Computational Methods in Genomic and Proteomic Data Analysis

Johansson, Peter

2006

Computational methods in genomic and proteomic data analysis

Peter Johansson

Department of Theoretical Physics
Lund University, Sweden

Thesis for the degree of Doctor of Philosophy

Thesis Advisor: Markus Ringnér

Faculty Opponent: Lodewyk Wessels
Netherlands Cancer Institute

To be presented, with the permission of the Faculty of Natural Sciences of Lund University, for public criticism in Lecture Hall F of the Department of Physics on Friday, the 2nd of June 2006, at 10.15 a.m.

DOKUMENTDATABLAD enl SIS 61 41 21

Organization **LUND UNIVERSITY** Department of Theoretical Physics Sölvegatan 14A SE-223 62 LUND	**Document Name** **DOCTORAL DISSERTATION** **Date of issue** May 2006 **CODEN:**
Author(s) Peter Johansson	**Sponsoring organization**

Title and subtitle

Computational methods in genomic and proteomic data analysis

Abstract

With the great progress of technology in genomics and proteomics generating an exponentially increasing amount of data, computational and statistical methods have become indispensable for accurate biological conclusions. In this doctoral dissertation, we present several algorithms designed to delve large amounts of data and bolster the understanding of molecular biology. MAPK and PI3K, two signaling pathways important in cancer, are explored using gene expression profiling and machine learning. Machine learning and particularly ensembles of classifiers are studied in context of genomic and proteomic data. An approach to screen and find relations in protein mass spectrometry data is described. Also, an algorithm to handle unreliable values in data with much redundancy is presented.

Summary in Swedish

Med modern mätteknik kan vi mäta cellers egenskaper för alla gener samtidigt. För att tolka den stora datamängden krävs analysmetoder och datorverktyg. Den här avhandlingen behandlar ett antal sådana verktyg avsedda att klargöra geners och proteiners inbördes samband. En metod att hantera datavärden av varierande kvalité presenteras, såväl som ett verktyg att visualisera samband i masspektrometri-data. Klassificering och då speciellt ensemblemetoder diskuteras och används för att undersöka två signalvägar, MAPK och PI3K, som är viktiga i cancer.

Key words

BRAF, cancer, classification, ensemble, hierarchical clustering, mass spectra, microarray, missing values, PI3K, PTEN, support vector machine

Classification system and/or index terms (if any)

Supplementary bibliographical information — **Language** English

ISSN and key title — **ISBN** 91-628-6852-7

Recipient's notes — **Number of pages** 98 — **Price**

Security classification

Distribution by (name and address)

Peter Johansson, Dept. of Theoretical Physics,
Sölvegatan 14 A, SE-223 62 LUND

Signature _____ **Date** 2006-05-09

For Anna (1970-1989)

This thesis is based on the following publications:

I P. Johansson and J. Häkkinen
Improving missing value imputation of microarray data by using spot quality weights
LU TP 05-40

II P. Johansson and M. Ringnér
An evaluation of using ensembles of classifiers for predictions based on genomic and proteomic data
LU TP 06-19

III S. Pavey, P. Johansson, L. Packer, J. Taylor, M. Stark, P.M. Pollock, G.J. Walker, G.M. Boyle, U. Harper, S.J. Cozzi, K. Hansen, L. Yudt, C. Schmidt, P. Hersey, K.A.O. Ellem, M.G.E. O'Rourke, P.G. Parsons, P. Meltzer, M. Ringnér, and N.K. Hayward
Microarray expression profiling in melanoma reveals a *BRAF* mutation signature
Oncogene **23**, 4060-4067 (2004)

IV L.H. Saal, P. Johansson, K. Holm, S.K. Gruvberger-Saal, P.O. Bendahl, S. Koujak, P.O. Malmström, L. Memeo, M. Ringnér, H. Hibshoosh, Å. Borg, and R. Parsons
An *in vivo* gene expression signature for PTEN/PI3K pathway activation predicts patient outcome in multiple tumor types
LU TP 06-18

V R. Alm, P. Johansson, K. Hjernø, C. Emanuelsson, M. Ringnér, and J. Häkkinen
Detection and identification of protein isoforms using cluster analysis of MALDI-MS mass spectra
Journal of Proteome Research **5**, 785-792 (2006)

During my PhD studies, I also contributed to the following publications:

* L. Packer, S. Pavey, A. Parker, M. Stark, P. Johansson, B. Clarke, P. Pollock, M. Ringnér, and N. Hayward
Osteopontin is a downstream effector of the PI3-kinase pathway in melanomas that is inversely correlated with functional PTEN
to appear in *Carcinogenesis*

* M. Ringnér, P. Edén, and P. Johansson
Classification of expression patterns using artificial neural networks
In *A Practical Approach to Microarray Data Analysis*
(eds. D.P. Berrar, W. Dubitzky and M. Granzow,
Kluwer Academic Publishers), pp. 201-215 (2002)

”Vårt umgänge med andra människor består huvudsakligen i att vi diskuterar och värderar vår nästas karaktär och beteende. Detta har medfört att jag frivilligt avstått från praktiskt taget all så kallad samvaro. Härigenom har jag blivit en smula ensam på min ålderdom. Min livsdag har varit full av hårt arbete och det är jag tacksam för. Det började som slit för brödfödan och slutade som kärlek till en vetenskap.”

Isak Borg

Contents

Introduction

"It's like driving a car at night. You never
see further than your headlights, but you can
make the whole trip that way."
Edgar Lawrence Doctorow

Work is important. When we meet strangers, our first question is "What do you do?" We are not asking about what they do for leisure as much as we ask what they do as *work*. When defining and summarizing a person in a few words, only one question may be more important: "Is that a miss or a bloke?" Of course, this latter question is not very often asked verbally. Most people would probably be offended if you questioned their sex, and in most cases a quick look is enough to reveal the answer anyway. Telling the profession of a person from a quick look is trickier though (unless she wears a uniform). And asking directly may be risky, because what you think is a good ice-breaker may just be an opening down to an icy-cold hole of water. Either your new friend starts whining about some kind of luxury problem such as colleagues stealing her ketchup or colleagues refusing to brew her daily cup of coffee. Or, if she is not that obsessed with work, she probably categorizes you as shallow, since she expects a socially skilled person to come up with something slightly more sophisticated than this cliché question.

When people ask me what I do for a living, I have three standard answers. Sometimes, I briefly answer: "Well, I'm a PhD student... at the Department of Theoretical Physics". Nineteen of twenty people respond with horror in their eyes and direct the conversation to something completely different. The twentieth person explains that physics is so amazingly interesting and starts to ask questions like "If the universe is finite, what is then outside?", "Is the cat dead or alive?", "How come, throwing tepid water on the aggregate, makes the sauna warmer?", or "Is one kilogram of ice more than one kilogram of water?". The twentieth person is so enthusiastic, it would be heart-breaking to explain

I'm not doing any physics, so I rather try to answer the questions asked.

My second answer is more of an attempt to explain what I do, rather than describing where my computer and desk happen to be located. However, I find it difficult to boil down years of work to one sentence and when I try, it often results in something pseudo-understandable. A sentence containing words like cancer and statistics. "Cancer and statistics, aha", they think and take the opportunity to ask whether sun bathing really is dangerous.

When I feel really enthusiastic about work, I try to be frank and tell them "Ok, to describe decently what I do, I will need 10 minutes. Have you got 10 minutes?" People must be very stressed because they never have 10 minutes.

Have you got 10 minutes? Anyway, this introduction describes what I have been up to the last years. The introduction starts with some basic molecular biology, then follows a discussion on hypothesis testing and machine learning. The introduction ends with a summary of the five papers this thesis is based upon.

Molecular biology

"Je n'avais pas besoin de cette hypothèse-là."
Pierre-Simon Laplace

The atom of life is the cell. All living organisms, from the grass in the garden to the birds in the sky, are built from cells. Each cell consists of various molecules including water, nucleic acids, and proteins. Proteins are important because they catalyze chemical reactions as well as being the building blocks in different compartments of the cell. Nucleic acids are important because they carry and mediate the genetic inheritance.

The genetic inheritance is encoded in deoxyribonucleic acids (DNA). Chemically the DNA molecule is a helix composed of two strands that are long chains of nucleotides with the bases adenine (A), cytosine (C), guanine (G), and thymine (T). These bases form complementary base pairs between A and T and between C and G, respectively, with one of the bases in each strand (Figure 1). Thus, any DNA molecule can be specified by a sequence of letters from a four-letter alphabet [1].

A key feature of DNA is the ability to replicate. Replication starts with the two strands being separated. Each of the two single strands works as template

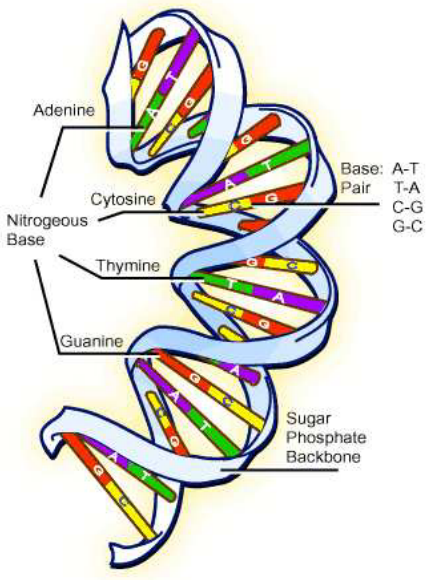

Figure 1: The DNA molecule consists of two helical strands connected via base pairs A-T and C-G, respectively. Reproduced with permission (Jane Wang) ©2006 bioteach.ubc.ca.

for the formation of a new DNA molecule. Nucleotides are added sequentially in such a way that base pairs form and thus the new DNA molecule is a perfect copy of the original. In this way the genetic information is transfered from mother cells to daughter cells, and from parents to their children. In higher organisms, the DNA is found in the nucleus of the cell, wherein it is packed in units called chromosomes and twisted around positively charged proteins called histones [2]. The DNA contains thousands of genes, specific sequences of nucleotides, serving as recipes for how to build a protein. The recipe is transmitted via an intermediate molecule, messenger ribonucleic acid (mRNA), very similar to the DNA molecule.

Although each cell in an organism has the same DNA, different types of cells do not look the same. Different patterns of genes being active lead to different proteins being produced giving each cell its specific qualities and functions. For example, the insulin gene is active in the pancreas and insulin is produced, whereas in all other organs the insulin gene is silenced. When a gene is active, *i.e.,* it is expressed, it works as a template for creating an mRNA strand in the same manner as it works as template for a new DNA strand during repli-

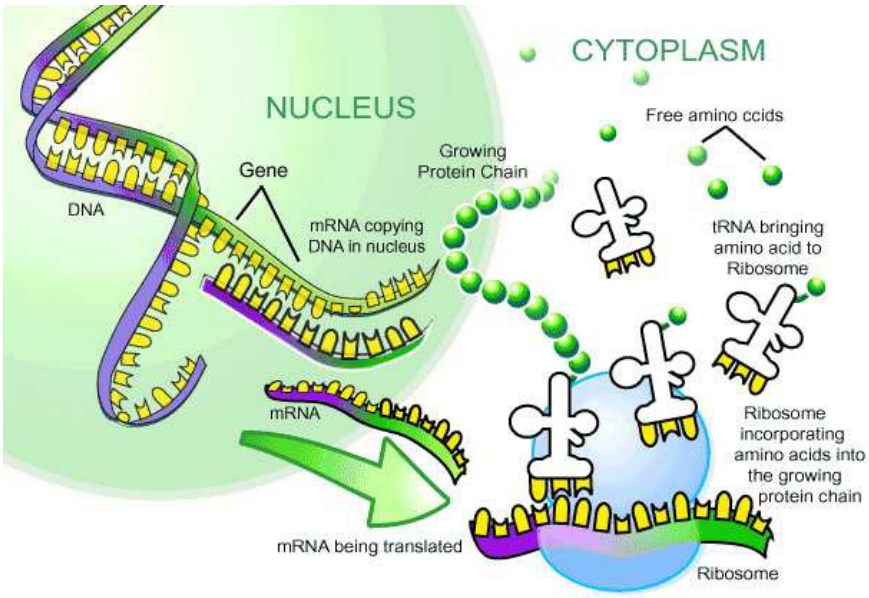

Figure 2: When a gene is expressed, DNA in the nucleus is transcribed into mRNA, which is transfered to ribosomes in the cytoplasm where it is translated into proteins. Reproduced with permission (Jane Wang) ©2006 bioteach.ubc.ca

cation [3]. This mRNA strand is moved from the nucleus to the ribosomes in the cellular cytoplasm where it serves as a template in protein production (Figure 2).

The ribosome is a neat little complex built from proteins and another kind of RNA called ribosomal RNA. Yet another kind of RNA, transfer RNA (tRNA), carry in amino acids. These complexes of tRNA and amino acids bind to the mRNA, and thereby the amino acids are attached to each other building a protein chain. As any combination of three tRNA molecules binds to a specific amino acid, the sequence of the mRNA uniquely defines what protein is produced.

The proteins are important because they are the doers in the cell. They have various roles including being building blocks in the cell; receptors in the cell membrane transmitting signals from outside to the inside of the cell; enzymes catalyzing chemical reactions in the cell; as well as being regulatory proteins. Regulatory proteins bind to the DNA and block a gene [4]. Alternatively, the protein might activate a gene, in other words, it triggers the gene to produce

mRNA [5–7]. This mRNA in turn serves as template for a protein, which may be an activator or blocker of another gene and so forth and so on. Activating one gene may result in a cascade of activated and deactivated genes, respectively, and one could picture these cascades as genes interacting in a large network.

Cancer

"I don't give a damn what the people say
I'm gonna do it, gonna do it my way
Gonna let it all out an do my thing
Boom boom boom an a bang bang bang"
Felix Buxton & Simon Ratcliffe

We are all made of cells - billions of cells, and every single cell is programmed to perform its specific functions. The cells are social in the sense that each cell knows its role and they work together in a complex network that is regulated by a sophisticated signaling system.

However, sometimes a cell breaks out from this system and behaves as bad as a rebellious teenager. A cancer cell is created that ignores the signals from the regulation system and starts to focus on one thing only, replication. It multiplies itself frenetically and as its daughter cells inherit the behavior, after a while there is a significant group of rebellious cells. Just like the teenager, after some time this group of cells gets the idea that home is sweet but not sweet enough. They start moving and spreading into other tissues. Their behavior is now more martial and asocial as they ignore the fact that they damage the tissue they infiltrate and invade. Eventually, they break into the transport system of the body and use it to migrate and colonize other parts of the body. Secondary tumors, metastases, arise, and if these tumors are not killed or removed, the normal cells will be so seriously damaged that the body cannot survive.

Taking the perspective of the cancer cells for a few moments, there are a number of obstacles we have to overcome. The whole idea of being a cancer cell is to multiply ourselves unimpededly, but the body has various defense mechanisms to prevent us from doing so [8]. The body sends signals telling us to kick back and relax a bit [9, 10]. We have no interest in calming down, so we need to be insensitive to these signals. If things get serious and we are considered a threat to the system, we will be told to go into apoptosis [11]. Apoptosis is just a paraphrase for suicide, which of course is unacceptable from our point of

view. We must avoid apoptosis, and can do that both by silencing those genes starting apoptosis, as well as activating anti-apoptosis genes. Our behavior is programmed in our genes, so we change our behavior by mutating important genes. Normally, cells have a system that checks for mutations and repair the DNA [12]. These guys are keeping back our purposes so we need to obstruct their work. Moreover, constant reproduction costs energy, so we need to start programmes to mobilize cell resources. All together, it is a long list of things we need to accomplish and will likely need multiple hits on the genome [13]. However, if we are supported by a couple of inherited gene defects, we are more likely to reach the ultimate goal of freedom and independence.

In breast cancer, for example, it is well-known that carrying a mutation in *BRCA1* [14] is a high risk factor. More than half of women carrying a *BRCA1* mutation will get cancer, whereas women without the mutation have a life time risk of 10% [15].

Genomic and proteomic expression data

"I like thinking big. If you're going to be thinking anything, you might as well think big."
Donald Trump

Until about ten years ago, studies of gene expression were limited to measuring gene expression levels of one or a couple of genes. With the microarray technology, a new tool was brought to the table allowing studies of thousands of genes in parallel. The underlying idea is that because mRNA molecules are instable and decay, the concentration of a specific mRNA reflects the activity of the corresponding gene. In order to measure the concentrations, the mRNA is extracted from the sample. By employing an enzyme, reverse transcriptase, the mRNA is transcribed into complementary DNA (cDNA). The cDNA is labeled by attaching a fluorescent molecule that absorbs and emits light at a specific wavelength. The cDNA is applied on the microarray, a small glass slide, on which thousands of spots have been printed. Each spot contains single stranded DNA matching a specific gene, and because of the base-pairing mechanism the applied sample cDNA binds to a specific spot containing the matching DNA. The microarray is then exposed to a laser beam that excites the fluorescent molecules, and by detecting and quantifying the emitted intensity from a spot, the amount of bound cDNA can be measured. Thereby, the gene expression can be determined for thousands of genes in parallel.

Peptide mass fingerprinting, first suggested by Yates and collaborators [16], is a strategy to identify many proteins in parallel. In short, trypsin is applied to the protein of interest, which results in the protein being cleaved at specific sites. The resulting mixture of peptides, protein fragments, comprise a unique identifier of the protein. The masses of the peptides are measured using a mass spectrometer that relies on the simple fact that heavy molecules accelerate slower than light molecules when exposed to an electrical field. In the spectrometer the peptide mixture of interest is mixed with a chemical called matrix and applied onto a metal plate. The matrix and peptide crystallize together on the metal plate and the metal plate is inserted into a vacuum chamber. The peptides are shot at by laser beams that promote the transition from solid phase to gas phase, after which the peptides accelerate in the applied electrical field and are detected in an ion detector, generating a histogram of time of flights. As heavier molecules accelerate slower, the histogram of time of flights can be translated into a histogram of masses. This histogram corresponds to a fingerprint of the protein and allows for identification of the protein by comparing it to theoretical fingerprints [17]. These theoretical fingerprints have been calculated by cleaving known proteins with trypsin theoretically and calculating the composition of peptide masses, the mass fingerprint.

Hypothesis testing

> "Information is not knowledge. Knowledge is not wisdom. Wisdom is not truth. Truth is not beauty. Beauty is not love. Love is not music and music is the best."
>
> *Frank Zappa*

Having measured the expression of all these genes and proteins is good, only a good start though, because without an interpretation of the data we have learnt nothing, and learning is what we are striving for, isn't it?

A standard question in microarray analysis is which genes are differentially expressed in two groups of biological samples. The groups may, for example, be samples from one kind of tumor versus samples from another kind, samples subjected to one kind of treatment versus samples subjected to another kind of treatment, or samples with a mutation in a specific gene versus samples without the mutation. This type of question is as old as statistics, and consequently the statistical literature is full of suggestions on how to measure the difference between two groups; for a review see [18]. Here, I will not go into details about

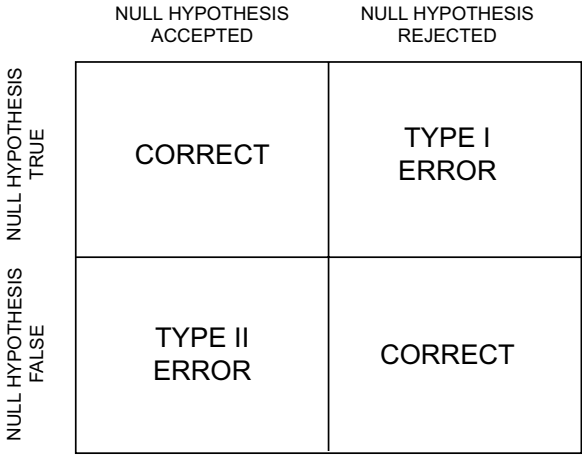

Figure 3: Illustration of the four possible results of a hypothesis test. A type II error occurs when the data is not strong enough to reject the false null hypothesis. A type I error occurs when a true null hypothesis is rejected. The significance level sets the balance between rejected and accepted and thereby the balance between type I and type II errors.

different methods, but sketch the basic concepts in hypothesis testing such as *null hypothesis*, *alternative hypothesis*, *significance level*, and *power*.

To describe these concepts I will use a very well-known example of hypothesis testing that is illustrated in tv series such as "LA Law", "Boston Legal", or "Perry Mason". Perry Mason, the hero of my childhood, is a lawyer who in every episode convinces the jury to "find the defendant not guilty", and the hypothesis testing I am talking about is of course the procedure of a trial. In a trial, the null hypothesis simply is the assumption that the defendant is innocent. In a scientific investigation, the null hypothesis often indicates that the treatment did not do anything or that the property of interest does not make a difference. The alternative hypothesis is the opposite, the hypothesis the researcher (believe in and) want to evaluate. In a trial the alternative hypothesis is the reason the defendant was arrested in the first place.

An important observation is that it takes infinite amount of evidence to prove a hypothesis, whereas it only takes one good piece of evidence to disprove it. For that reason it is every prosecutor's strategy to disprove the null hypothesis. If the null hypothesis is rejected, logically the jury will accept the alternative hypothesis and send the criminal to jail. The same strategy is employed in statistics. Given the evidence, the statistician calculates the probability the

evidence would appear this strong, if the null hypothesis were true. If this probability, the p-value, is small, the null hypothesis is rejected and consequently the alternative hypothesis is accepted. A standard threshold for rejection is a p-value cutoff of 0.05, which means that on average 5% of true null hypotheses are rejected. This is not perfect but means, what in statistics is called, type I errors occur. Alas, the same type of error occurs in the court. Although the null hypothesis shall only be rejected when evidence is convincing beyond reasonable doubt, it sometimes happen that innocent people are sent to jail. Most people find this error upsetting, but very few people would accept the only possible solution to avoid this travesty on justice. Because the solution is to re-write the law such that people are only sent to jail when we can be absolutely sure they are guilty, and being that strict means we cannot judge anyone, in other words, also criminals are set free. In statistics, accepting a false null hypothesis is referred to as a type II error. In a scientific investigation the balance between type I errors and type II errors may be set by the investigator, by choosing a significance level, *i.e.,* the threshold for the p-value. A smaller threshold leads by definition to fewer type I errors, and thus more type II errors. However, there are ways to decrease the number of type II error without changing the significance level. A trivial way is to collect more evidence in the first place and make the decision easier for the jury. Another way is to choose a jury that can interpret the data in a more clever and powerful way. This is applied in some legal systems, in which the jury is replaced by educated judges who know the law. In statistical testing this corresponds to choosing the most powerful test. A test is considered more powerful if it has less expected type II errors.

Another situation in which you apply hypothesis testing is when you play a good game of poker. Imagine you notice the new fellow around the table gets good cards a bit too often. Then you would ask yourself what the chances are he could get those cards by chance. If that chance is too small, it cannot only be good luck and the night might end with a smoking gun.

Do think twice though, before you shoot your new friend. The chance of getting the best hand, a royal straight flush, in one round may be small. However, if the night is getting late and you guys have played many rounds, the chance that one of your friends would get a royal in one of the rounds is not that small anymore. The same thing occurs in the microarray analysis. The chance that a specific gene gets a p-value less than say 0.01 is by definition only 1%. However, when we have measured 50,000 genes, the chance that at least one p-value is less than 0.01 is virtually 100%.

More exact, by pure chance we expect 1% of the genes to be discriminatory and have a p-value less than 0.01. Thus, a natural question is whether there

are more discriminatory genes than we would expect by pure chance. If there is a great overabundance of discriminatory genes, then the expression profiles of the two groups can be claimed to be different.

A more sophisticated way to investigate the difference between two groups is to employ machine learning methods. In machine learning an optimal decision rule is found by learning from data. This approach gives a more holistic picture than looking at a gene at a time. Methods such as nearest centroid classifiers [19], support vector machines [20], and artificial neural networks [21] have been found to be useful. When the machine manages to distinguish the groups this means there is a difference between the groups. If the machine fails, we can conclude the possible difference is more subtle. Another application for machine learning in this area is to really use the created predictors in clinical settings as a diagnostic tool.

Support vector machines

"Endast idioten har ett fritt val. Den intelligenta väljer det bästa."
Willy Kyrklund

In machine learning a machine is trained to distinguish training samples according to sample labels. A decision rule is found that may be applied on test samples to evaluate the machine, or the rule may be used to predict a sample with unknown sample label. The support vector machine (SVM) is a popular machine learning method. The embryo of what would become SVM was brought to the world in 1963 in the form of Vapnik's maximal margin classifier [22]. The method was later on improved by the usage of kernels [23], which made it applicable also on non-linear problems; and with the introduction of soft-margins [24] the method became famous under the name support vector machines.

The SVM method is built on kernel theory [25, 26], Kuhn-Tucker optimization theory [27], and Vapnik Chervonenkis risk minimization theory [28], which may frighten even the most enthusiastic newbie. However, as with cars, we do not need to understand the components to motivate the usage. Here, I will describe the basic properties of the SVM; for a more thorough introduction see [29].

For a linear classification method finding a classification rule is to find a hyperplane separating the two groups of training samples. In the first version of

SVM, the maximal margin classifier, the classification rule is found by considering two things. First, a condition for the classification rule is that the training samples are correctly classified, in other words, the found hyperplane does separate the two group of training samples perfectly. Second, among all hyperplanes fulfilling this condition, the hyperplane that maximizes the margin is chosen. The margin is the distance from the hyperplane to the closest training sample, and thus maximizing the margin is to maximize the width of the sample free strait around the decision hyperplane (Figure 4). Mathematically, this situation is equivalent to my favorite problem in mechanics. Imagine two parallel boards attached with numerous springs pushing the boards apart. However, when the boards reach certain points (the data points) forces are triggered in these points perpendicular to the board such that the boards never cross the points. For the static situation there are two obvious questions: 1) How are the boards positioned? 2) How large are the forces? The first question is obviously equivalent to finding the hyperplane in the maximal margin classifier, because in the static solution the potential energy from the springs is minimized which means the distance between the boards is maximized. Interestingly, the second question is often easier to answer. In fact, a good strategy to find the answer to question 1 is to first find the forces in question 2, and plug these forces into the equations of equilibrium (zero net force and zero torque). This strategy is exactly the strategy employed when training a support vector machine. Rather than maximizing the margin with the constraints described above, an easier dual problem is solved. The dual problem consists of minimizing a function of Lagrange multipliers that have been introduced to take care of the constraints. Lagrange multipliers appearing here having the same role as the forces should not be a surprise to the reader familiar with analytical mechanics, because in analytical mechanics forces often appear in shape of Lagrange multipliers [30], and all this comes together beautifully.

The maximal margin classifier in its simplicity has shown to work very well on high dimensional data such as genomic [20] and proteomic data [31]. There are a couple of reasons why it works so well. First, many problems in genomics and proteomics appear to be virtually linear and thus a linear method is appropriate. Second, a weakness of the maximal margin classifier is that it collapses if the training samples are not linearly separable. Remember, a condition for the decision rule is that the decision hyperplane perfectly separates the two groups of training samples. This weakness is not a problem in high dimensional data, because the high dimensionality makes data most likely linearly separable. Third, as a general rule in machine learning, when working with high dimensional data the number of dimensions needs to be reduced. Otherwise, the problem is under-determined and the resulting classifiers tend to have poor performance on test samples. The maximal margin, as any variant of SVM, has a built-in dimensional reduction. By construction the number of

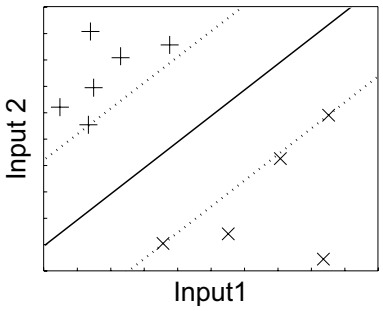

Figure 4: A dataset containing 11 data points with 2 inputs each. The two groups denoted + and x, respectively, are separated by a decision hyperplane (solid line). The margin is defined by the two dotted lines parallel to the decision hyperplane. The SVM is designed to maximize the margin without having data points between the dotted lines.

degrees of freedom equals the number of samples. More exactly, the normal to the decision hyperplane is a linear combination of the training points, which implies that we are working in the sub-space defined by the training points. In other words, the SVM decision rule can be pictured as projecting the data point down to the the normal of the decision hyperplane. The fact that this normal always belong to the sub-space defined by the training data points allows splitting this projection in two parts. First the data point is projected down to this sub-space, followed by a projection from the sub-space to the normal. Hence, directions orthogonal to the sub-space are ignored by the decision rule, which makes sense because the training points have no variation in these directions and thus contain no information. The maximal margin is very neat in its simplicity and lack of user parameters. However, SVMs would not have reach its status of fame and popularity in the machine learning community unless two tricks were added allowing non-linear classification and mislabeled data.

In 1992 Boser and colleagues [23] suggested a way to create non-linear SVMs by applying the kernel trick [32]. A key observation is that the maximal margin classifier does not depend on the data explicitly but only on the scalar products, $x_i^T x_j$, between data points. Boser and colleagues replaced the linear scalar product with a non-linear kernel function that corresponds to the scalar product in a feature space $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$. Thus the resulting algorithm finds the optimal hyperplane in feature space φ and this hyperplane may then correspond to a non-linear surface in the original space of data points. The beautiful thing is that the transformation into feature space is never needed

explicitely. Especially, as the feature space often is very high dimensional and thus it would have been computational expensive to do the transformation. One well-known example is the Gaussian kernel $K(x_i, x_j) = \exp(\frac{|x_i - x_j|^2}{\sigma^2})$ that corresponds to an infinite dimensional feature space. In general, when choosing a kernel it is not necessary to know what transformation it corresponds to, but one should know there exists a transformation, because otherwise the kernel matrix may become non-definite which implies training problems.

The next ingredient added to the SVM method was the soft-margin, which was added to avoid over-training. In machine learning over-training means the machine has adapted too detailed features from the training data leading to poor predictive power when applied on an unknown sample. The machine then has large generalization error because the rules it has learnt cannot be generalized to other samples outside the training set. One reason SVMs may get over-trained is the constraint in the maximal margin training that the classification on the training set must be perfect. It is easy to see that this might cause problems, particularly when working with noisy data and an outlier may ruin the predictive power completely. As the name suggests, soft-margins solve this by softening the constraints a bit and allowing violations. During training these violations are minimized at the same time as the margin is maximized and the balance between these two competing objectives is defined by the user.

Going back to the comparison to the boards connected with springs, we need to replace the boards because nothing could pass those boards. The situation in soft-margin SVMs resembles more of having a thick mattress that we squeeze in between the training points. We want the mattress to be as thick as possible, and the fact that it is indeed a soft mattress allows training points to compress the mattress pointwise. However, this compression costs and the thicker mattress we use, the more points we need to compress. In the end, the balance between having a thicker mattress and having less compressed points is determined by how soft the mattress is. A user defined parameter determines in the same manner, in an SVM, the balance between misclassifications and stiffness. A too stiff SVM may lead to poor generalization performance. On the other hand, making the SVM too soft means misclassifications are ignored completely during training and the SVM learns nothing. Machine learning has turned into machine ignorance.

Aims of the study

With the great progress of technology in genomics and proteomics generating an exponentially increasing amount of data, computational and statistical methods have become essential for accurate biological conclusions. As well as biology obviously benefits from development of computational methods, development of sensible methods is driven by relevant applications. This study therefore aimed at both developing algorithms, and applying computational methods to address biological questions. More specifically, the aims were

- to improve data preprocessing methods such as normalization and filtering.
- to develop and apply methods to explore large amounts of data and find relations, for example, between genes or between proteins.
- to utilize machine learning approaches for understanding biological systems.

Results and discussion

Paper I

In paper I, we present an algorithm for missing value imputation. Gene expression microarrays typically generate data of varying reliability; for instance, low-intensity data tend to be noise dominated. Therefore, microarray data analysis is commonly preceded by filtering according to some quality control criteria chosen by the investigator. Filtering leads to incomplete data that must be handled carefully because ignoring missing values might lead to a bias in analysis and inaccurate conclusions.

Many approaches have been suggested in the statistical literature [33]. Roughly speaking, methods appear in three groups. First, naive methods such as average imputation, in which each missing value is replaced by the average of the feature. A close relative is data deletion, in which calculations of statistics are based on available data, *e.g.*, calculation of correlation is based on available pairwise data. Second, maximum likelihood methods have been suggested, in which a model of the data is built followed by estimating the missing values in a maximum likelihood fashion. Third, regression methods in which a regression model is established for each feature predicting the missing value from the available features. In hot deck, a close relative to regression methods, a missing value in one feature is replaced by the corresponding value in the most similar feature.

The main idea in our approach is to, rather than to start from filtered data, embed the quality control estimate into the imputation method. We do not dichotomize values into missing or non-missing, but rather assign a continuous quality weight between zero and unity to each data value.

In other words, we suggest usage of a continuous quality weight instead of binary weights, and to examine the effects of this change, we extended two widely used methods to handle continuous weights. The two new methods: weighted average based on average imputation, and WeNNI based on a popular hot deck method named KNNimpute [34], were evaluated on replicate datasets. We found that the weighted approach improved the accuracy of imputation of data.

Conclusion: Including spot quality weights in estimation of missing values improves estimations.

Paper II

In paper II, we compare predictive power for ensembles of classifiers and for single classifiers in context of genomic and proteomic data. When designing a single classifier the aim and ambition is to select the optimal design and parameter setting for the classifier. All data is included in the training to construct the best possible classifier. In an ensemble several classifiers are constructed, and although none has as good predictive power as the optimal single classifier, the hope is that the average vote is more accurate than any single classifier. The underlying idea is that the classifiers in the ensemble compensate for each other's errors and agree on the correct decision. Clearly, to achieve this effect, there must be a diversity on opinion among classifiers. An ensemble of identical classifiers is effectively a single classifier. However, diversity should not be exaggerated. Including classifiers with poor predictive power, in its extreme random classifiers, would make the majority decision less distinct and deteriorate the predictive power of the ensemble.

In paper II, we evaluate three strategies to construct an ensemble of diverse high quality classifiers. We perform the evaluation parallel on four different datasets using two types of classifiers, nearest centroid classifiers and support vector machines. We use a cross-validation schema, whereby each classifier is trained on two thirds of training data and an ensemble of 30 classifiers is constructed. We examine the effect of feature selection, in other words, whether predictive power can be improved by using only features that individually discriminate the sample labels. We try feature selection in two ways. Either each classifier performs its own feature selection or the whole training dataset is utilized to select one consensus set of features. The former implies larger diversity as each

classifier selects different sets of features, whereas the latter possibly leads to a set of features more relevant for the task. We evaluated each strategy on four separate test datasets.

Conclusion: Ensembles of classifiers generally perform better compared to a single classifier. Feature selection improves the accuracy of prediction in most cases.

Paper III

In paper III, we use microarrays and SVMs to investigate gene expression patterns in 61 melanoma cell cultures. In many melanoma tumors, the MAPK pathway is activated by a mutation in genes *BRAF* or *NRAS*. However, these mutations rarely occur together, suggesting that a *NRAS/BRAF* double mutation would not yield any advantage for a tumor. For that reason we considered the possibility that *NRAS* and *BRAF* mutation, respectively, result in similar gene expression patterns. However, when we trained SVMs to discriminate samples carrying a mutation in either *BRAF* or *NRAS* from samples being wild type for both *BRAF* and *NRAS*, we got test performance comparable to random classifiers. Hence, we could not find a common expression pattern for the MAPK pathway.

On the other hand, when we took the three groups of samples, *BRAF* mutants, *NRAS* mutants, and double wild type samples, and trained SVMs to distinguish *BRAF* mutants from the other two groups, we got test performance significantly better than random classifiers. Moreover, when employing multidimensional scaling, we observed a separation between *BRAF* mutants and the other two groups. These findings suggest that the expression profiles in *BRAF* mutants and *NRAS* mutants are different, which means either *BRAF* or *NRAS* is signaling in an additional pathway on top of the common MAPK pathway.

Recently, Solit and colleagues [35] found that *BRAF* mutated melanomas are sensitive to treatment inhibiting MEK, whereas *NRAS* mutants showed much lower sensitivity to this treatment. This finding suggests, in line with our observations, that the whole *BRAF* mutation signaling is going through the direct downstream target *MEK*, whereas *NRAS* appears to be signaling through an additional pathway.

Conclusion: Our findings suggest that gene expression patterns in *BRAF* mutant samples are significantly different from gene expression patterns in *NRAS* mutant samples.

Paper IV

Paper IV is primarily concerned with examining the role of PTEN in breast cancer tumors. We used immunohistochemistry to determine expression levels of PTEN protein in 343 tumors, dichotomized into PTEN− (low level) and PTEN+ (high level) groups. Due to the known influence of estrogen receptor (ER) status and lymph node status on gene expression in breast cancer, we selected 105 tumors such that ER status and lymph node status were balanced in the two groups. The 105 tumors were applied on microarrays for gene expression profiling. Using the expression profiles, we constructed SVMs that could predict PTEN status with high accuracy. Moreover, we ranked the genes according to how well their expression level correlated with PTEN protein level. We identified a set of 246 discriminatory genes, which is a 15-fold overabundance compared to random chance.

Using these 246 PTEN associated genes in hierarchical clustering provided as expected two clusters containing PTEN+ and PTEN−, respectively. However, some samples appeared in the erroneous cluster, and interestingly these misclassifications correlated with mutations in *PI3K*, a component in the same signaling pathway as PTEN. More interestingly, these groups, suggested by clustering, correlated with survival. To further investigate this correlation between survival and expression of the 246 genes, we constructed nearest centroid classifiers to classify gene expression profiles according to which group they are most similar. We applied these classifiers on several publicly available datasets. For each dataset, we performed survival analysis on the groups suggested by the classifier and found that the groups correlate significantly with survival.

Conclusion: We have found a PTEN/PI3K associated gene expression signature that correlates with survival.

Paper V

In paper V, we present an algorithm to cluster protein mass spectra. We use lists of peptide peak masses extracted from the mass spectra. In order to cluster these peak lists, we introduced a score measuring the similarity between peak lists. The similarity score is calculated in two steps. First, a peak match score is calculated between pair of peaks reflecting the probability the two peaks originate from the same peptide. Second the two peak lists are aligned to find which peaks are matched, and individual match score are summed up to a total similarity score. Because the peak match score depends on mass differences in a smooth fashion, the similarity score is less sensitive to measurement errors, in contrast to bin-based approaches where a small change in mass may move a peak from a bin into the neighboring bin.

The suggested algorithm, SPECLUST, is available through a web interface (http://bioinfo.thep.lu.se/speclust.html), where peak lists can be transformed into dendrograms wherein similar proteins cluster together. The clustering gives an initial picture on how the different proteins relate to each other. Moreover, spectra can be analyzed within a cluster to see which peaks are overlapping between spectra and to reveal differences between spectra. In paper V, we point out numerous applications of this tool by using the approach on a dataset compiled from strawberry proteins.

Conclusion: The proposed algorithm for clustering of protein mass spectra is a useful tool to highlight peptides of interest for further investigations.

Future directions

As usual when questions are carefully answered, additional questions have arised during this study. Among the plethora of questions, some could be addressed by doing the following:

- Microarrays typically generate data of varying quality. Therefore, it is important to improve estimation of spot quality and incorporate spot quality weights into statistical tools. For SVMs kernels could be extended to utilize quality weights, and this choice should be evaluated and compared to using a weighted imputation approach (paper I) followed by a regular kernel.
- Further develop and validate methods to incorporate prior knowledge into statistical analysis. There are two aspects of this important field. One aspect is methods in which genes on the microarray are grouped according to *e.g.* ontology annotations and correlations between groups and sample labels are examined. Another aspect, in a sense orthogonal, is treating multiple sample labels. For instance, systematically analyze correlations between expression profiles and combinations of mutations.
- With the increasing number of spots printed on microarrays, it is getting more common to have reporters printed in replicate. Therefore, an important question is how to handle these replicates. Different strategies need to be evaluated. Is it preferable to merge replicate reporters to an average reporter? When merging and also applying imputation methods, should imputation be performed before merging or after? How is the reliability of a merged reported optimally estimated?
- Complement gene expression profiling with high-throughput proteomics to get a more complete picture of cells. Thus, statistical tools need to be developed to handle these data in a synergetic manner.
- The similarity score between peptide peak lists, suggested in paper V, can be viewed as a scalar product. Therefore, it might be worthwhile evaluating usage of the similarity score together with kernel-based methods such as multidimensional scaling, principal component analysis, and support vector machines. For SVM usage it is important to examine whether the similarity score is a valid scalar product in the sense of fulfilling Mercer's condition.

Acknowledgments

"A little nonsense now and then,
is cherished by the wisest men."
Willy Wonka

Many people have contributed significantly and I'm indeed very much obliged to all of you. Putting my big-grained goggles on, people have contributed in three ways. First, in a direct way by contributing in the quest for interesting findings and eternal glory. I've been privileged to work with sharp people, with whom you know some magic is gonna happen when you pass them the ball. Second, in a more indirect way by making the office a place you wanna be. This is equally important. You might be able to work alone, but working in boredom is doomed. They say one should not mix pleasure with business, but those people saying that have never taken part in a really functional team. Third, the wonderful people in the outer world who give me reasons to leave work. Without you, I would have run the race in full gallop and raised the flag of distress half way through.

In particular, I would like to thank:

Assistant Professor Ringnér who have given a deeper meaning to continuous guidance and support.

When trying to find words to thank Jari, my mind drives away to the legend of Marcus Wallenberg. The legend tells how he compared devaluation to urinating in your pants. First, it gets warm and nice, then starts the nastiness. I'm not saying Jari is like urinating in your pants, but sort of the other way around. He doesn't base his strategy on avoiding the immediate nastiness, because he knows that behind the corner waits a warm and nice feeling. He prefers temporary solutions before momentary ones. Thanks, for making me understand the concept of non-local optimization.

All my co-authours, and in particular Leisl and Lao. The close collaborations with you, exchanging ideas, interpretations, and experiences have been more than fruitful and inspiring. You've been an extra dimension; like what the sound is for TV. One can still watch without it, but it's not the same thing.

Carsten convinced me the three-letter combination: phd - is a good combination and gave me the opportunity to join his group. Patrik who always brings in sharpness in a discussion. Mr Miyagi spreads his wax on-wax off-attitude. Giorgio helped me with the number theory. Imaginary numbers might be beautiful, only useful in theory though. The hilarious Dhonte et Dhonde.

Stefan, Fredrik, and Michael among other virtual roommates in the attic, who appreciate a little nonsense every now and then. Spring who manages to stand me and my mess. Göran for typesetting paper IV.

Dewi. Any word seems too small. Kamu sangat hebat. Saya ingin kita selalu bersama. Peluk Besar.

My family; my parents for letting me become who I am, and making me believe that is a good thing. Pontus, who I can't imagine my world without.

My Massa, the dude, the beginning and the end. Markus, thanks for all great laughs and everything you've taught me on purpose, without purpose, and beyond the concept of purpose. It would be wrong to summarize. It would be ignoring the details. It would prove my ignorance. Although, I have to say it has been a trip, all the way from BWI to now - wherever we are - who cares? "Momentum is everything". Like Carson would put it "You're the business partner. You're the Dolce of Dolce and Gabana. You're the Pra of Prada." You have been to me as Arsene has been to Freddie. With the excellent team you lined up, it was just to run at full speed and then the ball was served as cheese on a silver plate. Creativity, stamina, and technical advice. All with an extreme sense of details. It's been a pleasure, Massa.

References

[1] Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid**. *Nature* 1953, **171**(4356):737–738.

[2] Kornberg RD: **Chromatin structure: a repeating unit of histones and DNA**. *Science* 1974, **184**(139):868–871.

[3] Travers AA: **Cyclic re-use of the RNA polymerase sigma factor**. *Nature* 1969, **222**(193):537–540.

[4] Jacob F, Monod J: **Genetic regulatory mechanisms in the synthesis of proteins**. *J Mol Biol* 1961, **3**:318–356.

[5] Eron L, Block R: **Mechanism of initiation and repression of in vitro transcription of the lac operon of Escherichia coli**. *Proc Natl Acad Sci U S A* 1971, **68**(8):1828–1832.

[6] Zubay G, Schwartz D, Beckwith J: **Mechanism of activation of catabolite-sensitive genes: a positive control system**. *Proc Natl Acad Sci U S A* 1970, **66**:104–110.

[7] Englesberg E, Irr J, Power J, Lee N: **Positive control of enzyme synthesis by gene C in the L-arabinose system**. *J Bacteriol* 1965, **90**(4):946–957.

[8] Hanahan D, Weinberg RA: **The hallmarks of cancer**. *Cell* 2000, **100**:57–70.

[9] Baker SJ, Fearon ER, Nigro JM, Hamilton SR, Preisinger AC, Jessup JM, vanTuinen P, Ledbetter DH, Barker DF, Nakamura Y, White R, Vogelstein B: **Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas**. *Science* 1989, **244**(4901):217–221.

[10] Harris H: **Cell fusion and the analysis of malignancy**. *Proc R Soc Lond B Biol Sci* 1971, **179**(54):1–20.

[11] Kerr JF, Wyllie AH, Currie AR: **Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics**. *Br J Cancer* 1972, **26**(4):239–257.

[12] Kastan MB, Zhan Q, el Deiry WS, Carrier F, Jacks T, Walsh WV, Plunkett BS, Vogelstein B, Fornace AJJ: **A mammalian cell cycle checkpoint pathway utilizing p53 and GADD45 is defective in ataxia-telangiectasia**. *Cell* 1992, **71**(4):587–597.

[13] Armitage P, Doll R: **A two-stage theory of carcinogenesis in relation to the age distribution of human cancer**. *Br J Cancer* 1957, **11**(2):161–169.

[14] Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC: **Linkage of early-onset familial breast cancer to chromosome 17q21**. *Science* 1990, **250**(4988):1684–1689. [Case Reports].

[15] Easton DF, Ford D, Bishop DT: **Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium**. *Am J Hum Genet* 1995, **56**:265–271.

[16] Griffin PR, MacCoss MJ, Eng JK, Blevins RA, Aaronson JS, Yates JRr: **Direct database searching with MALDI-PSD spectra of peptides**. *Rapid Commun Mass Spectrom* 1995, **9**(15):1546–1551.

[17] Larsen MR, Roepstorff P: **Mass spectrometric identification of proteins and characterization of their post-translational modifications in proteome analysis**. *Fresenius J Anal Chem* 2000, **366**(6-7):677–690.

[18] Kanji GK: *100 Statistical Tests*. Sage Publications Ltd 1993.

[19] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530–536.

[20] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data**. *Bioinformatics* 2000, **16**(10):906–914.

[21] Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks**. *Nat Med* 2001, **7**(6):673–679.

[22] Vapnik V, Lerner A: **Pattern recognition using generalized portrait method**. *Automation and Remote Control* 1963, **24**.

[23] Boser BE, Guyon IM, Vapnik VN: **A training algorithm for optimal margin classifiers**. In *In D. Haussler, editor, 5th Annual ACM Workshop on COLT*, ACM Press 1992:144–152.

[24] Cortes C, Vapnik V: **Support-Vector Networks**. *Machine Learning* 1995, **20**(3):273–297.

[25] Mercer J: **Functions of positive and negative type, and their connection with theory of integral equations**. *Proc. Roy. Soc. London* 1908, **83**:69–70.

[26] Riesz F, Nagy B: *Functional analyses*. Dover Publications 1955.

[27] Kuhn H, Tucker A: *Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press 1951 :481–492.

[28] Vapnik V, Chervonenkis A: **On the uniform convergence of relative frequencies of events to their probabilities**. *Theory Prob. Applic.Proc.* 1971, **17**(2):264–280.

[29] Cristianini N, Shawe-Taylor J: *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press 2001.

[30] Goldstein H, Poole C, Safko J: *Classical mechanics*. Addison Wesley 2002.

[31] Ressom HW, Varghese RS, Abdel-Hamid M, Eissa SAL, Saha D, Goldman L, Petricoin EF, Conrads TP, Veenstra TD, Loffredo CA, Goldman R: **Analysis of mass spectral serum profiles for biomarker selection**. *Bioinformatics* 2005, **21**(21):4039–4045.

[32] Aizerman M, Braverman E, Rozonoer L: **Theoretical foundations of the potential function method in pattern recognition learning**. *Automation and Remote Control* 1964, **25**:821–837.

[33] Little R, Rubin D: *Statistical analysis with missing data*. John Wiley and Sons. 1987.

[34] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays**. *Bioinformatics* 2001, **17**(6):520–525.

[35] Solit DB, Garraway LA, Pratilas CA, Sawai A, Getz G, Basso A, Ye Q, Lobo JM, She Y, Osman I, Golub TR, Sebolt-Leopold J, Sellers WR, Rosen N: **BRAF mutation predicts sensitivity to MEK inhibition**. *Nature* 2006, **439**(7074):358–362.