If it is very unlikely to occur by chance the null hypothesis is rejected in favor of the alternative hypothesis that there actually is an effect. Each investigator may choose how unlikely the effect must be for the null hypothesis to be rejected. The point is that quantitative rather than qualitative judgment can be cast, which makes communication of scientific results much easier.

The key step in turning qualitative judgment into quantitative in the above procedure is to capture the notion of chance in a mathematical formalism. There are two different schools of thought regarding probability, frequentist and Bayesian. The main differences are outlined below.

Bayesian Probabilities

In the frequentist school the probability of an event is defined as the frequency with which the event occurs in an infinite number of trials. In the Bayesian view probability reflects ignorance on part of the investigator: probability is interpreted as a degree of truth, or plausibility. Although this notion may seem too vague to be formalized, R.T. Cox demonstrated that the Bayesian calculus of probabilities can be derived from a set of basic desiderata (desidered properties) on how a measure of plausibility should behave (Cox 1946), stated by Jaynes (Jaynes and Bretthorst 2003) as:

(I) Degrees of plausibility should be represented by real numbers

(II) The measure should qualitatively correspond with common sense

(III) The measure should be consistent

where consistent means that all possible ways of reasoning should give the same result, always taking into account all evidence, and that equal states of knowledge are represented with equivalent assignments of plausibility. For a good introduction to Bayesian probability in the sense we use it, the reader should see (Jaynes and Bretthorst 2003). In this brief review we shall use the usual P to denote probability measures. In contrast to conventional probability theory P is not a measure of the size of some set of outcomes, but rather a measure of the degree of truth in a statement. Thus $P(q)$ should be interpreted as the degree of truth in the statement that the parameter q takes some value.

Bayes' Theorem

To illustrate Bayesian probabilities, consider the following law of probability:

$$P(q, D) = P(q \mid D)P(D) = P(D \mid q)P(q). \qquad (1)$$

From (1) it follows that

$$P(q \mid D) = \frac{P(D \mid q)P(q)}{P(D)} \qquad (2)$$

which is known as Bayes' theorem. Now, suppose q is a parameter such as the weight of an object and D is a set of measurements of the weight. Although in full accordance with the laws of probability, the left hand side of (2) is a forbidden quantity in frequentist statistics since q is not a random variable. In other words, although not exactly known, the object has a well defined weight which is a property of the object. Weight is not subject to chance. In the Bayesian view, probabilities denote a degree of belief and there is nothing strange about (2). Furthermore, the function $P(q|D)$ expresses the plausibility of q taking different values and can be used for estimating the value of q. For instance, choosing the most probable value of q is called the maximum a posteriori estimate. The function may also be used for constructing a credibility interval for the parameter q, which we do for error rates in Paper IV.

Prior and Posterior Probability

The function $P(q)$ in (2) is called the *prior*, and $P(q|D)$ the *posterior*. These names allude to the entry of data into the calculations, i.e. the functions describe uncertainty about q prior and posterior to seeing data. $P(D|q)$ is known as the likelihood function which incidentally forms the basis of likelihood-based statistics (a field of classical statistics). The denominator of (2), $P(D)$, is simply a normalization constant which ensures that the left hand side sums to one. It may be calculated by summing up $P(D|q)P(q)$ for all possible values of q, a technique known as *marginalization*. Here we may note an important fact: if $P(q)$ is independent of q (i.e. a constant), which corresponds to all values of q being equally likely, $P(q|D)$ is directly proportional to the likelihood function $P(D|q)$. Then, when selecting an estimate of q, there would be no difference between using a Bayesian treatment or likelihood-based statistics.

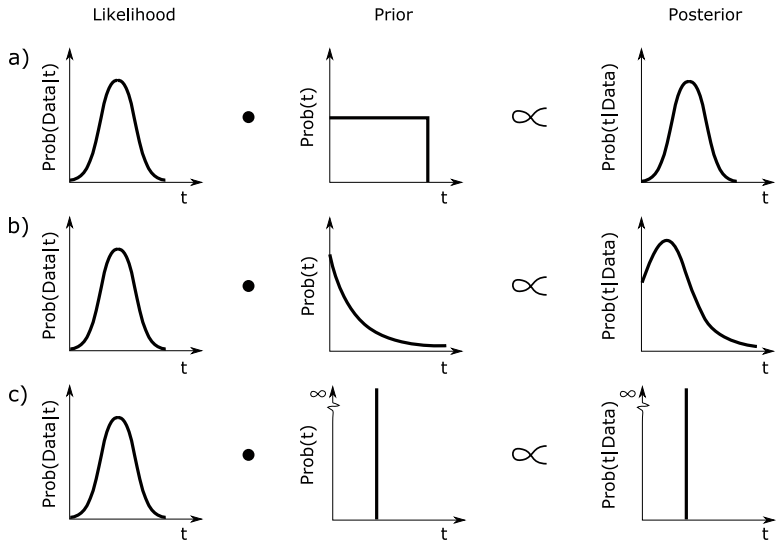

Figure 2. Illustration of how the posterior density is affected by different priors for the same likelihood. a) If the prior is uninformative, the posterior will be directly proportional to the posterior (same shape). b) A prior suggesting that smaller values of *t* are more likely will shift the probability mass towards smaller values. c) When the prior specifies one and only one value (represented by a Dirac impulse function), data cannot change the information.

The prior is a source of controversy as it on the surface introduces subjectivity into the analysis that is not present in frequentist statistics: two researchers might draw different conclusions from the same dataset if their prior knowledge differs. This is not as serious as may appear at first. With a bit of thought it is obvious that if an investigator possesses different prior information the data should be interpreted differently. If there is prior information excluding certain values of a parameter it doesn't matter if some value has a high likelihood, those values should be excluded in the posterior as well. Figure 2 graphically illustrates the interaction between prior and likelihood in estimation of a continuous parameter.

Although it is only natural for two investigators with different prior information to draw different conclusions, an objective analysis require that two investigators with the same prior express it as the same probability function using some procedure. Such procedures are available, e.g. Laplace indifference principle, transformation group invariance and maximum entropy (Jaynes and Bretthorst 2003). We illustrate how these principles provide objectivity by using Laplace indifference principle. It states that if any set of outcomes are considered equal by the prior information at hand, all outcomes in that set should be assigned equal probabilities. Consider the toss of a coin. What probabilities should be assigned to the outcomes Heads

and Tails respectively? Since the only available information is that there are two possible outcomes that are mutually exclusive, the only consistent assignment would be P(Heads) = P(Tails) = ½.

The Maximum Entropy Principle

In Papers II and IV we use the maximum entropy principle for expressing prior information. Entropy is a measure of uncertainty, much like probability is a measure of chance or plausibility. For example, returning to the coin toss, if the outcome was known to be Heads prior to tossing, there would be no uncertainty. Intuitively, the largest degree of uncertainty about the outcome is the fair coin with P(Heads) = P(Tails) = ½. Given a set of probabilities p_i of the different possible outcomes, the entropy function H is defined as:

.

$$H = -\sum_i p_i \log p_i \tag{3}$$

For the case of the coin toss, the maximum entropy is obtained when P(Heads) = P(Tails) = ½ as desired, an assignment consistent with Laplace indifference principle. The unit of the uncertainty measure is determined by the base of the logarithm in (3). For example, if base 2 is used, uncertainty will be measured in bits. The measure originated within communication theory where a measure of information was needed for mathematical analysis of communication channel capacity (Shannon 1948). Its functional form was derived from a set of basic desired properties in much the same way as the Bayesian calculus was derived. Specifically, Shannon argued that a measure *H* of uncertainty should:

(I) Be a continuous function of the probabilities. Otherwise arbitrarily small changes in the probability distribution could lead to a large change in the amount of uncertainty.

(II) Should correspond qualitatively to common sense in that we are more uncertain when there are more possibilities than when there are few.

(III) Be consistent

(Jaynes and Bretthorst 2003) where consistent is given the same definition as was given above for the derivation of the Bayesian calculus of probabilities. It can be shown that the functional form of the entropy function is the only one satisfying these desiderata, and there is a straightforward extension to probability density functions, the differential entropy functional.

The principle of maximum entropy dictates that if a set of constraints on a variable is given, e.g. a known mean value, the uncertainty about the parameter should be expressed as the probability distribution that maximizes

the entropy function and thus the measure of uncertainty. In other words, by using the maximum entropy principle one ensures that no additional, implicit information is added when the prior information is expressed as a probability function. Incidentally, the functional form of the maximum entropy probability distribution function for a given mean and variance is the standard Normal distribution[1], something which is often touted as an explanation for the success classical inferences has had using the Normal distribution even when the true distribution doesn't follow it.

Bayesian Inference

A point of radical departure between frequentist statistics and Bayesian inference is that of hypothesis testing. Using Bayesian inference it is possible to calculate the probability that hypothesis *i* is true given the data as

$$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{P(D)} \qquad (4)$$

As in classical statistics, the decision as to which hypothesis to declare true is left to the investigator. However, in Bayesian inference the decision is based on whether the hypothesis is sufficiently probable given the data, not what the risk is of making an error if it is declared true. Now, to illustrate an important point, consider the denominator of (3), $P(D)$. It can be calculated as

$$P(D) = \sum_i P(D \mid H_i)P(H_i) \quad . \qquad (5)$$

Thus, in a Bayesian treatment it is not possible to calculate the probability of a hypothesis being true without fully specifying the alternate(s). Since the probability of observing data under the alternate hypothesis never is calculated in classical tests, it is possible to draw some erroneous conclusions. A low p-value does not necessarily mean that data supports the alternative hypothesis; the p-value under the alternate may be exactly equal, in which case the data is not informative.

[1] Strictly speaking this is only true for if the probability density function has support (non-zero density) for all real numbers, a distinction that is important in Paper IV.

Computational Techniques

Bayesian methods are not yet widely accepted. Besides being criticized for being subjective, it is very common for multidimensional integrals to arise in Bayesian calculations. These integrals appear when the model contains many parameters, only a few of which are of interest. For instance, when comparing two models as in (4), the parameters of the models are not of interest. This is handled by integrating over all parameters (marginalization). Unfortunately, the integrals are rarely amenable to analytical treatment and numerical integration becomes very costly when there are many variables to be integrated (the number of points at which the integrand must be evaluated grows exponentially with the number of parameters if each parameter is discretized in the same number of steps). There are several solutions to this problem. One solution is to use conjugate priors (Gelman 1995). This simply entails choosing functional forms of the prior which makes the integrals analytically treatable. From a purist point of view, however, this amounts to changing the problem to fit the calculations. Another possibility is to employ Monte Carlo integration schemes (Gelman 1995) which escape the problems associated with calculating high dimensional integrals numerically by stochastically seeking out the parameters that contribute the most to the integral. However, such schemes are computationally intensive and require monitoring convergence to a stationary distribution. A more palatable approach is the use of approximation techniques and heuristics.

By virtue of the Central Limit Theorem, the posterior will tend to a Gaussian form as more samples are collected. Thus, one strategy is to use a quadratic approximation of the log-likelihood at the maximum of the posterior. This is known as the Laplace approximation (Gelman 1995) and has been applied with great success in many applications. In calculating the Laplace approximation one must obtain the maximum of the posterior as well as the Hessian evaluated at the maximum. An even simpler heuristic is the Bayesian Information Criterion (BIC), also known as Schwartz Information Criterion (Hastie, Tibshirani et al. 2001), used in Paper II.

Bayesian Information Criterion

As it turns out, the Laplace approximation can be further approximated. The determinant of the Hessian can be bounded, which results in an even simpler criterion, requiring only the maximum of the posterior to be located. Specifically, the BIC for a model (hypothesis) H is

$$BIC(H) = \log p(D \mid \theta_{MAP}, H) - \frac{k}{2}\log n \qquad (6)$$

where the first term is the log likelihood function of the model evaluated at the maximum a posteriori parameter setting θ_{MAP}, k is the number of parameters in the model and n the number of observations. BIC has been used as a criterion for model selection outside the Bayesian community. Although the approximation only is valid for large sample sizes, it can be motivated from a pragmatic standpoint as a measure of fit of the model (the likelihood evaluated at the maximum of the posterior), penalized by the number of parameters of the model. The latter part can be construed as an application of Occams razor, trading between model fit and complexity.

Reconciling Bayesian and Frequentist Probability

It must be noted that for Bayesian inference to have use in real world applications, a higher degree of belief must on average correspond to higher frequency, i.e. if probabilities do not correspond to frequencies, why would it make sense to base our decisions on them? On the other hand, frequentist statistics needs to embrace Bayesian views. If the investigator has prior information that contradicts the result of the statistical test, she is likely to doubt the test. Bayesian inference allows this prior information to be described and quantitated (Kendall 1949).

On an ending philosophical note, frequentist probabilities, just like Bayesian, are mathematical representations of real world phenomena in the same way as the points, lines and circles of geometry are mathematical representations of everyday objects. It can be argued that randomness and chance in its very nature reflects ignorance on part of the investigator. That the most useful description is statistical does not mean it is impossible to describe the process in detail. For example, statistical mechanics successfully describes matter, e.g. the distribution of molecules' kinetic energy in a volume of gas. Nevertheless, it could, in principle, be described by conventional mechanics. It is our lack of knowledge that leads to a statistical description. Thus, in our view, whether "true" randomness exists in nature is a moot point since it is indiscernible from lack of knowledge.

Machine Learning

The concept of machine learning arose in the artificial intelligence community. In practice it involves running an algorithm with some dataset as input which outputs a model describing the data. The algorithms can be divided into supervised and unsupervised learning algorithms. Unsupervised learning algorithms construct models that highlight relationships between samples and variables. Supervised algorithms take samples with group labels and construct a model that describes the differences between samples with different labels, i.e. a classifier. Machine learning algorithms come in many different shapes, many of which are inspired by statistical theory. Popular unsupervised algorithms are hierarchical and k-means clustering and principal components analysis. Examples of supervised machine learning algorithms include k-Nearest Neighbor, decision trees, linear discriminant functions, neural networks and support vector machines (Hastie, Tibshirani et al. 2001). Such algorithms are developed in parallel in many different communities, artificial intelligence, statistics and pattern recognition to name a few. This is reflected in the different terminologies in use. For instance, in statistics a model for predicting group labels is a discriminant function, in pattern recognition a classifier. Furthermore, the terms variable, attribute and feature are used interchangeably for denoting a value that has been recorded for each sample. Below I will use the terminology used in the community in which the algorithm originated in.

The relative values of heuristic machine learning algorithms and those derived from assumptions about the functional form of the data distribution are debatable and there is an emerging view that statistically founded algorithms come up short when applied to the high-dimensional and structured data available today (Breiman 2001). However, algorithms derived from principles of mathematical statistics have their own value since usually at least some of their properties can be proven mathematically.

When using supervised learning for the mere purpose of predicting labels it would seem that whatever algorithm produces the most accurate labeling would be most desirable. However, if one would like to learn something from the resulting model, it must be possible to interpret it. The interpretation will of course depend on the formalism the output model is described in. Most classifiers used in microarray analyses, such as the one used in Paper V, were derived using a vector space representation of the samples, that is each sample is described by some vector x in R^n which can be interpreted geometrically. In Paper III however, we use a classifier derived from the theory of rough sets (Pawlak 1982) which produces a model described in terms of rules.

Regardless of how the model is expressed, an important aspect of classifier design is how to evaluate the classifiers performance, which is the subject of Paper IV. Below I will briefly outline how vector space classifiers

can be interpreted, the idea behind rough set based classification, some aspects of performance evaluation of classifiers and finally describe how two popular unsupervised learning algorithms work.

Vector Space Classifiers

Many classifiers assume samples are described by a vector and can be described as real valued vector functions f: $R^n \rightarrow R$. For a binary classifier of some classes $\mathbf{C}_1$ and $\mathbf{C}_2$, we may assume without loss of generality that the classifier predicts class $\mathbf{C}_2$ if $f(x) > 0$, class $\mathbf{C}_1$ otherwise. The set of points x for which $f(x) = 0$ is called the decision boundary. Figure 3a-b visualizes the decision boundary for classification from two variables as well as the difference between a linear classifier and a non-linear classifier. The shape of the decision boundary is computed from the design data using an algorithm that selects parameters of the function f that minimizes the error rate or some other criteria on the set of samples used for learning. Non-linear classifiers are known to be more sensitive to outliers than linear classifiers and in general require more data for learning. Thus it is common to use linear classifiers when predicting from microarray data.

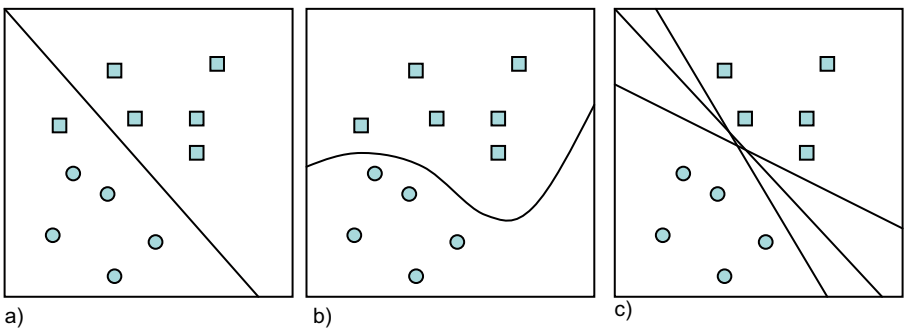

Figure 3. Graphical visualization of a classifiers decisions boundary when the examples are described by two variables. Squares and circles indicate samples from different classes. a) A linear decision boundary with one misclassified example. b) Non-linear decision boundary that separates the examples without errors. c) There could be many different choices of decision boundary which all classify the examples perfectly.

The differences between different linear classifiers such as the support vector machine (SVM), partial least squares-discriminant analysis (PLS-DA) and the diagonal linear discriminant (DLD) (Hastie, Tibshirani et al. 2001; Webb 2002) lie in how the coefficients are computed: linear support vector machines choose coefficients that maximize the margin between design data from the different classes; diagonal linear discriminant choose coefficients

optimal when variables in each of the classes follow independent Normal distributions; PLS-DA builds a linear discriminant on a small number of (hidden) latent variables that it assumes the observed features are correlated to.

When there are more variables than samples available for design, there are typically an infinite number of choices that minimize the error rate to zero on the design set (see Figure 3c). The linear SVM and PLS-DA methods have been designed with this in mind and makes what would appear to be rational choices. For instance, in many real-world problems with high dimensionality many of the features will actually be correlated to an underlying variable suggesting that PLS-DA is a good choice. It is for example reasonable to expect gene expression patterns to be correlated. The DLD on the other hand may suffer greatly by using variables for discrimination that appeared informative by chance. A general strategy for overcoming this is feature selection where informative features are chosen prior to designing the classifier. The simplest strategy for doing this is applying some test of how well each of the features separates the classes on their own choosing the top-ranked features. In Paper V we examine if unlabeled data can be used to boost supervised feature selection.

Rough Set Classification

In the rough set classifier the model is represented as a set of rules, each stating conditions the example should fulfill to obtain a given label. For a full introduction to rough sets in classification, see e.g. (Ohrn and Rowland 2000). Briefly, rough set classifiers are based on the mathematical theory of rough sets for describing uncertainty in data. In contrast to probability theory that provides a measure of uncertainty, rough set theory is concerned with computing *what* is uncertain. However, the workings of rough set classifiers can be explained without a formal introduction to the theory. Given a dataset D, where each object is described by a set of discrete valued attributes (features) A, the algorithm computes minimal subsets of A that suffice to distinguish as many objects in D as the entire set of attributes A can. Consider e.g. the dataset in Table 1.

Attribute 1	*Attribute 2*	*Attribute 3*	*Label*
Blue	Wet	Funny	Crunchy
Blue	Wet	Funny	Crunchy
Red	Wet	Boring	Smooth
Blue	Dry	Boring	Smooth

Table 1: Fictive data set for illustration of the rough set classifier methodology. See text for details.

Each observation is labeled with values in {Crunchy, Smooth} and is described by three attributes {Attribute 1, Attribute 2, Attribute 3}, valued in {Red, Blue}, {Dry, Wet} and {Boring, Funny} respectively. Now we ask what the minimal subsets of attributes are that retain the same discriminative power as all three attributes. Furthermore, in devising a classification scheme we are not interested in discriminating between observations belonging to the same class (same label). Now, from inspection it is obvious that only Attribute 3 could be used on its own to discriminate between the two classes. Furthermore, we note that Attribute 1 and 2 together could be used to discriminate between the classes. Thus, {Attribute 3} and {Attribute 1, Attribute 2} are the minimal subsets that retain the full discriminatory power of the full attribute set. Each such minimal subset is termed a reduct. It is important to note here that even when there is overlap between different classes, the reducts are still well defined.

Computing all reducts is computationally expensive and heuristics such as genetic algorithms must be applied for large datasets. Furthermore, instead of computing reducts which distinguish all members of one class from those of another class (a full reduct) it is common to compute reducts which discriminate one object from a class from all other of another class (object based reducts). A further development is approximate reducts in which the restrictions are loosened; the idea is to compute reducts which distinguish an object (or set of objects) from at least some user specified fraction of objects from other classes.

Regardless of the manner they were computed a rule may be formed from each reduct such as "IF Attribute 1 = Blue and Attribute 2 = Wet THEN Crunchy". In a resulting rough set classifier there are typically many such rules and it can be difficult to appreciate any general characteristics of them. Nevertheless, each of the rules is easy to interpret and general rules, i.e. rules which apply to a large set of examples, can be very valuable. When a new example is to be classified, its attributes are checked against each of the rules' left hand side and matches are noted. In order to arrive at a final classification a voting scheme is employed which corresponds to the practice of boosting (Hastie, Tibshirani et al. 2001) in which a large number of classifiers are built and the final classification is formed from the consensus.

The primary motivation for employing a rough set classifier is that the model has a rather pleasant and intuitive interpretation. It generates a minimal description of objects in a set (i.e. a class) in terms of a set of values of attributes. That being said, the method requires the attributes to take discrete values, thus continuous valued features requires discretization. However, this will not be covered here since in this work rough set classifiers have only been used for discrete, binary valued attributes. In Paper III we use rough sets classification for computing minimal subsets of *cis*-regulation descriptors that explain gene expression.

Performance Evaluation

Regardless of how the classifier was built its performance must be evaluated on unseen data. Performance of classifiers is usually measured by the error rate: the probability that a sample is misclassified. If the design data were to be used for performance evaluation the estimate is very likely to be positively biased since most learning algorithms output the classifier that minimizes the error rate on that particular data set. The straight-forward solution is to use a hold-out dataset for test. If the hold-out dataset is very large the empirical error rate in the test set will be a good estimate of the true error rate. However, in many bioinformatics applications there are typically few samples available for test and the error rate estimate is uncertain. The uncertainty about the error rate q after misclassifying k out n samples in a hold-out set can be described as a Bayesian probability density function as:

$$P(q \mid k,n) = \frac{P(k \mid n,q)P(q)}{P(k \mid n)} \propto \binom{n}{k} q^k (1-q)^{n-k} \tag{7}$$

where we have assumed that we have no prior information about the error rate, that is $P(q)$ is uniform on the interval [0,1], and that the n tests were

independent of each other. The function $P(q|k,n)$ can be used to obtain useful numbers such as an estimate of what error rate the *true* error rate is smaller than with some probability, or a credibility interval around the expected error rate. Proper estimates require much data. Suppose the true error rate of the classifier is 0. When using (7) to state with 95% confidence that the classifier performs no worse than guessing (50% error rate), only 4 samples are needed. However, about 30 samples are needed to state that the error rate is lower than 10%, 60 samples for below 5% and some staggering 300 samples to state that the error rate is below 1% with 95% confidence. Many microarray datasets contain on the order of 20 samples in total and the trade-off between how good the classifier will be (number of samples allocated to design) and how certain one is about the performance (number of samples allocated to validation) becomes crucial.

There are a number of computational techniques for alleviating the problem, such as cross-validation and bootstrapping (resampling). Cross-validation (Hastie, Tibshirani et al. 2001) is the most commonly used method for alleviating this problem, presumably because of its computational simplicity. The basic strategy is to divide data in to k blocks. One of the blocks is left out from classifier design which is performed on the remaining k-1 blocks and the resulting classifier is tested on the remaining block to produce an error rate estimate. This procedure is then repeated k times. The mean of the individual error estimates is an unbiased estimator of how well the particular learning algorithm performs on the problem.

There are a number of problems with this strategy however. For instance, commonly only the mean error is reported, should the variance be large it indicates that there is a high risk of building a bad classifier. Also, although the test sets are independent, the classifiers tested are not since they all share k-2 blocks of data with other classifiers tested. Furthermore, if k is small in comparison to the number of samples, the performance estimate may very well be pessimistic: performance increases greatly with increasing design sample size for small design sets. On the other hand if k is taken equal to the number of samples, a special case called leave one out cross-validation, the classifiers will become very similar and consequently the performance estimates correlated.

In Paper IV we investigate a different route for obtaining better performance estimates than what a straight-forward hold-out test can provide for small sample sets. Specifically, we study whether tighter bounds can be obtained by updating the prior $P(q)$ with descriptive statistics obtained from three independent hold-out tests.

Unsupervised learning

A common task in bioinformatics is to identify subgroups within data. This can be accomplished using unsupervised learning algorithms that output a model of the data that identify relationships between samples and variables. Unsupervised learning algorithms in common use in bioinformatics are clustering algorithms such as k-means clustering and agglomerative hierarchical clustering (Hastie, Tibshirani et al. 2001).

In k-means clustering the algorithm's objective is to divide the samples into k coherent clusters by finding the partitioning of the samples that minimize the mean distance within the clusters. Each sample is initially assigned to one of the clusters (e.g. at random). Then each of the samples is reassigned from cluster i to cluster j if and only if the mean distance between the sample and other samples in cluster j is smaller than in cluster i. This is iterated until no sample can be reassigned or a limit on the number of iterations is reached. Of course, the distance function must be specified. Common choices for real valued features include the Euclidean metric and angular separation, for binary features the Manhattan distance is a natural choice. The main advantage of k-means clustering is the speed of the algorithm, the main drawback that the output depends on the initial assignment. It is good practice to check the output clusters for stability by re-running the algorithm with a different initial assignment.

Agglomerative hierarchical clustering algorithms sequentially clusters objects together by choosing the closest pair of objects, where objects may be either individual observations or clusters formed in a previous step. The process stops when all observations are joined into a single cluster. It is common to present the results as a binary tree which graphically represents the computational process, the dendrogram. Distance between pairs of observations is determined by the metric in use. The distance between two clusters is determined by another function, the linkage function. There are three linkage functions in wide-spread use: average, single and complete linkage. Average linkage function calculates the distance between two clusters as the average pair-wise distance between observations in one of the clusters to observations in the other cluster. Single linkage computes the smallest distance between any pair samples from the clusters, complete linkage the largest distance. It is well-known that cluster structure is greatly affected by the choice of linkage and metric function. There is a large literature available debating the appropriateness of different settings, but, by and large, the choice is arbitrary and left to the investigator.

High-Throughput Data Sources

A high-throughput analysis processes a large number of samples rapidly where as massively parallel analyses perform a large number of analyses simultaneously on a single sample. For instance, even though an mRNA microarray measures tens of thousands of transcript concentrations in a single analysis, the analysis of each sample can be quite laborious. However, high-throughput and massively parallel techniques alike generate vast amounts of information and similar computational challenges are faced in the analysis. In this thesis we have used primary data from mRNA microarrays, genome-wide location analysis and microculture cytotoxicity assays.

mRNA microarrays

By now, mRNA microarrays is a standard part of the molecular biology tool chest. Although there are a number of different approaches, the basic principle is the same: short segments of DNA are attached to a surface in spots (Schena, Shalon et al. 1995). A sample of mRNA is reverse transcribed and labeled with fluorophors and hybridized to the spots on the chip. Then the signal intensity of the fluorophors in each spot is recorded. The signal is roughly proportional to the amount of the corresponding mRNA in the sample.

Commercial interests have brought quality control in the production of chips and few laboratories produce their own microarrays today. With modern microarrays, reproducibility is high and only few vendors recommend technical replicates. The technology is still expensive however, running upwards 5 kSEK per chip. Thus the number of experiments that can be performed is limited within reasonable economic constraints. In addition, in e.g. clinical use, the number of samples is limited by the available material. This is a problem when the measurements are used for exploratory purposes. The risk of finding spurious correlations increases as the number of correlations tested grows.

Another issue is the samples themselves. Typically, one is only interested in a subpopulation of cells in the sample. Careful protocols are needed for sample preparation to not draw false conclusions, e.g. a marker for contamination of normal tissue in a tumor sample may appear linked to characteristics of the tumor. Techniques that allow the study of individual

cells such as laser capture microdissection (Emmert-Buck, Bonner et al. 1996) and single cell PCR (Emmert-Buck, Bonner et al. 1996; Fink, Seeger et al. 1998) are emerging, but those techniques are still expensive and time-consuming.

In Paper II and III we use cDNA microarray data from dividing *Saccharomyces cerevisiae* cultures (Spellman, Sherlock et al. 1998), in Paper I we analyze microarray data from two different cancer cell line panels (Dhar, Nygren et al. 1996; Weinstein and Pommier 2003).

Genome-Wide Location Analysis

Also called ChIP-on-chip, genome-wide location analysis is a technique for determining the location of binding sites for DNA-binding protein. The basic idea is to cross-link any DNA-binding proteins to the chromatin and hybridize the cross-linked complex to a DNA microarray. Labeled immunoglobulins targeting the DNA-binding proteins are then allowed to bind to the hybridized complexes, thus providing a view of where the transcription factor binds in the genome. Importantly, the analysis only provides information about the general vicinity of the binding site; the specific locus is not resolved. In Paper III we use the genome-wide location analysis data of 251 transcription factors in *Saccharomyces cerevisiae* (Lee, Rinaldi et al. 2002; Harbison, Gordon et al. 2004).

Microculture Cytotoxicity Assays

There are several different methods for measuring the cytotoxic and/or cytostatic effect of compounds in vitro, such as the MTT assay (Mosmann 1983) and fluorometric microculture cytotoxicity assay (FMCA) (Larsson and Nygren 1989). Both methods measure the cytotoxic effect of a compound by comparing the cell count in a treated culture to that in a control culture. Since actually counting the number of cells would be far too laborious a surrogate for the cell count is used. In the case of the FMCA method the conversion of fluorescein diacetate into its fluorescent derivate fluorescein by living cells is used. This allows the analysis to be carried out in a massively parallel high throughput fashion by growing cell microcultures in a microtitre plate. The activity of a compound is reported as either the fraction of cells surviving at some fixed concentration or the estimated concentration for which half of the cells die (known as the inhibitory concentration 50, IC50).

In Paper I we use drug-response data for cytotoxic compounds generated by using the FMCA method on a panel of cancer cell lines (Dhar, Nygren et al. 1996) as well as drug-response data generated in a screening program

using the MTT assay at National Cancer Institute, USA (Alley, Scudiero et al. 1988; Shoemaker, Monks et al. 1988).

Applying Domain Knowledge in Integrative Analyses

This thesis investigates different ways that heterogeneous data obtained from high-throughput analyses can be merged to analyze and generate sophisticated hypotheses about the molecular underpinnings of biological systems. Merging of heterogeneous data is commonly called *data fusion*, a term that originated in signal processing where the heterogeneous data is collected from different sensor systems in order to make better decisions than could be made by using only a single sensor system. A related term is integrative analysis, used to denote analyses that merge different types of data. In our integrative analyses we focus on how to use domain knowledge, specifically how to create algorithms that can help the investigator interpret data in view of existing knowledge. Below I will highlight these aspects of each of the Papers as well as summarize the main results.

Genome-Wide Correlation analysis of Gene expression and Chemosensitivity

In Paper I we study data fusion between drug activity and gene expression in the context of cancer chemotherapy. It has been shown that the cytotoxic activity profile of a compound across a cancer cell line panel can be used to identify the compounds' mechanism of action by comparing the profile to profiles of compounds with known mechanisms of action (Dhar, Nygren et al. 1996). Identifying the mechanism of action is an important problem since high-throughput screening of chemical libraries for cytotoxic activity is a routine operation that generates many leads for future therapies. In parallel, mRNA microarrays analysis is being used to find molecular pathways whose regulation reflects cancer drug response. The underlying idea is that correlations between drug activity and gene expression in a cancer cell line panel can identify molecular pathways conferring resistance or sensitivity (Rickardson, Fryknas et al. 2005). Many drugs depend on the same chemoresponse mechanisms, and furthermore, many genes are co-regulated. Thus, it is only natural to seek to lift the analysis to correlations between *subsets* of similar genes and drugs. A visual method for subset analysis was suggested by Weinstein and coworkers who used it to find a set of compounds that were actively transported out of the cells by Pgp/Mdr-1, a well-known system for cellular detoxification (Weinstein, Myers et al. 1997). In brief their method was as followed: the genome-wide pattern of correlations for all pairs of drug sensitivity and gene expression profiles is displayed in an array of colored blocks where the Pearson correlation coefficient between drug activity and gene expression is indicated by the color of the blocks. Each row corresponds to a drug and each column to a gene. The rows and columns correspond to the order of the leaves in dendrograms obtained from hierarchical clustering of the correlations between drugs and genes, respectively. Using this presentation makes it possible to visually identify groups of similar genes and drugs with strong correlations between the corresponding drug sensitivity and gene expression profiles as a coherent region in the map of correlations as shown in Figure 4.

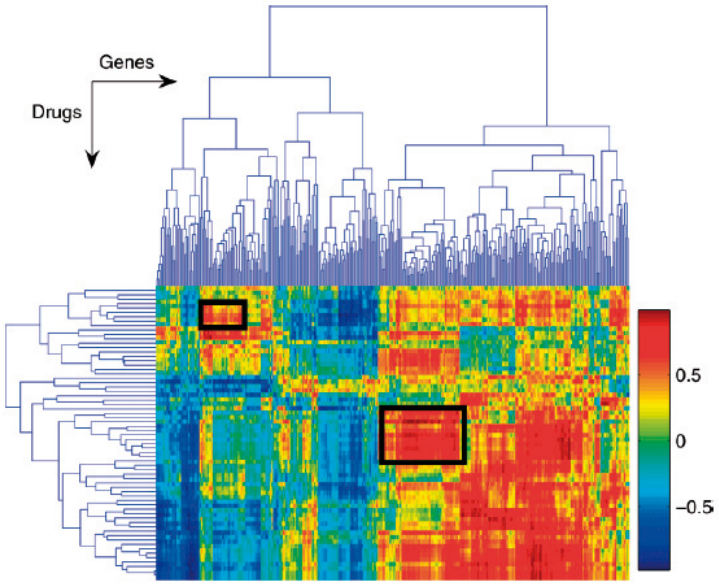

Figure 4. Example of a visualization of gene expression-drug activity correlations that can be used for identifying associations between drugs and molecular pathways. The framed rectangles denote examples of related drugs that all correlate to a set of co-regulated genes. See text for details.

However, since the clustering is based on the correlation coefficients there is little information in the coherent regions. In fact, they are likely to appear even if the primary data contained only noise. Furthermore, several researchers have pointed out that drugs with similar mechanisms of action tend to cluster together whether clustered based on the correlations with a set of gene expression profiles or directly based on drug sensitivity profiles. However, the resulting dendrograms are not identical (Scherf, Ross et al. 2000; Dan, Tsunoda et al. 2002).

In Paper I we demonstrate that the distance between two drugs described by the correlations of their activity profiles with a set of genes' expression patterns is mathematically equivalent to using a different distance function (metric) for the original activity profiles that depends on the gene expression data. We then analyze this new metric and show that by clustering correlation coefficients instead of primary data, statistical dependencies due to genetic relationships between the cell lines are reinforced and information is lost. Thus we recommend that the visualization is based on independent clustering of drug activity profiles and gene expression respectively, and that if there is domain knowledge suggesting that there are strong correlations between cell lines, a more informative analyses could be obtained by employing the Mahalanobis distance (Webb 2002).

Using Semantics of Time Profiles: Applications to the *S. cerevisiae* Cell Cycle

In Papers II and III we study the cell cycle in the budding yeast *Saccharomyces cerevisiae*. The organism is of great economical value because of its use in fermentation processes, particularly for brewing and baking. However, it is also a very important as a eukaryotic model organism, and this is also how we view it: as a simple, molecularly well annotated model organism for evaluating our analysis methods.

In many microarray studies the gene expression patterns recorded are clustered using some unsupervised learning algorithm. The resulting clusters are then inspected to see if some biological insight can be delivered by cross-checking against domain knowledge such as annotations of gene function to see if some gene function is overrepresented in the cluster. It is at this stage the data gains semantics, that is, starts to have some meaning in the view of the investigator. In the analyses of Paper II and III we define which groups of genes are relevant to the system ourselves in a problem specific manner: we use time-series data to tell us whether a gene is periodically or not during the cell cycle under a given experimental condition. The fact that a gene is periodically expressed implies that it is regulated by the cell cycle. It is this implied meaning of periodic expression which motivates our studies of associations between sequence motifs and periodic expression. In Paper II this distinction is quite trivial, however, in Paper III it becomes important as it facilitates the generation of sophisticated hypotheses about cell cycle regulation.

Assigning Semantics to mRNA Microarray Time Profiles: Bayesian Inference for Periodicity Detection

One of the first applications of microarrays was to identify genes involved in the cell cycle of *S. cerevisiae* (Spellman, Sherlock et al. 1998) and this was done by identifying genes that were periodically expressed during the cell cycle. Since the original analysis, many different algorithms have been proposed for the purpose of identifying periodically expressed genes most of which rely on supervised learning methods, that is they require a training set consisting of genes known to be periodically expressed. Thus we saw a need for a detector that did not require a training set but was able to use an estimate of the period time. Such a detector would be valuable for cell cycle studies in poorly annotated organisms where the relevant genes to be used for training aren't known. Furthermore, there are other periodic processes that are less well studied than the cell cycle, such as circadian rythms and glycolytic oscillations. In Paper II we develop a model-based detector of periodicity in the Bayesian formalism that is able to use uncertain information about the period time of the process. In the case of the cell cycle this information would be an estimate of the cell division time. We demonstrate its applicability on simulated and compare it to two other detectors that do not require a training set (Wichert, Fokianos et al. 2004; de Lichtenberg, Jensen et al. 2005). Our detection algorithm has been independently benchmarked by Adhesmäki and coworkers (Ahdesmaki, Lahdesmaki et al. 2007) who confirmed that it was optimal for the input signals we assume in our derivation, but that it was quite sensitive to outliers in the data. This might offer an alternative explanation for the weak performance we see on a set of benchmark sets (de Lichtenberg, Jensen et al. 2005) which we initially attributed to our unbiased analysis (in contrast to the supervised methods we did not use any information from these benchmark sets).

We apply the detector to microarray time series from synchronized yeast cultures and investigate whether previously described upstream sequence motifs (Hughes, Estep et al. 2000) could account for the periodicities observed. This was done by analyzing whether our detector of periodicity could predict the presence of different upstream sequence motifs. In doing so we merged two very different data sources, DNA sequence and quantitative time series of gene expression. The genes detected as periodically expressed were found to have a statistically significant overrepresentation of known cell-cycle regulated sequence motifs. One known sequence motif and 18 putative motifs, previously not associated with periodic expression, were also overrepresented.

In the same manner we analyzed functional annotations to cell cycle and periodicity, i.e. whether periodicity could predict domain knowledge about cell cycle involvement. The domain knowledge we used took the form of

Gene Ontology annotations (Ashburner, Ball et al. 2000). In practice an ontology is a controlled vocabulary of terms that may be combined according to a strict syntax. The Gene Ontology has three different branches: localization, molecular function and biological process. The terms are organized in a top-down fashion[2] and a distinction is made between lower level terms that are parts of the parent term e.g. M-phase is *part of* the cell cycle or whether the relationship constitutes a subclass e.g. the insulin like growth factor 1 receptor *is a* tyrosine kinase. Little studied entities will be annotated with more general terms. In addition, each term carries a tag denoting the type of evidence used for the annotation (literature, inferred by computation, inferred by a mutant phenotype et cetera). When we applied our detector to mRNA expression time profiles from *S. cerevisiae* shows that the genes detected as periodically expressed only contain a small fraction of the genes annotated to the biological process of cell cycle as inferred from mutant phenotype. For example, when the probability of false alarm was equal to 7%, only 12% of the cell cycle genes were detected.

When a labeled dataset of genes that are periodically expressed is available, it would make sense to use that information in a detector based on supervised learning. However, as shown by simulations, the detector we propose is useful in situations when the only domain knowledge available is vague prior information about the period time of the process for which one wants to find the relevant genes.

Revealing Cell Cycle Control Mechanisms

A seminal article by Beer & Tavazoie showed the possibility of predicting gene expression from upstream sequence features (Beer and Tavazoie 2004). The models they constructed generated testable mechanistic hypotheses on gene regulation and their paper stimulated research into alternative methods for predicting gene expression from sequence. For instance, Hvidsten and Wilczynski with coworkers used a rough set model to explain mRNA expression clusters in terms of the presence of upstream sequence motifs (Hvidsten, Wilczynski et al. 2005) and transcription factor binding data obtained from ChIP-on-chip data (Wilczynski, Hvidsten et al. 2006). The main drawback of these approaches and others is that the generated hypotheses take the form of rules stating 'IF *gene has some genomic feature* THEN *the gene's expression pattern clusters into some cluster*'. Thus the hypotheses have no useful biological meaning a priori; meaning must be inferred by finding general characteristics of the genes that are members of 'some cluster' by using e.g. Gene Ontology (Ashburner, Ball et al. 2000) annotations.

[2] Formally, the Gene Ontology is a directed acyclic graph (DAG).

In Paper III we demonstrate how more useful hypotheses can be generated by assigning semantics to the time profiles a priori. In our prototype application we study the regulation of periodically expressed genes in the yeast *S. cerevisiae*. Specifically, we target cell cycle regulation using prior knowledge about the shape of the time profiles that are characteristic to the process and study the well known phenomenon that there are genes that only appear as periodically expressed during the cell cycle when some synchronization methods are used (Shedden and Cooper 2002). This is done by dividing the genes into different classes depending on which synchronization methods produced a detectable periodicity. Each gene is described by novel descriptors of *cis*-acting regulation that are based on statistical associations between upstream sequence motifs inferred from sequence (Hughes, Estep et al. 2000) and experimentally determined transcription factor binding sites (Lee, Rinaldi et al. 2002; Harbison, Gordon et al. 2004). These novel descriptors thus integrate two heterogeneous data sources: sequence derived motifs and ChIP-on-chip data. This allows us to generate sophisticated hypotheses that suggest which combinations of transcription factors binding and sequence motifs effect cell cycle regulation when a particular synchronization method is used. We are able to demonstrate that targeting periodically expressed genes enriches the model with more known cell cycle regulators than when clustering is used for grouping the genes. Furthermore, when analyzing the combinations of transcription factors and sequence motifs we find evidence for a hierarchical additive structure of gene regulation. The presence of this structure in the organization of gene suggests it is less rich than the initial studies that found diverse and complex rules (Beer and Tavazoie 2004). Indeed it was recently demonstrated that it is possible to generate predictions as accurate as those of Beer & Tavazoie using a model that doesn't take combinatorial information into account (Yuan, Guo et al. 2007), in other words a less rich model than that originally used by Beer & Tavazoie.

The generic structure of our method could be used study any process where there is prior knowledge about time profile shapes, such as e.g. in infections which proceed through discernable phases and generates rich hypothesis. Furthermore, since the rough set classifier we employ has features which allows the investigator to constrain how the subsets used to discriminate between different classes of genes are computed we believe it could be a useful tool in future studies.

Improving Error Rate Estimation

It has been pointed out that many of the performance estimates reported from tumor classification studies using mRNA microarrays are so uncertain that it cannot be excluded that the classifiers perform no better than random guessing (Simon, Radmacher et al. 2003). Given the great interest in developing diagnostic (Fryknas, Wickenberg-Bolin et al. 2006), prognostic (van 't Veer, Dai et al. 2002) and predictive (Hess, Anderson et al. 2006; Potti, Dressman et al. 2006) tests using microarray analysis it is crucial to obtain good performance estimates of the classifier before a decision could be made to put it into clinical use.

Theoretical approaches have made significant progress towards determination of bounds on the error rate of supervised classifiers. However, the estimates obtained from a conventional holdout test using Bayesian inference still deliver tighter bounds than these new approaches. For sample sizes less than a few hundred and no prior knowledge about the true performance even the Bayesian estimates become unacceptably uncertain in many applications.

In Paper IV we use simulations to show how improved estimates can be obtained based on the maximum entropy principle. These intervals, maximum entropy empirically based credibility intervals (MEECIs), are based on the results from a few non-overlapping designs and tests which provides information about how well the classification algorithm performs on the particular problem domain. This domain knowledge may then be used as a prior in the Bayesian framework when obtaining the final estimate. In practice, the improvement can be used to reduce the uncertainty about the unknown performance. Alternatively, the improvement can be used to keep a fixed level of uncertainty based on a smaller number of examples.

Extracting Information from Unlabeled Data

The first problem faced when designing a classifier is feature selection. Selecting features is a particularly pressing problem when designing classifiers from microarray data. There are thousands of features to consider, only a small subset of which discriminates between the classes. Since there are far fewer samples than features the risk of including non-informative features in the classifier is high. In Paper V we investigate a method for eliminating non-informative features when there is additional unlabeled data available. The method is tailored for the situation when the classifier should be applied to data from a different distribution than the design data. This situation is faced when designing predictors of chemosensitivity from cell line panel microarray data that are to be applied to patient samples (Potti, Dressman et al. 2006; Lee, Havaleshko et al. 2007). If successful, a patient's chemotherapy could be tailored on an individual basis.

The reason for designing the classifier using cell line data is that it is possible to measure the effect of a single drug using cell lines where as in the clinic, patients are almost exclusively treated using a combination of drugs, and there is no way of determining which drugs were actually effective. However, one would not expect all of the genes that discriminate between sensitive and resistant cell lines to be relevant for prediction in patient material so there is a need for methods that can select only the features relevant in the patients. Furthermore, many publicly available datasets are not labeled, and even if there is labeled data available one typically has to use it all for estimating the performance of the classifier. Thus there is a need for algorithms that can design a classifier tailored for the patient samples without using the class labels. The unlabeled dataset thus constitutes prior knowledge about the distribution of gene expression which the algorithm is able to use for selecting relevant features.

The algorithm we study uses a list of candidate features obtained by supervised feature selection on the design set. It then constructs a classifier using the candidate list and predicts class labels on the unlabeled data set. A new list of candidates is computed by applying the supervised feature selection scheme to the unlabeled data using the predicted class label as true labels. Those which were also selected initially are saved, and a new classifier is built using those features in the design data. The process is then iterated, adding features from the initial selection that also discriminate between predicted class labels. In this manner we are able to integrate unlabeled data into a supervised feature selection scheme. We demonstrate using simulated as well as real data that the proposed method can eliminate false positives and in some cases dramatically improve classifier performance.

Final comments

We have proposed new methods for learning molecular biology that provide better analyses and estimates by integrating domain knowledge into the analysis. This is important since the throughput of biological analyses is increasing and automated analyses are called for, analyses that can help interpret data in the view of the investigator's prior knowledge. In the words of Theodosius Dobzhansky:

> Scientists often have a naive faith that if only they could discover enough facts about a problem, these facts would somehow arrange themselves in a compelling and true solution.

The subject of this thesis is bioinformatics, a relatively young, inter-disciplinary field of research. Bioinformatics attracts researchers from many other fields such as computer science, statistics, signal processing and machine learning who find interesting problems to which they can apply their methods. At the fringe of bioinformatics we find the consumers of their results: the experimentalists. Due to the language barriers that arise when different disciplines intersect much research has been devoted to what can be done rather than what should be done. However, I believe the language barriers between experimentalists and bioinformaticians are starting to break down and soon even high-level bioinformatics analyses will be employed in everyday research by experimentalists.

Svensk sammanfattning

Dagens molekylärbiologer har tillgång till kraftfulla mättekniker som kan mäta tusentals molekylärbiologiska analyter parallellt. En av de första metoderna som blev allmänt tillgänglig var mRNA-mikroarrayen som kan mäta ett provs innehåll av tiotusentals olika mRNA i en enda analys. Sedermera har ytterligare massivt parallella metoder tillkommit. Till exempel ChIP-on-chip där kromatinimmunoprecipiteringstekniken som används för att identifiera var i arvsmassan en transkriptionsfaktor binder parallelliserats med hjälp av mikroarraytekniken, arraybaserad komparativ genomhybridisering som kan detektera förändringar i geners kopietal samt SNP-arrayer som kan identifiera polymorfier i arvsmassan. Dessa massivt parallella tekniker brukar användas för att generera hypoteser om vilka molekylära system som ger upphov till en fenotyp eller är inblandade i en sjukdoms etiologi. Till exempel kan cellprover från sjuka och friska individer jämföras med hjälp av mRNA-mikroarrayer för att se vilka geners uttryck som skiljer sig mellan sjuka och friska. På så sätt skapas en bild av vilka gener som är inblandade vilken sedan kan testas i uppföljande experiment. Denna avhandling som består av fem delarbeten undersöker hur data som genererats med hjälp av massivt parallella tekniker kan kombineras för att skapa mer sofistikerade hypoteser om den underliggande biologin. Eftersom så många hypoteser kan genereras är det viktigt att kunna begränsa analysen till hypoteser som är relevanta givet den bakgrundskunskap som finns om biologin. Bakgrundskunskapen kan antingen ta formen av annoteringar av gener där tidigare experiment har utrett deras funktion eller vara resultatet av en tidigare upparbetning av relevanta data. Våra tillämpningar fann vi inom två besläktade forskningsfält: i två av delarbetena studerar vi cellcykelns reglering, i de resterande tre delarbetena avhandlar vi frågeställningar som uppstått i samband med utveckling och val av kemoterapi mot cancer (en sjukdom som beror på felaktig cellcykel-reglering).

I det första delarbetet undersöker vi en metod som används för att identifiera samreglerade gener som ger upphov till resistens mot eller är nödvändiga för en lyckad behandling med en grupp av cytostatika. I laboratoriet studeras cytostatika med hjälp av cellinjemodeller. Genom att mäta den cytotoxiska effekten av en substans i flera olika cellinjemodeller och korrelera den mot respektive cellinjes genuttrycksmönster kan gener

inblandade i kemoresponsen identifieras. När flera substanser testas kommer varje substans ge upphov till ett korrelationsmönster gentemot genuttrycket i cellinjerna. Liknande substanser som har samma verkningsmekanism kommer att ge upphov till liknande korrelationsmönster. Genom att visualisera dessa korrelationsmönster är det möjligt att identifiera substanser med samma verkningsmekanism och samtidigt associera dem till en grupp samreglerade gener. I vår undersökning visar vi att eftersom analysen inte tar hänsyn till släktskapet mellan cellinjerna kan resultatet vara missledande. Detta behov av att ta hänsyn till bakgrundskunskap är ett återkommande tema.

Delarbete två och tre studerar cellcykelns reglering i jästsvampen *Saccharomyces cerevisiae*, ett ofta använt modellsystem för eukaryota celler. I delarbete två föreslår vi en periodicitetsdetekor som kan användas för att identifiera gener som är cykliskt uttryckta. Vår detektor utvecklades för att använda bakgrundskunskap om periodtiden för att avgränsa möjligheten att en gen är cykliskt uttryckt. Vi utvärderar detektorn på simulerade data och tillämpar den sedan på mRNA-tidsserier (mätta med mikroarray) från synkroniserade jästkulturer. Sedan analyserar vi huruvida tidigare föreslagna inbindningsplatser för transkriptionsfaktorer i jästsvampens arvsmassa kan förklara varför vissa gener detekteras av periodicitetsdetektorn. På så sätt generar vi hypoteser om vilka inbindningsplatser som används för att reglera genuttryck under cellcykelns gång. Vi undersöker också huruvida bakgrundskunskap i form av tidigare känd inblandning i cellcykeln kan förklara geners cykliska uttryck. I det tredje delarbetet skapar vi nya särdrag för generna med hjälp av associationer mellan inbindning av transkriptionsfaktorer och sekvensmotiv i arvsmassan. Sedan studerar vi huruvida dessa särdrag kan förklara fenomenet att vissa gener endast förefaller vara cykliskt uttryckta när en viss (eller vissa) synkroniserings-metod(er) används. På detta sätt genereras hypoteser om hur cellcykeln regleras. En viktig skillnad mot tidigare föreslagna metoder är att dessa hypoteser uttrycks i termer som är meningsfulla för experimentalisten: varje hypotes föreslår vilka transkriptionsfaktorer som behöver binda var i närheten av gen för att den ska bli cykliskt uttryck under ett givet experimentellt förhållande.

I delarbete fyra och fem återvänder vi till sammanhanget kemoterapi mot cancer, om än våra fynd är mer allmängiltiga. Problemen som avhandlas uppstår i samband med utveckling av prediktiva test för kemoterapisvar som baseras på multivariat analys av mRNA-mikroarraydata. Eftersom mikroarrayanalyser är kostsamma finns det vanligtvis få prov tillgängliga för att utveckla och testa det prediktiva testet. En avvägning måste göras mellan hur god uppskattning av testets prestanda kan göras och hur bra testet kommer att bli (multivariata metoder kräver i regel mycket data för parameteranpassning).

Delarbete fyra behandlar problemet med att uppskatta hur bra ett prediktivt test kommer att fungera. Specifikt så undersöker vi huruvida information från flera oberoende test på data som använts för att designa det prediktiva testet kan användas för att bättre uppskatta testets prestanda. Informationen uttrycks som en sannolikhetstäthetsfunktion som speglar vilket prestanda som är troligast och kan användas för att avgränsa vilka prestanda som är troliga när ett slutgiltigt test utförs på testdata. Med simulerade data visar vi att metoden vi använder kan ge rättvisande och förbättrade estimat av prestanda.

Avslutningsvis så undersöker vi i delarbete fem en metod som kan använda data från prover med där behandlingsutgången är okänd (omärkta prover) för att välja ut relevanta gener vid konstruktionen av prediktiva test från mikroarraydata. En sådan metod är nödvändig då man önskar skräddarsy en kemoterapi utifrån ett prediktivt test eftersom data som används för att konstruera testet av nödvändighet kommer från cellinjemodeller, men testet ska användas på prover från patienter. Anledningen till att testet måste konstrueras baserat på cellinjemodeller är att det saknas data på genuttryck för patienter som behandlas med enstaka läkemedel eftersom man i praktiken alltid använder flera läkemedel i kombination. Därmed vet man inte vilka av patienterna som hade nytta av vilket läkemedel i kombinationerna. Vidare är det troligt att prover från cellinjemodeller skiljer sig från patientprover eftersom vissa av generna som förefaller vara prediktiva i cellinjemodellerna kanske inte ens är uttryckta i patientproverna. Vi föreslår och utvärderar en lovande metod för att välja ut de gener som är prediktiva i både cellinjer och patientprover med hjälp av omärkta patientprover som är mer tillgängliga än märkta patientprover.

Acknowledgements

I'd like to extend my sincere gratitude to:

My main supervisor, Jan Komorowski for giving me this opportunity and for providing a stimulating research environment as well as allowing me to explore research with my other collaborators.

My co-supervisor Mats Gustafsson. Your never ending enthusiasm for research and belief in me has been truly inspirational and instrumental to this thesis. I cannot express my thanks without getting sentimental so I'll simply shrug it off and leave it at that like real men do.

Anders Isaksson, for fulfilling my remaining needs for supervision. You have always been available for discussions of scientific things small and great and a valuable co-author. I never quite got that bird thing though.

Rolf Larsson, who has generously hosted me in the Cancer Pharmacology and Informatics group the last couple of years as well as provided many interesting research topics. And a Genesis album.

Torgeir Hvidsten, co-author extraordinaire for slugging it out with me and introducing me to the nobler aspects of grown men running around chasing a ball on a grass field. Forza!

Hanna Göransson, wizard of microarray bioinformatics. I don't think you realize how much your readiness to listen and discuss problems big and small has mattered to me over the years. I can only hope to have reciprocated.

Mårten Fryknäs, for our many interesting discussions. Far too many have been about science for us to be friends in private, far too few for us to be mere colleagues.

My room mates, Caroline Haglund, Linda Rickardson and Malin Wickström. Caroline, not only does your pysselskills rule, but you've never been anything but nice to me. I suppose you're the designated good cop… which

brings me to you Linda. Thanks for always reminding me that some people think värmländska is an awful sounding dialect. On a more serious note though, thanks for being a good co-author and contributing the screening data we used. And finally, speaking of serious notes… Malin: even though you haven't convinced me that depressing singer-songwriters can be a good thing, if they have anything to do with your personality I'm all for them. Seriously though, thank you all for answering all my questions about pharmacological things small and great. More importantly, you've made coming to work fun. I will miss you.

Gunilla Nyberg-Olsson and Christin Grønnslett at the LCB (presently and formerly respectively) for skillful administrative support and good company, even though your jobs are best done when gone unnoticed.

Present and past postdocs, PhD and masters students at the Linnaeus Centre for Bioinformatics for many stimulating discussions and just overall fun company, especially my 'contemporaries': Helena Strömbergsson, Stefan Enroth, Eva Freyhult, Alice Lesser, Johan Kåhrström, and Robin Andersson.

Present and past members of the Computational Medicine group, the Array Platform and staff at the Clinical Pharmacology group, who've made for a nice working environment. And yes, secretly I too prefer cozy lighting in the lunch room.

The BMC Computing Department: Emil Lundberg, Gustavo Gonzales-Wall and Nils-Einar Eriksson for occasionally providing technical support and more importantly, comic relief.

Finally, though it has nothing to do with this thesis:

Friends and foes: if you don't know who you are by now, you won't find out here.

To my family: Bo, Ann-Christine, Erica, Jörgen, Gustaf, and Elvira… I do love you all very much.

References

Ahdesmaki, M., H. Lahdesmaki, et al. (2007). "Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data." BMC Bioinformatics **8**: 233.

Alberts, B. (2002). Molecular biology of the cell. New York, Garland Science.

Albertson, D. G. and D. Pinkel (2003). "Genomic microarrays in human genetic disease and cancer." Hum Mol Genet **12 Spec No 2**: R145-52.

Alley, M. C., D. A. Scudiero, et al. (1988). "Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay." Cancer Res **48**(3): 589-601.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Beer, M. A. and S. Tavazoie (2004). "Predicting gene expression from sequence." Cell **117**(2): 185-98.

Breiman, L. (2001). "Statistical modeling: The two cultures." Statistical Science **16**(3): 199-215.

Buck, M. J. and J. D. Lieb (2004). "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments." Genomics **83**(3): 349-60.

Chee, M., R. Yang, et al. (1996). "Accessing genetic information with high-density DNA arrays." Science **274**(5287): 610-4.

Cox, R. T. (1946). "Probability, Frequency and Reasonable Expectation." American Journal of Physics **14**(1): 1-13.

Dan, S., T. Tsunoda, et al. (2002). "An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines." Cancer Res **62**(4): 1139-47.

de Lichtenberg, U., L. J. Jensen, et al. (2005). "Comparison of computational methods for the identification of cell cycle-regulated genes." Bioinformatics **21**(7): 1164-71.

Dhar, S., P. Nygren, et al. (1996). "Anti-cancer drug characterisation using a human cell line panel representing defined types of drug resistance." Br J Cancer **74**(6): 888-96.

Dressman, H. K., A. Berchuck, et al. (2007). "An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer." J Clin Oncol **25**(5): 517-25.
Emmert-Buck, M. R., R. F. Bonner, et al. (1996). "Laser capture microdissection." Science **274**(5289): 998-1001.
Fink, L., W. Seeger, et al. (1998). "Real-time quantitative RT-PCR after laser-assisted cell picking." Nat Med **4**(11): 1329-33.
Fodor, S. P., R. P. Rava, et al. (1993). "Multiplexed biochemical assays with biological chips." Nature **364**(6437): 555-6.
Fryknas, M., U. Wickenberg-Bolin, et al. (2006). "Molecular markers for discrimination of benign and malignant follicular thyroid tumors." Tumour Biol **27**(4): 211-20.
Gelman, A. (1995). Bayesian data analysis. London ; New York, Chapman & Hall.
Hanahan, D. and R. A. Weinberg (2000). "The hallmarks of cancer." Cell **100**(1): 57-70.
Harbison, C. T., D. B. Gordon, et al. (2004). "Transcriptional regulatory code of a eukaryotic genome." Nature **431**(7004): 99-104.
Hastie, T., R. Tibshirani, et al. (2001). The elements of statistical learning : data mining, inference, and prediction. New York, Springer.
Hess, K. R., K. Anderson, et al. (2006). "Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer." J Clin Oncol **24**(26): 4236-44.
Hughes, J. D., P. W. Estep, et al. (2000). "Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae." J Mol Biol **296**(5): 1205-14.
Hvidsten, T. R., B. Wilczynski, et al. (2005). "Discovering regulatory binding-site modules using rule-based learning." Genome Res **15**(6): 856-66.
Jaynes, E. T. and G. L. Bretthorst (2003). Probability theory : the logic of science. Cambridge, UK ; New York, NY, Cambridge University Press.
Kendall, M. G. (1949). "On the Reconciliation of Theories of Probability." Biometrika **36**(1-2): 101-116.
Larsson, R. and P. Nygren (1989). "A rapid fluorometric method for semiautomated determination of cytotoxicity and cellular proliferation of human tumor cell lines in microculture." Anticancer Res **9**(4): 1111-9.
Larsson, R. and P. Nygren (1993). "Prediction of individual patient response to chemotherapy by the fluorometric microculture cytotoxicity assay (FMCA) using drug specific cut-off limits and a Bayesian model." Anticancer Res **13**(5C): 1825-9.
Lee, J. K., D. M. Havaleshko, et al. (2007). "A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery." Proc Natl Acad Sci U S A **104**(32): 13086-91.

Lee, T. I., N. J. Rinaldi, et al. (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science **298**(5594): 799-804.
Mosmann, T. (1983). "Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays." J Immunol Methods **65**(1-2): 55-63.
Ohrn, A. and T. Rowland (2000). "Rough sets: a knowledge discovery technique for multifactorial medical outcomes." Am J Phys Med Rehabil **79**(1): 100-8.
Pawlak, Z. (1982). "Rough Sets." International Journal of Information and Computer Science **11**: 341-356.
Potti, A., H. K. Dressman, et al. (2006). "Genomic signatures to guide the use of chemotherapeutics." Nat Med **12**(11): 1294-300.
Rickardson, L., M. Fryknas, et al. (2005). "Identification of molecular mechanisms for cellular drug resistance by combining drug activity and gene expression profiles." Br J Cancer **93**(4): 483-92.
Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.
Scherf, U., D. T. Ross, et al. (2000). "A gene expression database for the molecular pharmacology of cancer." Nat Genet **24**(3): 236-44.
Shannon, C. E. (1948). "A Mathematical Theory of Communication." Bell System Technical Journal **27**(3): 379-423.
Shedden, K. and S. Cooper (2002). "Analysis of cell-cycle gene expression in Saccharomyces cerevisiae using microarrays and multiple synchronization methods." Nucleic Acids Res **30**(13): 2920-9.
Shoemaker, R. H., A. Monks, et al. (1988). "Development of human tumor cell line panels for use in disease-oriented drug screening." Prog Clin Biol Res **276**: 265-86.
Simon, R., M. D. Radmacher, et al. (2003). "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification." J Natl Cancer Inst **95**(1): 14-8.
Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell **9**(12): 3273-97.
van 't Veer, L. J., H. Dai, et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530-6.
Webb, A. R. (2002). Statistical pattern recognition. West Sussex, England ; New Jersey, Wiley.
Weinstein, J. N., T. G. Myers, et al. (1997). "An information-intensive approach to the molecular pharmacology of cancer." Science **275**(5298): 343-9.
Weinstein, J. N. and Y. Pommier (2003). "Transcriptomic analysis of the NCI-60 cancer cell lines." C R Biol **326**(10-11): 909-20.
Wichert, S., K. Fokianos, et al. (2004). "Identifying periodically expressed transcripts in microarray time series data." Bioinformatics **20**(1): 5-20.

Wilczynski, B., T. R. Hvidsten, et al. (2006). "Using local gene expression similarities to discover regulatory binding site modules." BMC Bioinformatics **7**: 505.
Yuan, Y., L. Guo, et al. (2007). "Predicting gene expression from sequence: a reexamination." PLoS Comput Biol **3**(11): e243.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 401

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, 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 Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-8477

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2008

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 314

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, 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 Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-8461

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2008